

LIRICS semantic role annotation: design and evaluation of a set of data categories

Volha Petukhova and Harry Bunt

Department of Communication and Information Sciences
Tilburg University, Netherlands
{v.petukhova,harry.bunt}@uvt.nl

Abstract

In this paper we report on the analyses of alternative approaches to semantic role annotation (FrameNet (FrameNet, 2005), PropBank (Palmer et al., 2005) and VerbNet (Kipper, 2006)) with respect to their models of description; granularity of semantic role sets; definitions of semantic roles concepts; and consistency and reliability of annotations, and we propose a methodological basis for identifying and analysing semantic roles, including a data-driven account of defining semantic role concepts. We present evaluation results of the defined concepts for semantic role annotation concerning the redundancy and completeness of the tagset, and concerning the reliability of annotations in terms of inter-annotator agreement.

1. Introduction

Semantic roles have often proved to be useful labels for stating linguistic generalisations of various sorts. There is, however, a lack of agreement on their defining criteria, which causes serious problems for semantic roles to be a useful classificatory device for predicate-argument relations. These criteria should (a) support the design of a semantic role set which is complete but does not contain redundant relations; (b) be based on semantic rather than morphological, lexical or syntactic properties; and (c) enable formal interpretation.

In the LIRICS¹ project, alternative approaches to the annotation and representation of semantic role information were analysed; methodological principles for characterising well-defined concepts were developed; and a set of semantic roles and their definitions was designed in ISO 12620 format.

This paper is organised as follows. First, we briefly discuss the results of comparative analyses of recent projects concerned with semantic role annotation such as FrameNet (FrameNet, 2005), PropBank (Palmer et al., 2005), VerbNet (Kipper, 2006) and LIRICS (Bunt and Romary, 2002). We then describe annotation experiments carried out in order to evaluate the set of semantic roles proposed in the LIRICS project, and discuss the quantitative and qualitative results. Finally, we point out some interesting issues arising from the annotation and evaluation tasks.

2. Approaches to semantic role annotation

In an early stage of the LIRICS project several approaches and existing projects were analysed and compared with respect to (1) description model; (2) semantic granularity; (3) definitions of semantic roles; and (4) consistency and reliability of annotation.

2.1. Description models

FrameNet (FrameNet, 2005) is designed as an ontology of frames, which are representations of prototypical situations

(events or states). There are more general and more specific situations (e.g. communication events and reporting events respectively). The higher-level frames are considered as characterising the basic structural properties of events and relations in the more specific frames. Each frame provides its set of semantic roles which corresponds to categories of entities or concepts that occur in an event or state. FrameNet has a rich set of relations between frames, e.g. an *is-a* relation between a parent frame and a child frame that implies full inheritance of semantic roles, and where a child frame has at least one difference. This hierarchically structured set of semantic roles could in principle be extended to support annotations that are useful for various applications. In contrast to FrameNet, PropBank (Palmer et al., 2005) and VerbNet (Kipper, 2006) have a verb-dependent model of description of semantic relations. PropBank is a practical approach to semantic annotation, which adds semantic role information to the syntactic structures of the Penn Treebank. The main purpose of PropBank is to provide a description of every verb in the Penn Treebank corpus and to define semantic roles per verb sense based on the number of arguments. Arguments are numbered as Arg0, Arg1, etc. depending on the valency of the verb in question. PropBank's framesets are verb-specific. PropBank limits itself to annotating the literal meaning of a verb². VerbNet (Kipper, 2006) is based on the assumption that syntactic frames associated with a particular verb of a particular class (based on Levin's verb classes) reflect underlying aspects of meaning. VerbNet refined and extended Levin's verb classes, their number growing to 247 classes that cover 5257 verb senses.

Different approaches to semantic role annotation maintain different levels of semantic granularity. VerbNet accounts for high-order generalisations about verb lexical meanings, and defined an exhaustive set of 23 general ("high-level") roles. In addition, there are roles like Theme1 and Theme2, Patient1 and Patient2, which are used for a few classes where there seems to be no distinction between the arguments. FrameNet, by contrast, defines semantic roles rela-

¹Linguistic Infrastructure for Interoperable Resources and Systems (<http://lirics.loria.fr>).

