

Enriching the ISST-TANL Corpus with Semantic Frames

Alessandro Lenci*, Simonetta Montemagni†, Giulia Venturi◇, Maria Rosaria Cutrullà*

* Department of Linguistics, Università di Pisa (Italy)

† Istituto di Linguistica Computazionale “Antonio Zampolli”, Pisa (Italy)

◇ Scuola Superiore Sant’Anna di Studi Universitari e di Perfezionamento, Pisa (Italy)

lenci@unipi.it, simonetta.montemagni@ilc.cnr.it, giulia.venturi@ilc.cnr.it, kena1987@hotmail.com

Abstract

The paper describes the design and the results of a manual annotation methodology devoted to enrich the ISST-TANL Corpus with Semantic Frames information. The main issues encountered in applying the English FrameNet annotation criteria to a corpus of Italian language are discussed together with the choice of anchoring the semantic annotation layer to the underlying dependency syntactic structure. We also describe an experiment to measure inter-annotator agreement and a first case study to extend and specialise FrameNet annotation to a corpus of legislative texts.

Keywords: Semantic annotation, FrameNet, Multi-layer annotated corpus

1. Introduction

Recently, a number of projects have been focused on the creation of FrameNet resources for new languages, and so far two main directions have been followed. Some research groups are mainly working at the manual annotation of FrameNet-like databases, as is the case with Spanish (Subirats, 2009) and Japanese (Ohara, 2008). Other teams, instead, have been investigating ways to automatize and speed-up the database development, mainly by exploiting existing annotated resources for English: such approaches have been applied, for example, to Swedish (Johansson and Nugues, 2006) and Chinese (Chen and Fung, 2004). For German, both approaches were tested, namely the NEGRA corpus was manually annotated with frame information and the stochastic tool for automatic annotation called *Shalmaneser* was developed (Erk and Pado, 2006). As for Italian, small projects aimed at investigating the applicability of the FrameNet paradigm started independently in several research centers. Given the specific expertise and research goals of each group, different approaches have been adopted so far. It is the reason why since 2010 the *iFramenet* national project has started, aiming at releasing a unified Italian FrameNet resource. It is a joint initiative devoted to coordinating ongoing efforts to develop lexical and annotated resources for Italian based on frame semantics. In the project Wiki portal¹, the different groups can share the annotated corpora and information about news, meetings, resources etc. are collected.

In this paper, we present the work that has been jointly carried out at the University of Pisa (Department of Linguistics) and at the Institute of Computational Linguistics “Antonio Zampolli” (ILC-CNR) where a frame-based annotation has been added to an already existing treebank, ISST-TANL. ISST-TANL is a dependency annotated corpus originating as a revision of the ISST-CoNLL corpus (Montemagni and Simi, 2007), in turn derived from the Italian Syntactic-Semantic Treebank or ISST (Montemagni et al., 2003). ISST-TANL includes 3,109 sentences (71,285

tokens) and consists of articles from newspapers and periodicals. In addition, we report about a recent case study devoted to extending the adopted frame annotation methodology to corpora of legislative texts.

The paper is organized as follows: Section 2. shortly reports the basics of Berkeley FrameNet project; in Section 3., the adopted frame annotation methodology is presented with particular attention to *i*) the anchoring of frames to syntactic dependency structures, *ii*) language-specific issues and *iii*) the annotation of figurative language instances, such as metaphors; in Section 4., details about the annotation process are reported as well as the quality control results; Section 5. illustrates the results of a case study aimed at extending the annotation methodology to deal with legislative texts. Finally, in Section 6., current research directions are illustrated, including the exploitation of the ISST-TANL Corpus enriched with Semantic Frames for Natural Language Processing tasks.

2. The FrameNet project

The FrameNet resource² is a lexical resource for English, based on *Frame Semantics* (Fillmore, 1985) and supported by corpus-evidence. The goal of the FrameNet project is to document the range of semantic and syntactic combinatory possibilities of each word in each of its senses. Typically, each sense of a word is associated with a Semantic Frame, conceived in Ruppenhofer et al. (2010) as “a script-like conceptual structure that describes a particular type of situation, object or event along with its participants and properties”. For example, the APPLY_HEAT frame describes a common situation involving participants such as COOK and FOOD, called Frame Elements (FEs), and is evoked by Lexical Units (LUs) such *bake*, *blanch*, *boil*, *broil*, *brown*, *simmer*, etc. As shown by the following example, the frame-evoking LU can be a verb (bolded in the example) and its syntactic dependents (those between square brackets and whose semantic role is specified in the subscript) are its FEs:

¹<http://sag.art.uniroma2.it/iframe/doku.php>

²<https://framenet.icsi.berkeley.edu/fndrupal/>

- [*Matilde* _{Cook}]**fried** [*the catfish* _{Food}] [*in a heavy iron skillet* _{Heating-instrument}].

