

Annotation of Semantic Relations in Patent Documents

Valentina Bartalesi Lenzi
Rachele Sprugnoli
CELCT / Via Sommarive 18, 38100
Povo (TN), Italy
{bartalesi, sprugno-
li}@celct.it

Emanuele Pianta
FBK / Via Sommarive 18, 38100 Povo
(TN), Italy
pianta@fbk.eu

Abstract

This paper presents the theoretical bases and quantitative results of an activity consisting in manually annotating part-whole and motion relations in patent documents. The aim of this activity was creating a gold standard for the evaluation of an automatic relation extraction tool developed by FBK-irst within the PATExpert project. For this purpose, we took the annotation scheme created for the Relation Detection and Recognition task of the Automatic Content Extraction (ACE) Program as a starting point, adapting attribute values to our specific needs. A number of syntactic and semantic aspects have been considered in the annotation, e.g. syntactic class, modality, and semantic role of the relation arguments. The annotation has been inspired by some of the major theories regarding meronymic and motion relations.

1 Introduction

Patents have a big impact on industry, business, and law communities both at the national and international level: they affect the economy providing incentives for innovation, dissemination of scientific knowledge and technology transfer. For this reason, advanced patent documentation processing techniques are in great demand.

In recent years, a particular effort in the patent processing research field concerned patent document retrieval systems and many workshops have been organized on this topic (e.g. NTCIR 2002-2007, CLEF-IP09, and TREC-CHEM09).

In order to improve retrieval performance, the PATExpert European project (Wanner et al., 2008) used semantics-driven processing tech-

niques. Indeed, the goal of the project was to change the traditional approach to patent processing from textual to semantic, in the Semantic Web perspective. In this context, the aim of our work was to create a gold standard for the evaluation of an automatic relation extraction tool developed by FBK-irst. We choose *part-whole* and *motion* relations because of their numerical relevance in patent documents. The automatic annotation of such relations by the PATExpert system should allow a user to ask queries like: “Retrieve all patent sentences mentioning an optical head that includes a Wollaston prism” or “Retrieve all patent sentences mentioning the movement of a cam member from the first to the second position”.

This paper is structured as follows. Section 2 describes the annotated corpus with an overview of annotation tool, formats, and annotation scheme. Section 3 details the annotation of part-whole relations with a subsection dedicated to the lexico-syntactic patterns in which meronymic relations can be expressed. Section 4 illustrates the annotation of motion relations with a focus on motion verbs and associated prepositions. Both Sections 3 and 4 provide examples extracted from the annotated corpus and quantitative data about the annotation. Finally, in Section 5 we draw some conclusions.

2 Annotation Process

Our work concerned the manual annotation of: (i) part-whole relations in a corpus of about 50,000 words; (ii) motion relations in a corpus of about 100,000 words. These two corpora are made up of 16 patents each, equally divided into two technology areas: 8 about optical devices and 8 regarding machine tools. The annotation process involved two annotators and required 4.5

person/months. Specific guidelines were developed during the project, describing annotation tags and attributes, and illustrating what and how to annotate with examples. After a training phase, annotators have been interacting and negotiating common solutions to controversial annotations.

We used Callisto, a freely distributed software developed by the MITRE Corporation, as annotation tool (<http://callisto.mitre.org>). Input files were raw texts in UTF-8 encoding, while output data were made available in the ACE Program Format (APF).

The following attributes of the ACE Event Task (Dodgington et al, 2004) have been adopted for the annotation:

- *Extent*, the smallest or closest possible portion of text in which the relation is expressed;
- *Arguments*, representing the two entities involved in the relation;
- *Semantic Type* and *Semantic Subtype* of the relation, which categorize the relation on the basis of its meaning;
- *LexicalCondition*, the syntactic class of the lexical item expressing the relation;
- *Modality* of the relation, which indicates if the relation is asserted or not.

In order to adapt this task to our specific needs, we modified some attribute values of ACE Event Task. More specifically: (i) values of the *Semantic Type* were only *part-whole* or *motion*; (ii) values of the *Semantic Subtype* have been used to characterize the semantic roles of motion relations (see Section 4); (iii) values of the *LexicalCondition* attribute were defined in a different way for part-whole and motion relations, see sections 3.1 and 4.1; (iv) *Modality* values were redefined: instead of the Asserted and Other values of the ACE Event Task, we introduced Asserted, Negated and Possibility values.