²PropBank covers about 4 659 framesets. FrameNet defines 700 frames.

tive to the frames to which they belong, and are not selected from a pre-defined universal set. Therefore, very different types of semantic roles are defined: from very general to very specific ones. They are linked by relations between the frames they belong to. PropBank has a very fine granularity due to the fact that it distinguishes between the roles of each verb argument. There are 6 role-types (e.g. Arg0 is consistently assigned to an AGENT-type meaning) for core arguments and 11 frame-independent modifier roles (e.g. ArgM LOC: location). Table 1 lists the semantic roles defined within these three projects and shows the role mapping between the compared projects and LIRICS³.

2.2. Semantic role definitions

As for definitions of semantic roles, FrameNet defines semantic roles indeed in a semantic way irrespective of any syntactic information (such as the number of a verb's arguments and their syntactic role in a sentence). However, FrameNet is not fully satisfactory in several respects. There is no consistency in the use of semantic role names where two extremes were observed: the use of 'classical' general roles like *Agent*, and very concept-specific roles, e.g. *Judge* in comparable frames. There is also some inconsistency in the definitions of semantic roles and their defining criteria are quite vague: there often two or more different definitions for one and the same semantic role, e.g. 16 slightly different definitions for *Speaker*. This is due the fact that FrameNet assigns no special significance to the names of frames or the names of the semantic roles; the only important thing is that frame names are unique and conceptually defined, and that semantic roles are defined relative to the frames to which they belong.

In PropBank the semantic role definitions are verb-specific, e.g. for the roset *report.01* the roles are defined as follows: *Arg0*: reporter, *Arg1*: thing reported, *Arg2*: entity reported to. Due to the use of verb-specific roles high annotation consistency is achieved and the tagset was proved to be reliable, Kappa scores of 0.9 measuring the inter-annotator agreement (Palmer et al., 2005). However, the classification of individual verbs into higher level classes as in FrameNet is far from trivial. Serious attempts are made and progress can be noted in establishing a systematic mapping from PropBank semantic roles to FrameNet semantic roles using VerbNet in the SemLink⁴ project (Loper et al., 2007).

The VerbNet role set is very much comparable with the one defined in LIRICS, which has 29 semantic roles. Looking at the semantic role definitions, however, we should notice that VerbNet's roles are not truly semantic concepts; they are partly defined as syntactic or lexical structures and the set does not capture the semantic differences between the roles. For example, VerbNet defines *Agent* as "*generally a human or an animate **subject**, used mostly as a volitional agent, but also used in VerbNet for internally controlled **subject** such as forces and machines*". With the term 'subject' used in the sense of grammatical subject, this defini-

tion automatically excludes passive constructions, e.g. *The tree was hit by **the truck***, where 'the truck' is an internally controlled machine but is not in the subject position. For clarity's sake, we strongly suggest to avoid in the semantic role definitions terms which are not truly semantic. Syntactic, lexical or part-of-speech information could be provided outside the definition in notes, elaborations or annotation guidelines. Another problem with this definition is that it relies on the internal properties of participant (e.g. animacy) rather than describing the way this participant is involved in an event. Surely certain properties of entities enable these entities to play a particular role in an event, e.g. being animate enables a participant to initiate and carry out an event which makes it an *Agent*; however, this property does not necessary make a participant an *Agent*. For example in (1):

(1) *Edison customers receive electric service since April 1985.*

'Edison customers' are animate participants in a 'receiving'-event. We may assume that they act volitionally, as nothing suggest that they were forced to 'receive electric service'. Nevertheless, 'Edison customers' is obviously not the *Agent* but the *Recipient* in this event. Finally, some VerbNet roles seem to be only applicable to certain verb classes. For example, *Experiencer* is used for "*a participant that is aware or experiencing something and used by classes involving psychological verbs, verbs of perception, touch, and verbs involving the body*", and *Stimulus* is "*used by verbs of perception for events or objects that elicit some response from an Experiencer*". This brings redundancy in the defined set of roles, since this information is covered by another, more general role. For example, *Experiencer* is in fact either *Patient* in events, which is "*a participant in an event that undergoes a change of state, location of condition, that is causally involved or directly affected by other participants, and exists independently of the event*" (Schiffrin and Bunt, 2007), e.g. *Mary was surprised by the party*; or else it is *Pivot* in states, which is "*a participant in a state that is characterised as being in a certain position or condition throughout the state, and that has a major or central role or effect in that state*" (Schiffrin and Bunt, 2007), e.g. *I am afraid of spiders*.