The type of representation produced by FrameNet is a network of “situation–types” (frames) organized across inheritance relations between frames (representing the so–called *frame–to–frame relations*, i.e. ‘Inheritance’, ‘Subframe’, ‘Causative of’, ‘Inchoative of’ and ‘Using’), as opposed to a network of meaning nodes, as in the case of WordNet (Fellbaum, 1998). The design of these relations is overtly aimed at keeping robust the frame network. As claimed by Ruppenhofer et al. (2010), “having relations to semantically similar frames allows frames (and thus their Lexical Units) to be associated despite being separated”.

In FrameNet, FEs can also be specified with Semantic Types (i.e. ontological categories) specifying the semantic classes of their fillers. Most of these semantic types correspond directly to “synset nodes” of WordNet, and can be mapped onto already existing ontologies. FrameNet currently contains more than 1,123 frames, covering 12,280 Lexical Units; these are supported by more than 188,778 FrameNet–annotated example sentences.

3. Frame annotation of the ISST–TANL treebank

Our final goal is the creation of an annotated corpus by adding a layer of FrameNet frames and frame elements to the syntactic structures of the ISST–TANL Treebank. The corpus has been designed with a specific view to the development of methods and techniques for automatic semantic role labelling: this represents a qualifying feature of our approach to frame annotation, which influenced many of the annotation choices discussed in the following sections. A similar approach to semantic annotation is reported, for instance, for the construction of the PropBank (Palmer et al., 2005) for English as well as of the SALSA corpus for German (Burchardt et al., 2006), with a main difference: whereas in the PropBank and SALSA corpora semantic frames are anchored to constituency–based syntactic representations, in our case the underlying syntactic annotation layer is dependency–based.

In this section we summarise the main criteria which guided the annotation process of the ISST–TANL Treebank with frame information. In Section 3.1. we discuss the criteria which were defined for anchoring semantic frames to syntactic dependency structures, whereas Sections 3.2. and 3.3. discuss respectively the approach followed for dealing with Italian–specific peculiarities and with figurative usages.

3.1. Anchoring frame annotation to the underlying syntactic structure

As pointed out above, our semantic frame information is annotated on top of linguistic annotations that cover morpho–syntax and dependency syntax. Among the advantages of such a choice it is worth mentioning here the fact that previous levels of annotation can drive the annotation process thus resulting in an increase in efficiency and quality for any new annotation.

In anchoring frame information to syntactic dependencies, we followed the “syntactic locality principle” proposed by

Ruppenhofer et al. (2010) for FrameNet annotation according to which all FEs, with the only exception of subjects, should be realized by constituents that are part of the maximal phrase headed by the frame–evoking word. However, for FrameNet Ruppenhofer et al. (2010) foresee two types of situations in which non–local constituents are annotated as frame elements and motivate such a choice on lexicographic grounds, due to the fact that these non–local constituents contain valuable information about the semantic type of the frame element which cannot be recovered from the co–indexed locally occurring constituents or from empty elements. More specifically, non–local constituents are annotated as frame elements when the frame–evoking word is syntactically governed by a raising or control predicate (case 1) or it occurs inside a relative clause (case 2). Although this choice is fully justified from the lexicographic point of view, if looked at from a different perspective it may turn out to be problematic: as Dolbey (2009) pointed out, this annotation strategy may cause “difficulties for end users who want to perform automatic processing that includes information from FrameNet’s annotation collection”. In order to overcome this type of problem in exploiting frame annotated resources for semantic role labelling, in our corpus the syntactic locality constraint is always enforced.

Consider the annotation example reported in Figure 1, where the KILLING frame has been annotated on top of the syntactic dependency tree of the sentence *Un fulmine ha ucciso ieri tre alpinisti sul Sassolungo, in Alto Adige* (‘A lightning killed three climbers on Sassolungo yesterday, in South Tyrol’). It can be noticed that FEs (namely, CAUSE, TIME, VICTIM and PLACE) do coincide here with dependency subtrees governed by the immediate dependents of the frame–evoking word (*uccidere* ‘kill’).