3 Annotation of Part-Whole Relations

The first part of our project concerned the annotation of parthood relations, that are the relations between a segment or a portion in which an entity is divided and the entity itself.

Parthood relations are studied from the philosophical and linguistic point of view. In the first perspective that are labeled as mereological relations while, in the second approach, they are often referred to as meronymic relations. As described in Keet (2006), there are differences and overlapping between the two perspectives. In our work, we choose the meronymic perspective tak-

ing the Winston et al. (1987) taxonomy as a reference. In this taxonomy meronymic relations are divided in six classes capturing the different ways in which parts contribute to the composition of the whole:

- *Component-Object*: an integral object is formed by some component parts that have a particular relationship to one another and to the whole. The components have a specific functionality, can be removed from the integral object and are patterned within the whole which they comprise. E.g. “a handle is part of a cup”.
- *Portion-Mass* (also known as Portion-Object (Odell, 1994)): the whole is considered as a homogeneous mass and its portions are similar to each other and to the whole. E.g. “a slice is part of a pie”.
- *Place-Area*: spatial relation among regions in a geographical (e.g. “an oasis is part of a desert”) or geometrical sense (e.g. “the side of a building is part of that building”).
- *Member-Collection*: that represents the membership in a collection. E.g. “a tree is part of a forest”.
- *Stuff-Object*: the relation between the stuff of which an object is made of and the object itself. E.g. “water is partly hydrogen”.
- *Feature-Activity*: captures the relation between phases of an activity and the activity as a whole. E.g. “paying is part of shopping”.

Only three of these types of meronymic relations are coded in WordNet 3.0: *Component-Object*, *Member-Collection*, *Stuff-Object*. The *Portion-Mass* relation is not present, while *Place-Area* is conflated in the *Component-Object* relation, and *Feature-Activity* is expressed as an entailment relation between verbal concepts (<http://wordnet.princeton.edu>).

Given the characteristics of the patent documents in our corpus, we choose to annotate only the following part-whole relations: (i) *Component-Object*, e.g. “tool clamping device comprising a draw bar”; (ii) *Portion-Mass*, e.g. “excessive portion of the adhesive; (iii) *Place-Area*: in our corpus we found examples of this relation only in the geometrical sense, thus we annotated the relation between the extent of a physical object and part of that extent: e.g. “a surface of an optical disc”. Because of the specialized subject dealt in the patent corpus, the presence of the other types of relations is only marginal.

During the annotation, a special effort has been put in making a clear distinction between part-whole relations and other relations, which

according to Winston et al. (1987) can be expressed in a similar way and can be confused with them. In particular, meronymy needs to be distinguished from the following relations: (i) *Topological Inclusion*: relation between a content and its container, e.g. “there is a foam in the adhesive layer”; (ii) *Attribution*: relation between an object and its attribute, e.g. “the thickness of the Wollaston prism”; (iii) *Attachment*: two components are attached to each other in some way, e.g. “the front panel attached to the casing”; (iv) *Ownership*: relation between a person or an institution and something that they own, e.g. “a woodworker have several boards”.

Although the transitivity of meronymic relations is one of the most discussed issues in the relevant literature (Lyon 1977, Cruse 1986, Vieu 2006), this problem did not affect our work because we annotated only explicit relations and not the ones that can be inferred. For example, in the sentence “that said optical element comprises a Wollaston prism made up of two crystal optical elements” we annotated a parthood relation between “a Wollaston prism” and “that said optical element” and another between “two crystal optical elements” and “a Wollaston prism”, but not the inferable relation between “two crystal optical elements” and “that said optical element”.

3.1 Part-Whole Lexico-Syntactic Patterns

Meronymic relations can be expressed by a variety of lexico-syntactic patterns. Some of these patterns are unambiguous (e.g. “X is part of Y”) but others are ambiguous and indicate part-whole relations only in specific contexts.