Based on the above considerations it was decided for the LIRICS project to define semantic roles:

- as neither syntactic nor lexical structures but as semantic categories;
- by virtue of distinctive semantic properties, since differences between individual roles are semantic;
- that are not restricted to only a few specific verb (noun, adjective) classes;
- not as primitives but rather as relational notions that link participants to an event, and describe the way the participant is involved in an event, rather than by internal properties (e.g. does it act intentionally, is it affected, changed, manipulated by the other participants in an event, does it come into existence through the event, etc.).

³The analyses displayed in this table were made due to the SemLink project (Loper et al., 2007)

⁴For more information and downloads visit <http://verbs.colorado.edu/semLink/>

VerbNet	PropBank	FrameNet	LIRICS
Agent	Arg0, Arg1	Agent, Speaker, Cognizer, Communicator, Ingestor, Deformer, etc.	Agent
Actor	Arg0	Avenger, Communicator, Item, Participants, Partners, Wrongdoer	Agent
Actor1	Arg0	Arguer1, Avenger, Communicator, Interlocutor1, Participant_1, etc.	Agent
Actor2	Arg1, Arg2	Addressee, Arguer2, Injured_Party, Participant2, Partner2	Partner
Attribute	Arg1, Arg2	Attribute, Dimension, Extent, Feature, etc.	Attribute
Beneficiary	Arg1, Arg2, Arg3, Arg4	Audience, Beneficiary, Benefitted_party, Goal, Purpose, Reason, Studio	Beneficiary
Cause	Arg0, Arg1, Arg2, Arg3	Addressee, Agent, Cause, Communicator, etc.	Cause, Reason
Destination	Arg1, Arg2, Arg5	Addressee, Body_part, Context, Goal, etc.	Final_Location
Experiencer	Arg0, Arg1	Cognizer, Experiencer, Perceiver, etc.	Pivot
Extent	Arg2	Difference, Size_change	Amount, Distance
Instrument	Arg2	Agent, Fastener, Heating_instrument, Hot_Cold_source, etc.	Instrument
Location	Arg1, Arg2, Arg3, Arg4, Arg5	Action, Area, Fixed_location, etc.	Location
Material	Arg1, Arg2, Arg3	Components, Ingredients, Initial_entity, Original, Resource, Undergoer	Source
Patient	Arg0, Arg1, Arg2	Addressee, Affliction, Dryee, Employee, Entity, Executed, etc.	Patient
Patient1	Arg0, Arg1	Concept_1, Connector, Fastener, Item, Item_1, Part_1, Whole_patient	Pivot
Patient2	Arg2, Arg3	Concept_2, Containing_object, Item_2, Part_2	Patient
Predicate	Arg1, Arg2	Action, Category, Containing_event, etc.	-
Product	Arg1, Arg2, Arg4	Category, Copy, Created_entity, etc.	Result
Proposition	Arg1, Arg2	Act, Action, Assailant, Attribute, etc.	-
Recipient	Arg1, Arg2, Arg3	Addressee, Audience, Authorities, Recipient	Goal
Stimulus	Arg1	Emotion, Emotional_state, Phenomenon, Text	Theme
Theme	Arg0, Arg1, Arg2	Accused, Action, Co-participant, Co-resident, Content, Cotheme, etc.	Theme
Theme1	Arg0, Arg1	Cause, Container, Phenomenon_1, Profiled_item, Theme	Pivot
Theme2	Arg1, Arg2, Arg3	Containing_object, Contents, Cotheme, etc.	Theme
Time	ArgM_TMP	Time	Time
Topic	Arg1, Arg2	Act, Behavior, Communication, Content, etc.	Theme
Asset	Arg1, Arg3	Asset, Category, Measurement, Result, Value	Amount
Value	Arg1	Measurement, Result, Value, Asset, Category	Amount
Source	Arg2, Arg3	Role, Victim, Patient, Source, Path_start, etc.	Initial_location
-	-	Setting, ContainingEvent	Setting
-	-	Means	Means
-	ArgM_Manner	Manner	Manner
-	ArgM_Purpose	Purpose	Purpose

Table 1: *Semantic roles in different projects.*

LIRICS defines semantic roles as relational notions which link a participant to some real or imagined situation ('event'). For each role we first made a list of entailments associated with each semantic role, starting with the most frequently used ones (e.g. *Agent* and *Theme*), and looked further for non-arbitrary boundaries between roles to design a set which is ideally complete and does not contain redundant relations. These entailments were converted into a set of properties, e.g. [+/- intentionality], [+/- independent existence], etc. Table 2 illustrates the differences between the *Theme* and *Result* roles.