In defining the criteria for anchoring frame annotation to syntactic dependencies, another important issue had to be tackled, concerning the selection of the appropriate text span to be marked as the instantiation of a given frame element. As far as FrameNet is concerned, Ruppenhofer et al. (2010) state what follows: “we tag whole constituents that realize frame elements relative to our target words, rather than just tagging the head words of these constituents”. We proceed similarly though in a dependency–based framework, i.e. the text span of FE instantiations corresponds to the text span covered by the subtree headed by the dependent of the frame–evoking word. In the example in Figure 1, the textual spans associated with the CAUSE, TIME, VICTIM and PLACE FEs are respectively *Un fulmine* (‘A lightning’), *ieri* (‘yesterday’), *tre alpinisti* (‘three climbers’) and *sul Sassolungo, in Alto Adige* (‘on Sassolungo, in South Tyrol’). The motivation underlying this choice was, again, to produce a consistent and “cleanly” annotated corpus, thus enhancing its potential for machine–learning tasks.

Let us consider now a typical context for which the syntactic locality constraint is relaxed in FrameNet, i.e. when the frame–evoking word is the head of a relative clause. Figure 2 shows the annotation performed for the frame USING occurring inside a relative clause within the sentence *Alcuni poliziotti, presi di mira come bersaglio dai “punk tiratori scelti”, che usavano fionde di precisione, hanno risposto*

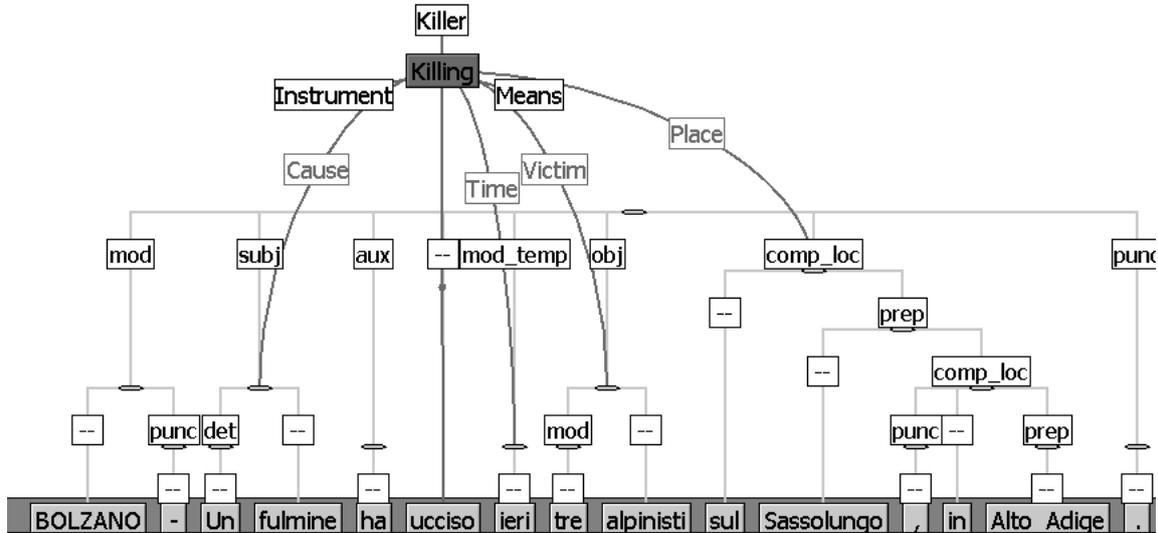


Figure 1: A sentence of ISST-TANL annotated with semantic frames using the SALTO tool.

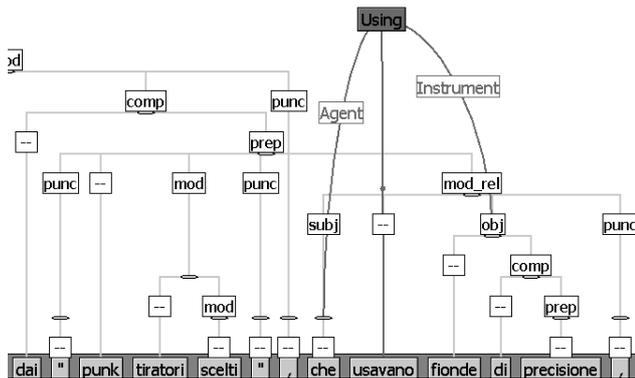


Figure 2: Frame annotation of a relative clause.

rilanciando al mittente i sampietrini (‘Some police officers, targeted by “punk snipers” who used precision slingshots, have responded by re-launching the cobblestones to the sender’). It can be noticed that, by enforcing the locality constraint, all identified FEs represent local dependencies, also in the case of the relative pronoun, which is the filler of the AGENT role. However, if on the one hand the relative pronoun as such is a semantically empty element, on the other hand semantic information about the frame element corresponding to the relative pronoun can be recovered by exploiting the mapping between syntactic and semantic annotations, in particular through the dependency (*mod_rel*) linking the head of the relative clause to its antecedent.