Girju et al. (2006) groups lexico-syntactic patterns in four clusters on the basis of their semantic similarity. Each of these patterns presents ambiguity problems. See the following list:

1. *Genitives and the verb “to have”*. There are two kinds of genitives in English: the *s*-genitive and the *of*-genitive. Both constructions encode different semantic relations in addition to the part-whole one, such as Possession (“Alice’s house”), Kinship (“Alice’s sister”) and Source/From relations (“president of Italy”) (Moldovan and Badulescu, 2005). Also the verb “to have” conveys various kinds of relations, as the 19 senses related to that verb in WordNet 3.0 point out: e.g. to possess (“how many houses does she have?”), to make (“she has a party”), and to consume (“I’d like to have another slice of cake”).
2. *Noun compounds*. Noun compounds are defined as any noun phrase composed of a noun

head and one or more nominal modifiers (e.g. “linen bag”). The correct interpretation of the relation between the modifier(s) and the head noun is a complex process: the semantic relation is implicit and highly context-dependent (Downing, 1977); world knowledge is often necessary; more than one kind of relation can be encoded by the same compound (e.g. “linen bag” can express a Stuff-Object relation: “bag made of linen”, or a Purpose relation: “bag for linen”).

3. *Prepositions*. Also prepositional constructions can express several semantic relations, depending on the context and on the meaning of the two nouns (Litkowski and Hargraves, 2005). For example, here are some possible semantic relations expressed by the preposition “in”: Locative (“Alice is in New York”), Temporal (“Alice was born in 1980”), and Meronymy (“the engine in the car”).
4. *Other*. This last cluster includes verbs different from “to have” (e.g. “to consist, include, comprehend”) and expressions like “X is member of Y” and “X is a branch of Y”.

Although the classification of Girju and colleagues has been taken into account in order to recognize and distinguish meronymic relations, we adopted a classification which is more clearly based on the syntactic function of the item expressing the relation. For this purpose we used the *LexicalCondition* attribute to indicate the syntactic element that provides justification for the tagging of each relation. Five values, denoting five different syntactic classes, can be assigned to that attribute:

- a. *Verbal*: all verbs, including “to have”, that determine a part-whole relation. E.g. “the rotary holder *has* a shaft”, “tool clamping device *comprising* a draw bar”.
- b. *Preposition*: all prepositions, including “of”, that introduce prepositional phrases. E.g. “optical recording layers *of* an optical disk”, “a portion *with* a radial surface”.
- c. *Possessive*: possessive adjectives and pronouns indicating that a component is part of the whole. The anaphoric reference was solved and the preceding discourse entity to which the possessive refers was annotated. For example, in the sentence “optical disks bonded at *their* faces” we annotated the relation between “faces” and “optical disks”.
- d. *PreModifier*: noun modifiers that precede the head noun and can be the part or the whole of that head. In our corpus the part is always in the second position. E.g. “the substrate *side*”.

e. *Adverbial*: adverbs that suggest a part-whole relation. E.g. “the optical disc is read from one side *thereof*” (see also “wherein”). Even if “thereof” and similar words could be classified as post-positions from a distributional point of view, in this paper they are classified as adverbs according to current dictionaries.

Comparing our classes with the ones proposed by Girju et al. (2006), some intersections and differences can be noticed: (i) our class *Verbal* groups part of the first (“to have”) and part of the fourth (verbs different from “to have”) Girju’s clusters; (ii) the *Preposition* class comprises the third and part of the first (*of*-genitives) clusters; (iii) *PreModifier* corresponds to the second cluster (noun compounds). Adverbial and Possessive syntactic classes seem to be not considered in the Girju classification. The expression “X is member of Y”, inserted in the fourth cluster, and s-genitives do not occur in our corpus.

3.2 Quantitative Data

A total of 1,015 part-whole relations have been annotated. Table 1 shows that the most frequent *Modality* value is the Asserted one (“a recording medium film 3 having track grooves”), while a small group of relations is annotated as Possibility (“the DVD-RAM may have a rom area”) and Negated (“the rom area has no reflective layer”).

Modality	Count
Asserted	963 (94.88%)
Possibility	46 (4.53%)
Negated	6 (0.59%)

Table 1. Occurrences of the *Modality* attribute

Table 2 reports data about the *LexicalCondition* attribute: a balanced distribution between the two most frequent syntactic classes can be noticed. The high frequency of prepositions is due to the fact that patent documents often describe highly structured objects. Prepositions are used to build complex noun phrases describing each single part of those objects

LexicalCondition	Count
Verbal	499 (49.16%)
Preposition	440 (43.35%)
PreModifier	37 (3.65%)
Adverbial	20 (1.97%)
Possessive	19 (1.87%)

Table 2. The *LexicalCondition* attribute

The Verbal *LexicalCondition* was analyzed in more detail, by collecting all verbs in the corpus expressing part-whole relations. Table 3 provides the list of verbs ranked according to their frequency. The top ranking verbs are not domain-specific (e.g. “have”), while patent-specific verbs have lower frequency (e.g. “incorporate”).