Theme	Result
- intentionality	- intentionality
- affectedness	- affectedness
+ independent existence	- independent existence

Table 2: *Semantic properties for THEME and RESULT roles.*

Thus, *Theme* differs from *Result* in that a *Result* does not exist independently of the event, it is rather the product of the event described by the verb, whereas a *Theme* existed

before the event started, e.g. *Elene read a book* and *Elene wrote a book*.

In this way the set of 29⁵ 'high-level' roles was constructed (Schiffrin and Bunt, 2007).

2.3. Granularity of semantic roles

The LIRICS meta-model (see Figure 1) has two levels of granularity: coarse (high-level) and fine (low-level). For the latter level the FrameNet approach was used, namely the idea of hierarchical structure due to the links to conceptual frames (inheritance relations). A certain low-level semantic role inherits all the properties of the relevant high-level semantic role except for at least one, which would reflect (a) more specific entailment(-s) of a particular predicate or class of predicates. For example, the *Agent* role is defined in LIRICS as:

⁵LIRICS defines 11 roles which are central to any event, e.g. *Agent*, *Theme*, *Patient*, etc., 10 adjunct roles, e.g. *Time*, *Location*, *Manner*, etc., and 8 sub-roles for *Time* and *Location*, e.g. *Duration*, *Frequency*, *Path*, etc. For definitions and illustrative examples of each individual semantic role see (Schiffrin and Bunt, 2007) and (Bunt et al., 2007)

- participant in an event,
- who initiates and carries out the event intentionally or consciously,
- and who exists independently of the event.

For the verbs of communication (communication events) the participant who plays the *Agent* role would be *Communicator* (see (FrameNet, 2005)) and would be defined as:

- participant in an event,
- who initiates and carries out the **communication** event intentionally or consciously **using written, spoken or nonverbal language or combination of those**,
- and who exists independently of the event.

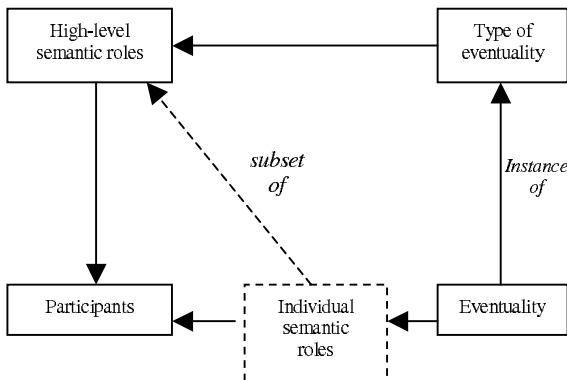


Figure 1: LIRICS metamodel for semantic role annotation.

This shows that the *Communicator* has all the properties of the *Agent* plus what is specific for this class of predicates. If we go one more level down we can define more specific roles, again benefiting from the FrameNet hierarchy. For a particular sub-class of verbs of communication, for example, *Speaker* would be defined as a participant who initiates and carries out the communication event intentionally or consciously **using speech**. Finally, at the verb-specific level *Speaker* could be *Sayer*, *Teller*, *Orator*, *Broadcaster*, etc. Here, the semantic roles defined by PropBank could be used. Figure 2 shows the possible hierarchy according to the model in 1.