This example shows that, within a multi-layered annotation scheme, frames characterised by FEs that are either semantically empty or have a null instantiation are no longer a problem, since the underlying semantics can be recovered starting from the syntax-semantic mappings. Although this can be seen as requiring further efforts for recovering this type of information, we should consider that in this way the prerequisites are created for exploiting this frame anno-

tated corpus for machine-learning purposes.

3.2. Italian-specific issues

In applying FrameNet frames to Italian texts, some language-specific issues had to be addressed. This is the case, for instance, of null subjects which in Italian occur quite frequently, being Italian a pro-drop language. Null subjects belong to the class of implicit constructionally licensed omissions for which the standard treatment described in Ruppenhofer et al. (2010) consists in introducing a frame element indicating which semantic role the missing element would fill, if it were present. The same strategy were adopted to deal with null subjects in the Spanish FrameNet (Subirats, 2009). Our solution for the treatment of null subjects departs from standard FrameNet. Figure 3 exemplifies the representation of a frame element corresponding to a null subject for the frame BEING_EMPLOYED within the sentence *Erano in gita e lavoravano come volontari per costruire una centrale* (‘They were on a trip and they worked as volunteers to build a power plant’). It can be seen that the EMPLOYEE frame element is linked to the frame-evoking word *lavoravano* since its associated inflectional features (in particular, person and number) convey relevant information about the omitted subject.

3.3. Annotation of metaphors

As described in Ruppenhofer et al. (2010), the FrameNet treatment of metaphors makes a distinction between “productive” and “lexicalized” metaphors, indicated by whether annotation is done with respect to the source domain of a metaphor (i.e. the literal frame) or with respect to the target domain (i.e. the frame that more directly encodes what the speaker was trying to say) respectively; in the case of productive metaphors, a specific tag is added to the annotated frame. In SALSA (Burchardt et al., 2006) a different annotation strategy is pursued, according to which both source and target frames are annotated with respect to the metaphoric frame instance due to the complementary in-

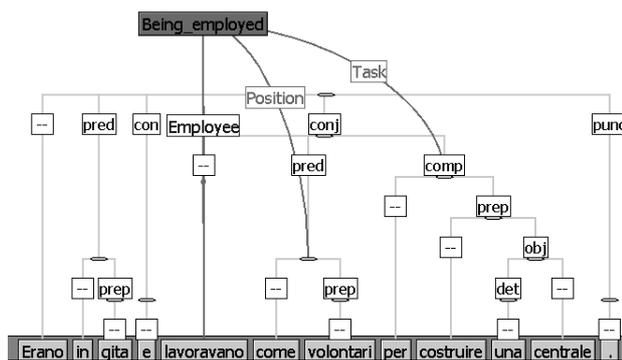


Figure 3: Frame annotation with null subject.

formation captured by them: “the source frame models the syntactic realization patterns of arguments, while the target frame captures the understood meaning”.

In the annotation of metaphors, we departed from both approaches, and we have always marked the metaphoric context as an instance of the literal frame. The latter has then been marked with a specific tag aimed at recovering all metaphoric usages within the annotated corpus. The reasons underlying this choice are of different nature. First, as pointed out by Burchardt et al. (2006), the literal frame represents the most suitable choice as far as the anchoring to the underlying dependency structure is concerned. Second, due to the difficulty of discriminating between “productive” and “lexicalized” metaphors during the annotation process we decided to assign the literal frame and postpone the marking of such a distinction to a post-annotation phase based on the quantitative and qualitative analysis of all metaphoric usages encoded in the corpus.

4. The annotation process

Frame annotation of the ISST-TANL Treebank was carried out manually with the SALTO tool (Erk et al., 2003). The treebank was first automatically converted into the TIGER/SALSA XML format, and then loaded onto SALTO, together with the ontology of semantics frames and FEs derived from the Berkeley FrameNet. The conversion process had to cope with the fact that, differently from the SALSA project (Burchardt et al., 2006), here the annotation of frame-semantic information had to be carried out on top of dependency-based syntactic representations. Annotation has been performed according to the so-called FrameNet lexicographic mode. We sampled a set of medium frequency verbs and for each of them we extracted the sentences in the treebank in which these verbs occurred. The task of the annotators was to *i*) identify the frame evoked by the target LU in the sentence, *ii*) annotate the FEs that were overtly realized in the sentence.