Verbs	Count
have	116
provide	94
include	85
comprise	74
form	72
locate	13
coexist	11
characterize	10
make	8
compose	7
incorporate	6
place	3

Table 3. Verbs determining part-whole relations

Table 4 presents the percentage of occurrence of prepositions involved in part-whole relations with examples from the corpus. Notice that preposition “of” has frequency above 80%.

Prep.	%	Examples
of	80.34%	a cam face 290 <i>of</i> a spring disk
in	9.57%	a recording pit <i>in</i> the data area
on	5.42%	a tab 106 <i>on</i> the slide block
with	3.03%	a bore 382 <i>with</i> a stop shoulder
within	0.68%	the region <i>within</i> the data area
at	0.48%	the pin is <i>at</i> the lower end
into	0.48%	the pipes P <i>into</i> one unit

Table 4. Prepositions in part-whole relations

The quantitative data extracted from our corpus can be compared with the ones reported in Girju et al. (2006). Interestingly, their fourth cluster (verbs different from “to have”) covers only 6.36% of the part-whole patterns discovered in their corpus, whereas, even ignoring the verb “to have”, the occurrences of the Verbal *LexicalCondition* in our corpus are more than 38%. This high percentage of verbs is justified by the nature of our documents. While in not specialized texts part-whole relations belong mostly to the background knowledge, in patents they are explicitly expressed and emerge from verbs used in the detailed technical description of the invention.

As far as *noun-noun* compounds are concerned, the corresponding cluster in Girju has a coverage of 16.07%, while in our corpus this pattern is not very frequent (3.65%). Again this can be explained by the specific needs of patent writing, and more in general technically writing: *noun-noun* constructions are potentially more ambiguous (both syntactically and semantically) than *noun-prepositional-phrase* constructions, and we can assume that they are intentionally avoided by patent writers.

4 Annotation of Motion Relations

The second cycle of annotation concerned motion relations. While part-whole relations are inherently binary, motion relations can have more than two arguments (e.g. “the optical head 7 is moved in a radial direction by optical head moving means”), or just one argument (e.g. “while the optical head is raised”). This is problematic from a practical point of view, because the Callisto interface only allows for annotating two-place relations. To solve this problem we decided to use what in Callisto are two-place relations to annotate the link between a lexical item expressing a motion relation and one of its arguments. The *Semantic Subtype* attribute has been used to indicate the semantic role of each argument involved in the relation. In the rest of this section, the term “relation” is used to refer to a complex textual object relating a number of one or more semantic roles within a motion event.

The analysis of motion concepts illustrated in Talmy (1985) is taken as a starting point for many studies about motion. Talmy assumes that the concept of motion necessarily includes four components: (i) a figure, that is an individuated object; (ii) the motion of this object; (iii) a path, along which the motion takes place, consisting of an initial, a medial and a final portion, and (iv) a ground with respect to which the motion is conceptualized. The manner and the cause of motion can be used in addition as optional components.

Talmy’s theory is also the basis for the analysis of motion relations developed within the FrameNet project (Baker, et al. 1998) which was taken as a reference point for our work. FrameNet defines a *Motion_Scenario* where the concept of motion is expressed by many different frames. In our annotation we took into consideration two main frames, i.e. *Motion* and *Cause_motion*, and, in addition, four other frames that inherit from the *Motion* one: i.e. Ar-

iving, *Departing*, *Traversing* and *Motion_directional*.

A working list of semantic roles has been selected from the above mentioned frames. The following list presents definitions and examples for each annotated semantic role (the semantic role filler is underlined):

- *Theme*: the entity that changes location, “the optical pickup moves in the radial directions”;
- *Source*: the position the Theme occupies before its change of location, “an optical head moves from the current position”;
- *Goal*: the position the Theme ends up in, “an optical head to move to a target track from the current position”;
- *Path*: the space covered by the Theme between the Source and the Goal, “the collar moved along the drill”;
- *Direction*: the line in which the Theme moves, “the pick-up moves in that direction”;
- *Distance*: the extent of the Motion, “the optical pickup moves to a certain extent”;
- *Manner*: the way in which the motion takes place, “the optical head moves at a low speed”;
- *Cause*: the object whose action causes the motion of the Theme, “the object moved by these motors”.