2.4. Completeness and redundancy of semantic role set

The LIRICS set of semantic roles was evaluated with respect to redundancy, completeness and reliability (see Section 3). We tested defined semantic roles on redundant information both by looking at annotated data while searching for boundaries between semantic roles to avoid overlapping information and analysing the set of defined properties, eliminating roles with the same properties. This led to removing some roles like *Recipient*, *Stimulus* and *Experiencer*. *Recipient* has the same properties as *Goal*, *Stimulus* overlaps with *Theme*, and *Experiencer* either with *Patient* in events or *Pivot* in states, but the latter roles are more broader concepts and not just restricted to mental, psychological or perception events/states, like *Stimulus* or *Experiencer*. The completeness of the defined set of roles was measured both theoretically by comparing our observations

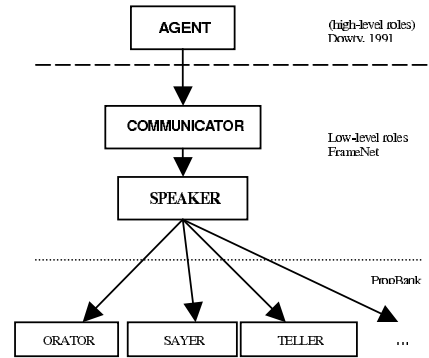


Figure 2: Possible AGENT-roles hierarchy.

with the semantic role sets defined in various other projects (Petukhova et al., 2007) and empirically, as described in Section 3.

It should also be noted here that, once we have analysed the redundancy and completeness of the set of high-level roles, this does not need to be done again for the low-level roles, since the low-level roles inherit the relevant properties from the high-level ones.

3. Evaluation of LIRICS semantic role set

The LIRICS set of semantic role was evaluated for completeness and reliability in terms of inter-annotator agreement. For this purpose multilingual test suites were constructed for English, Dutch, Italian and Spanish. For English FrameNet and PropBank data was used. We selected three unbroken FrameNet texts (120 sentences) and isolated sentences (83 sentences). The PropBank data consists of isolated sentences (355 sentences). For Dutch 15 unbroken texts were selected from news articles, with a total of 260 sentences. News articles were also selected to construct Italian test suites (101 sentences), all taken from the Italian Treebank corpus. For Spanish, the LIRICS test suite consists of 189 sentences taken from the Spanish FrameNet corpus.

3.1. Annotation tasks

The semantic role annotation task involved two main activities:

- Identification and labeling of markables: expressions that represent the entities involved in semantic role relations. Markables come in two varieties:
 - anchors, which correspond to one of three situation (or 'eventuality') types: events, states and facts; every semantic role must be 'anchored' to a situation of one of these types. Anchors are realised mainly by verbs but sometimes by nouns or adverbs.
 - situation participants. They are realised mainly by nouns, noun phrases and pronouns, but also by various types of subordinate clauses.
- Identification and labeling of links: relations between participant and anchor markables.

The annotators were instructed to annotate all possible anchors and related participants including those of subordinate clauses and embedded NP constructions. For example:

(2) [Vicar Marshall_{Agent, e1}; Pivot, s1] admits_{e1} [to mixed feelings_{s1} [about this issue_{Theme, s1}] Theme, e1].

In 2 'Vicar Marshall' is the *Agent* of the 'admitting' event but also the *Pivot* of the 'having mixed feelings' state.

The annotations were made using the GATE annotation tool from the University of Sheffield⁶. GATE provides annotators with a graphical interface for indicating which pieces of text denote relevant concepts (the 'markables').

The annotators were PhD or Master students of linguistics, native speakers of Dutch, Italian and Spanish respectively, and their level of English knowledge was evaluated as proficient. The annotators had little previous experience in annotation and should be considered as 'naive' annotators; they received one afternoon of training in annotation using the LIRICS data categories and the annotation tool. Annotators were provided with Annotation Guidelines for semantic role annotation (Bunt et al., 2007), which contains information on the use of annotation tool, a description of annotation task, examples illustrating the use of data categories, simple decision trees to support choices to be made by annotators, and discussion of some difficult cases.

For Dutch and English all test suite material was annotated independently by at least three different annotators, in order to investigate the usability of the tagset in terms of inter-annotator agreement. Annotations were carried out in two phases: collaborative, where annotators were allowed to discuss their decisions and difficulties; and individual, where annotators made their annotations independently.

3.2. Results

Since the selected test suites were mainly news texts, the results were well comparable for all languages both quantitatively and qualitatively. In order to evaluate the proposed data categories quantitatively we estimated the coverage of defined tags by the annotated corpora.

It may be observed from Table 3 that all the LIRICS semantic roles were covered by the test suites at least for one language. The percentages indicate that their frequencies are comparable for the various corpora. The LIRICS set of semantic role categories can be considered as complete⁷.