All previous semantic roles belong to the *Motion* and *Cause_motion* frames, that are the most general frames indicating inchoative and causative motion respectively. In particular, *Goal*, *Path*, *Source*, *Theme* and *Manner* are semantic roles in both frames, but the last one is a non-core element. *Direction* and *Distance* are specific to the *Motion* frame, while *Cause* is a core element of the *Cause_motion* frame. The *Area* semantic role, indicating the region in which the movement takes place without a single linear path, is a core element of both frames. In our annotation scheme we did not introduce this semantic role because our aim was to annotate only linear movement.

In annotating motion relations, we did not distinguish among motion frames because this was not relevant for the aim of the PATExpert relations extraction task; however, in some cases, it is possible to recognize a specific frame by looking to the semantic roles involved in relation: e.g. in the sentence “the coil spring pushes the rockers” the presence of “the coil spring”, that is the *Cause* role filler, indicates that the motion verb “to push” belongs to the *Cause_motion* frame.

4.1 How Motion is Expressed

For each semantic role we annotated the *LexicalCondition*, which indicates how the motion is syntactically expressed. By analyzing the data, we found seven possible values rather than the five identified for the part-whole relations:

1. Verbal-Direct: verbs accompanied by a direct argument. E.g. “the puller *moves* to a target position”.
2. Verbal-Prep: verbs with prepositional phrases. E.g. “the operative gear 65 *moves* along the guides”.
3. Nominal-Prep: nominal heads followed by prepositional phrase. E.g. “*the movement* of the guide”.
4. Nominal-Poss: nominal heads modified by possessives. E.g. “its *movement*”.
5. Nominal-PreMod: pre-modifiers followed by nominal heads. E.g. “the guide *movement*”.
6. Nominal-Head: nominal phrases where the head expresses an attribute of the motion. E.g. “*the movement* distance”.
7. Adverbial: adverbs encoding manner of motion. E.g. “the guide quickly *moves*”.

4.1.1 Motion Verbs

To identify motion verbs we took into consideration the lexical units reported for the frames cited in the previous section: Motion (“the puller is *moved* in the cross direction”), Cause_motion (“the lens actuator is *driven* by the control unit”), Arriving (“the cam member *reaches* the second position”), Departing (“the reading and / or writing head *departing* from the first position”), Traversing (“the guide rod *passes through* a guide hole”), and Motion_directional (“the front guide pins *rise* up along the slanting grooves”).

A motion interpretation can be assigned also to some verbs not included in the mentioned frames. This is the case, for example, of the verb “to raise” that in our corpus is clearly involved in a motion relation (“the drill guide assembly 20 is *raised* to a higher position”) whereas in FrameNet this verb is included only in the Causation, Building, Cause_change_of_position_on_a_scale frames. We also annotated the verb “to displace”, that is relatively well attested (e.g. “the cam member is *displaced* from the first position to the second position”), although it is not present at all in FrameNet.

Occurrences of motion verbs within fictive motion sentences (Talmy, 1996) were not taken into account, even if many examples of radiation paths are present in documents concerning opti-

cal devices (e.g. “the parallel rays of the laser beam *passed through* the beam splitter”).

Although, in English, verbs can conflate a manner component (Talmy, 1991) as in the case of “to slide”, we decided to annotate only linguistic elements that explicitly express manner of motion. Table 5 presents schematically the semantic roles identified in the sentence: “The cam member slides slowly along the restricting groove”.

Semantic Roles	Linguistic elements
Theme	The cam member
Manner	slowly
Path	the restricting groove

Table 5. Example of semantic role annotation

4.1.2 The Role of Prepositions

Spatial prepositions turned out to play a crucial role in identifying semantic roles fillers. The literature distinguishes between directional prepositions, like *into* and *through*, and locative (or stative) prepositions, like *in* and *under* (Sablayrolles, 1995). The difference lies in the contribution that the two types of prepositions make to the aspect of the verbal predicate. For example, in the sentence “to move the machining head 5 *along* a desired path”, the preposition “along” contributes to determining the atelic nature of the movement event, while in the “the drill bit slides *into* the drilling guide bushings” the preposition “into” determines the telic nature of the sliding event.

During the annotation, we faced the problem of correctly interpreting directional prepositions which can encode both goal and path roles (Folli, 2001). To solve this difficulty, we decided to adopt the classification of directional prepositions proposed by Zwarts (2006) (see Table 6).

Prepositions	Examples
Source: e.g. <i>from, out of</i>	- the optical pick-up 47 is moved <i>from</i> the inner side
Goal: e.g. <i>into, to, onto</i>	- the drill bit slides <i>into</i> the drilling guide bushings
Route: e.g. <i>via, through, across</i>	- guide pins which slide <i>across</i> slide bushings
Comparative: e.g. <i>towards</i>	- the slider moves toward the turn table 46
Constant: e.g. <i>along</i>	- the cam follower slides <i>along</i> the support shaft
Periodic: e.g. <i>up and down</i>	- the drill guide assembly 20 is moved <i>up and down</i>

Table 6. Classes of directional prepositions

According to this approach, the preposition *to* has only a goal meaning while the preposition *through* refers only to path. Holistic prepositions, like *around*, were not considered in our annotation because they refer to a non linear path that enclose the reference object.

4.2 Quantitative data

A total of 624 motion relations have been annotated. For each relation there are several semantic roles. As shown in Table 7, 1,324 semantic role fillers have been detected, with a clear prevalence of the Theme role. Notice that 34 relations do not have the Theme semantic role: for example, in the sentence “the movement caused by the driving mechanism”, only the Cause semantic role is present, i.e. “the driving mechanism”.

Semantic Roles	Count
Theme	590 (44.56%)
Goal	166 (12.54%)
Direction	159 (12.01%)
Cause	134 (10.12%)
Path	97 (7.32%)
Manner	81 (6.12%)
Source	63 (4.76%)
Distance	34 (2.57%)

Table 7. Occurrences of semantic roles

Tables 8 and 9 illustrate that, similarly to what we saw for part-whole relations, the Asserted value of the *Modality* attribute is the most frequent and, for what concerns the *LexicalCondition* attribute, Verbal-Direct and Verbal-Prep values have the highest percentages of occurrence.

Modality	Count
Asserted	1213 (91.61%)
Possibility	99 (7.48%)
Negated	12 (0.91%)

Table 8. Occurrences of the *Modality* attribute

LexicalCondition	Count
Verbal-Direct	631 (47.66%)
Verbal-Prep	452 (34.14%)
Nominal-Prep	118 (8.91%)
Adverbial	69 (5.21%)
Nominal-Head	36 (2.72%)
Nominal-Poss	13 (0.98%)
Nominal-PreMod	5 (0.38%)

Table 9. The *LexicalCondition* attribute

Motion verbs involved in relations with Verbal-Direct *LexicalCondition* have been extracted and reported in Table 10. Notice that the 631 verbal occurrences correspond to only 23 lemmas with 338 occurrences of “to move”. This data denotes a loss in lexical richness that often matches with a strong syntactic similarity among patent documents.

Verbs	Count
move	388
drive	59
position	37
reach	35
push	25
guide	22
displace	14
slide	11
pass, lift, carry, ride, depart, go, overrun, place, pull, raise, reposition, run, transfer, rise, transport	< 10

Table 10. Occurrences of motion verbs

5 Conclusions

In this paper we presented the results of an activity aiming at annotating part-whole and motion relations in a corpus of patent documents. The theoretical background that was taken into consideration in the design of the annotation scheme and in the practical annotation activity has been illustrated and discussed.

The annotation activity had to face a number of difficulties, the first of which was sheer text understanding given the technical nature, the domain and genre specificity of the corpus. From an annotation point of view the two hardest tasks turned out to be the identification of verbs encoding part-whole relations and the choice of the correct Semantic Role for each argument of motion relations.

The annotation of the part-whole relation that we performed can be compared with a similar work by (Girju et al., 2006). However our work concerns a genre- and domain- specific corpus, which brought to our attention a number of peculiarities that have been presented and tentatively explained. On the other side, to our knowledge, our annotation of motion relations cannot be compared to any other similar work. This activity has raised a number of interesting issues, related to the nature of motion concepts. One important issue is related the fact that motion events require a number of core semantic roles, which can vary

from one to 4/5. This should be compared with current studies on automatic relation extraction, which focus on binary relations.