To assess the usability and reliability of the defined tagset the inter-annotator agreement was measured in terms of the standard Kappa statistic (Cohen, 1960), the definition of which is based on the probability of inter-annotator agreement, denoted by P(A), and the agreement expected by chance P(E).

The obtained Kappa scores displayed in Table 4 were evaluated according to (Rietveld and van Hout, 1993) and interpreted as annotators having reached *substantial agreement* (scores between 0.61 to 0.8) on two main annotation

⁶See: <http://gate.ac.uk> for further details and <http://gate.ac.uk/documentation.html> for documentation.

⁷For completeness estimations comparing other projects we refer here to (Petukhova et al., 2007).

Data category	English	Dutch	Italian	Spanish
Identified roles	1795	1332	447	1357
/agent/	311 (17.3%)	186 (13.9%)	60 (13.4%)	258 (19%)
/partner/	5(0.3%)	9 (0.7%)	2 (0.4%)	3 (0.2%)
/cause/	39 (2.2%)	33 (2.5%)	2 (0.4%)	43 (3.2%)
/instrument/	10 (0.6%)	7 (0.5%)	7 (1.6%)	4 (0.3%)
/patient/	186 (10.4%)	137 (10.3%)	51 (11.4%)	119 (8.8%)
/pivot/	104 (5.8%)	85 (6.4%)	51 (11.4%)	154 (11.3%)
/theme/	501 (28%)	331 (24.8%)	117 (26.2%)	315 (23.2%)
/beneficiary/	40 (2.2%)	19 (1.4%)	7 (1.6%)	63 (4.6%)
/source/	16 (0.9%)	31 (2.3%)	7 (1.6%)	2 (0.1%)
/goal/	18 (1%)	13 (1%)	13 (2.9%)	5 (0.4%)
/result/	66 (3.7%)	54 (4%)	14 (3.1%)	24 (1.8%)
/reason/	36 (2%)	14 (1%)	9 (2%)	43 (3.2%)
/purpose/	49 (2.7%)	18 (1.4%)	7 (1.6%)	24 (1.8%)
/time/	135 (7.5%)	106 (8%)	13 (3%)	65 (4.8%)
/manner/	39 (2.2%)	33 (2.5%)	18 (4%)	44 (3.2%)
/medium/	4 (0.2%)	1 (0.1%)	2 (0.4%)	8 (0.6%)
/means/	8 (0.4%)	6 (0.5%)	0	2 (0.1%)
/setting/	47 (2.6%)	48 (3.6%)	16 (3.6%)	28 (2%)
/location/	41 (2.3%)	66 (5%)	24 (5.4%)	34 (2.5%)
/initial_location/	2 (0.1%)	1 (0.1%)	2 (0.4%)	5 (0.4%)
/final_location/	6 (0.3%)	10 (0.7%)	7 (1.6%)	43 (3.2%)
/path/	20 (1.1%)	9 (0.7%)	0	0
/distance/	1 (0.1%)	0	1 (0.2%)	0
/amount/	27 (1.5%)	19 (1.4%)	11 (2.5%)	17 (1.3%)
/attribute/	72 (4%)	88 (6.6%)	6 (1.3%)	45 (3.3%)
/frequency/	12 (0.6%)	8 (0.6%)	0	9 (0.7%)

Table 3: Tag occurrences and data categories distribution (in %) across the tested multilingual corpora

tasks: labelling semantic anchors and labelling semantic roles. The annotators exhibited *significant agreement* on the ratings of semantic roles and anchors ($\alpha < .01$).

To reveal and analyse problematic cases and confused categories and/or their definitions in detail we measured the annotators' performance on the individual semantic roles. Table 4 presents the Kappa scores obtained for each defined semantic role as well as disagreement ratio and its source. The averaged Kappa scores presented in Table 4 are obtained from three annotators pairs. All scores indicate that the annotators reached *substantial* (from 0.61 to 0.8) to *perfect* (from 0.81 to 1.00) agreement annotating individual semantic roles except for *Instrument*, where agreement is considered as *fair* (from 0.21 to 0.4), and for *Medium* and *Source*, where agreement is considered as *moderate* (from 0.41 to 0.6) (Rietveld and van Hout, 1993). The *Instrument* role was often confused by annotators with the *Means* role. *Instrument* is distinguished from *Means* by whether it is a participant that exists independent of the event and is manipulated by an agent or not; if it is, then it is an *Instrument*; if not, then it may be a *Means*. *Means* is defined as a procedure or method by which the event takes place, for example:

(3) *The far left had some good issues even if it did not have good programs for dealing with them.*

The NP 'good programs' was annotated by one annotator as *Instrument* of the event 'dealing', by another annotator

Task	Kappa	Disagreement ratio	Cases of confusion:
Semantic anchors	0.77	0.15	state vs event
Semantic roles	0.68	0.25	Agent vs Cause, Attribute vs Manner, Beneficiary vs Goal, Instrument vs Means, Purpose vs Reason, Theme vs Result, Location vs Setting, Theme vs Pivot, Theme vs Patient
Semantic Role	Kappa	Disagreement ratio	Confused with:
Agent	0.87	0.1	Theme; Pivot; Patient; Cause
Amount	0.77	0.2	Instrument; Source; Manner
Attribute	0.71	0.29	Theme; Manner; Result; Setting
Beneficiary	0.81	0.19	Patient; Goal; Theme
Cause	0.64	0.36	Agent; Theme; Patient
Final_Location	0.98	0.02	Setting
Frequency	0.94	0.06	Amount; Attribute
Goal	0.64	0.36	Beneficiary; Theme; Result
Instrument	0.3	0.72	Patient; Means
Initial_Location	0.9	0.1	Setting
Location	0.92	0.08	Setting
Manner	0.89	0.11	Attribute; Setting
Means	0.57	0.43	Patient; Manner; Instrument
Medium	0.76	0.24	Patient; Source; Setting
Partner	0.8	0.19	Patient; Theme
Path	0.76	0.23	Goal; Result
Patient	0.73	0.25	Theme; Result; Instrument; Agent
Pivot	0.65	0.33	Theme; Agent; Patient
Purpose	0.76	0.23	Theme; Reason
Reason	0.81	0.19	Theme; Purpose
Result	0.77	0.22	Theme; Patient; Goal
Setting	0.68	0.32	Manner; Location; Attribute
Source	0.52	0.48	Reason; Setting; Agent
Theme	0.67	0.28	Pivot; Result; Patient
Time	0.99	0.01	Manner; Setting; Theme
Distance	1.00	0	

Table 4: Inter-annotator agreement on semantic anchors and (individual) roles expressed in Kappa scores and ratio and cases of disagreement

as *Means*, and by the third annotator was not identified as a participant of the 'dealing' event. For *Instrument* some inconsistency in identification of this type of participants was observed between annotators; while Annotator 1 identified 11 participants with *Instrument* role and Annotator 2 identified 10 of those, reaching an agreement of 84%, Annotator 3 identified zero participants with the *Instrument* role and reached zero agreement with both other annotators.

The role *Source* was frequently confused with *Reason*:

- (4) *His doubts stemmed from the fact that several years earlier a Princeton University researcher, Arnold Levine, had found in experiments with mice that a gene called p53 could transform normal cells into cancerous ones.*

In this case two annotators assigned the *Source* role to the participant marked in bold and one annotator assigned the role *Reason*. We may assume that 'the fact...' is the *Reason* of 'his doubts', but the *Source* of the 'stemming' event, because *Reason* represents the set of facts or circumstances explaining why a state exists or an event occurs. *Source*, by contrast, is a participant in an event that is the non-locative and non-temporal start point of an action.

Spatial and temporal roles (*Location* and *Time*, and their

sub-roles) were easier to identify than others. These roles are usually less ambiguous, but some confusing cases do occur, for example, for *Location* vs *Setting*. *Setting* is distinguished from *Location* by whether it defines a set of circumstances of the occurrence of event or state, or not; if it does, then it is a *Setting*; if not, then it is a *Location*. *Location* is a participant that represents the place where an event occurs, or a state that is true. For example 5:

- (5) *It hopes to speak to students at theological colleges about the joys of bell ringing.*

The participant 'theological colleges' in 5 is ambiguous and can refer to a building, a school for advanced education, an organization, or students and teachers of these organisation. Some situations are ambiguous, e.g. *Reason* vs *Purpose*:

- (6) *Laws exist to prevent crimes.*

In this particular case it is not entirely clear without context whether 'preventing crimes' is a *Reason* of 'laws existence' or a *Purpose*.