A number of quantitative data about the frequency of the parthood and motion relations in our corpus, as well as how they are expressed in lexical and syntactic terms has also been provided.

As for future work, first of all we need to calculate the inter-annotator agreement. Then, we plan to improve the linguistic analysis of motion verbs, e.g. looking into classifications like the one presented by (Pustejovsky and Moszkowicz, 2008).

Acknowledgments

This paper has been realized with the support of the PATExpert European Project, and the Live-Memories project funded by the Autonomous Province of Trento.

References

- Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proc. of the COLING-ACL*, pages 86-90. Montreal, Canada.
- Cruse, Alan. 1986. *Lexical semantics*. Cambridge University Press, Cambridge.
- Doddington, George, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation. In *Proc. of Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 837-840, Lisbon, Portugal.
- Downing, Pamela. 1977. On the creation and use of English compound nouns. In *Language*, 53(4): 810-842.
- Folli, Raffaella, and Gillian Ramchand. 2005. Prepositions and results in Italian and English: an analysis from event decomposition. In H. Verkyul, H. Van Hout, and H. De Swartz, editors, *Perspectives on aspect*, Springer, Dordrecht, pages 81-105.
- Girju, Roxana, Adriana Badulescu, and Dan Moldovan. 2006. Automatic Discovery of Part-Whole Relations. *Computational Linguistics*, 32(1): 83-135.
- Keet, Catharina M. 2006. Introduction to part-whole relations: mereology, conceptual modelling and mathematical aspects. *KRDB Research Centre Technical Report KRDB06-3*, Faculty of Computer Science, Free University of Bozen-Bolzano, Italy.
- Litkowski, Ken C., and Orin Hargraves (2005). The Preposition Project. In *Proc. of the Second ACL-SIGSEM Workshop on The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 171-179, Colchester, England.
- Lyons, John. 1977. *Semantics*. Cambridge University Press, Cambridge.
- Moldovan, Dan and Adriana Badulescu. 2005. A semantic scattering model for the automatic interpretation of genitives. In *Proc. of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 891-898, Vancouver, BC, Canada.
- Odell, James J. 1994. Six different kinds of composition. *Journal of Object-Oriented Programming*, 5(8): 10-15.
- Pustejovsky, James D. and Jessica L. Moszkowicz. 2008. Integrating Motion Predicate Classes with Spatial and Temporal Annotations. In *Coling 2008: Companion volume Posters and Demonstrations*, pages 95-98, Manchester, England.
- Sablayrolles, Pierre. 1995. The Semantics of Motion. In *Proc. of the 7th Conf. of the European Chapter of the Assoc. for Computational Linguistics (EACL 1995)*, pages 281-283, Toulouse, France.
- Talmy, Leonard. 1985. Lexicalization Patterns: Semantic Structure in Lexical Forms. In T. Shopen, editor, *Language Typology and Syntactic Description III: Grammatical Categories and the Lexicon*, CUP, Cambridge, pages 57-149.
- Talmy, Leonard. 1991. Path to realization: a typology of event conflation. In *Proc. of the Seventeenth Annual Meeting of the Berkeley Linguistics Society*, pages 480-519, Berkeley, CA.
- Talmy, Leonard. 1996. Fictive motion in language and "ception". In P. Bloom, M.A. Peterson, L. Nadel, and M.F. Garrett, editors, *Language and space*, MIT Press, Cambridge, pages 211-276.
- Vieu, Laure. 2006. On the transitivity of functional parthood. *Applied Ontology*, 1(2):147-155.
- Wanner, Leo, Sören Brüggemann, Joan Codina, Barrou Diallo, Enric Escorsa, Mark Giereth, Yiannis Kompatsiaris, Symeon Papadopoulos, Emanuele Pianta, Gemma Piella, Ingo Puhmann, Gautam Rao, Martin Rotard, Pia Schoester, Luciano Serafini, and Vasiliki Zervaki. 2008. Towards Content-Oriented Patent Document Processing. *World Patent Information Journal*, 30(1): 21-33.
- Winston, Morton, Roger Chaffin, and Douglas Hermann. 1987. A taxonomy of part-whole relations. *Cognitive Science*, 11(4):417-444.
- Zwarts, Joost. 2008. Aspects of a typology of direction. In S. Rothstein, editor, *Theoretical and Crosslinguistic Approaches to the Semantics of Aspect*, John Benjamins, Amsterdam, pages 79-106.