LIRICS defines semantic roles as a way a participant takes part in an event, and a participant's involvement is potentially manifold. Thus, a participant may have one or more semantic roles associated with an event. For example, for verbs like 'pay', 'supply' and 'provide', a participant, who receives something may have two roles, namely *Beneficiary*, but also *Goal*, for example:

(7) *Germany and China allegedly provided technical and material assistance to the Al-Fatah program.*

The participant *the Al-Fatah program* is clearly advantaged by the event (*Beneficiary*) and it also describes a terminal point which will be reached in the normal course of events or in all possible courses of events (*Goal*).

Overall, the results are encouraging and promising, considering the fact that annotations were made by 'naive' annotators with little experience in annotation work and very limited training. After a close inspection of the results, we concluded that some moderate Kappa values were mainly due to the fact that the annotation guidelines were not yet well-established. As an outcome of the results described above both the annotation guidelines and some of the semantic role definitions were improved.

4. Conclusions and future research

In conclusion we would like to highlight some benefits of the LIRICS description model and semantic role set. The LIRICS model incorporates important findings of other projects in the same area and makes a step forward by providing a complete set of semantic roles without redundancies, defined as purely semantic concepts by virtue of distinctive semantic properties. The LIRICS model encompasses different levels of granularity enabling hierarchical structures of semantic roles, making this model extendable and attractive for many applications. Finally, the LIRICS semantic role set can be used reliably for annotation purposes. It was established that annotators exhibit substantial agreement using the proposed data categories. Those categories which were frequently confused by the annotators underwent some revision. For example, the definitions of *Instrument* and *Means* were revised and the distinction between them was clarified in the property of *independent existence*, where the *Instrument* does exist independently from the event, whereas *Means* is a participant in an event that represents a procedure for performing the action in terms of component steps, or a method by which an intentional act is performed by an agent, and does not necessarily exist independently of the event.

In the future, more effective annotation guidelines will be designed where roles are organised in a taxonomy exploiting semantic features, allowing annotators to deal with different levels of granularity and perform a case-by-case decision. Finally, we aim to support annotators by incorporating other resources, such as the VerbNet index and Sem-Link, and provide systematic mappings of roles defined within other projects to those defined in LIRICS.

5. References

- H. C. Bunt and L. Romary. 2002. Requirements on multimodal semantic representations. *Proceedings of ISO TC37/SC4 Preliminary Meeting*, pages 59–68, Dordrecht, Seoul.
- H. C. Bunt and V. Petukhova and A. Schiffrin. 2007. LIRICS Deliverable D4.4. Multilingual test suites for semantically annotated data. <http://lirics.loria.fr>.
- J. Cohen. 1960. A coefficient of agreement for nominal scales. *Education and Psychological Measurement*, 20:37-46.
- D. Dowty. 1991. Thematic Proto-Roles and Argument Selection. *Language*, 67:547–619.
- ICSI. 2005. FrameNet. <http://framenet.icsi.berkeley.edu>.
- P. Kingsbury and M. Palmer and M. Marcus. 2002. Adding Semantic Annotation to the Penn TreeBank. *Proceedings of the Human Language Technology Conference*, San Diego, California.
- K. Kipper. 2002. VerbNet: A Class-Based Verb Lexicon. <http://verbs.colorado.edu/mpalmer/projects/verbnet.html>.
- E. Loper and Szu-ting Yi and M. Palmer. 2007. Combining Lexical Resources: Mapping Between PropBank and VerbNet. *Proceedings of the Seventh International Workshop on Computational Semantics (IWCS-7)*, pages 118–129.
- M. Palmer and D. Gildea and P. Kingsbury. 2002. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71-106.
- V. Petukhova and A. Schiffrin and H. Bunt. 2007. Defining Semantic Roles. *Proceedings of the Seventh International Workshop on Computational Semantics (IWCS-7)*, pages: 362–365, Tilburg.
- T. Rietveld and R. van Hout. 1993. *Statistical techniques for the study of language and language behavior*. Berlin: Mouton de Gruyter.
- A. Schiffrin and H. C. Bunt. 2007. LIRICS Deliverable D4.3. Documented compilation of semantic data categories. <http://lirics.loria.fr>.