Currently, we annotated 1,916 sentences from the ISST-TANL corpus, for a total number of 2,934 frame instances and 287 frame types. Three annotators with undergraduate training in linguistics participated in the annotation process independently. In a subsequent adjudication step, cases of disagreement between annotators have been resolved manually.

Annotation is still ongoing. In particular, it is currently focussed on increasing the ISST-TANL semantic annotation coverage since, for the time being, 246 frame types out of 287 annotated ones still have a low number (i.e. ≤ 10) of instances. This entails that about 85% of the annotated frame types are scarcely instantiated: this motivates the current effort to enlarge the number of annotated frame instances in the corpus. In addition, up to now, annotation mainly focused on verb LUs, with a small amount of nouns that have also been annotated. Current annotation efforts are thus aimed at also increasing frame annotation with respect to nouns.

Last but not least, in order to measure the annotation reliability, a sample of corpus sentences were annotated by all the three annotators and the inter-annotator agreement was computed, as described below.

4.1. Inter-annotator agreement

The inter-annotator agreement was computed on the basis of a subset of 420 sentences which were independently annotated by the three annotators. 106 different frame types were involved for a total number of 1,606 frame instances. Two different cases were considered: *i*) the three annotators have assigned the same frame to a given frame-evoking LU and *ii*) different frames have been assigned to a given LU. According to this strategy, we were able to compute that the annotators agreed on 208 sentences whereas they disagreed on 212 sentences. This entails that 49.53% of sentences were annotated in the same way by the three annotators while 50.47% of sentence were annotated by resorting to different frames.

In order to provide a unique value of the inter-annotator agreement, we calculated the kappa coefficient (Siegel and Castellan, 1988; Artstein and Poesio, 2008) on the sample of 420 sentences annotated by the three annotators. Even though, on the one hand, it is one of the most widely used measure to evaluate the reliability of annotations, on the other hand, it is also acknowledged that it is not fully appropriate for some tasks. As discussed in Burchardt et al. (2006), “kappa assumes a very restricted annotation process, in which a single label is chosen from a globally fixed pool for each annotated instance”. In other words, the kappa statistic assumes that annotation categories are discrete and mutually exclusive. However, this is not the case for semantic frames, some of which are semantically related thanks to their hierarchical organization designed in FrameNet. This is the case of those frames which are linked by frame-to-frame relations or which describe situation-types which are semantically similar. They can lead to misleading frame assignments, making the annotation process more subjective.

FrameNet frames greatly differ with respect to their granularity as well as for their semantic distance. In fact, in many cases there are multiple, very close frames that cover similar semantic areas, and therefore “compete” with each other, when a specific annotation decision must be taken. In such cases, inconsistent annotations are very likely. For example, the following ISST-TANL sentence was annotated differently by the three annotators:

- È improbabile che gli abitanti di Knin, sui quali

prima della resa pioveva un proiettile d'artiglieria ogni dieci secondi, abbiano ricordato l'anniversario dell'atomica di Hiroshima. ('It's unlikely that the inhabitants of Knin, on which before the surrender an artillery shell rained every ten seconds, commemorated the anniversary of the atomic bombing of Hiroshima'.)

The first annotator interpreted the LU *ricordato* ('commemorated') as an instance of the MEMORY frame, the second of the REMEMBERING_EXPERIENCE frame and the third of the REMEMBERING_INFORMATION frame. All the annotators found that the verb *ricordare* ('to commemorate') evokes a situation-type where a cognizer remembers a mental content whereas such a situation is differently described in the three suggested frames. Even though the three frames are semantically related, they depict the given situation in different manners. While the MEMORY and REMEMBERING_EXPERIENCE frames describe a situation where a cognizer calls up an episodic memory of past experience formed on the basis of past personal experience, the REMEMBERING_INFORMATION frame describes a more impersonal and more general situation where a "cognizer retains facts in memory and is able to retrieve them". In addition, it should be noted that the REMEMBERING_EXPERIENCE and REMEMBERING_INFORMATION frames are also linked by the frame-to-frame relation 'See_also'.

In order to empirically investigate how the frame hierarchic organization affected the annotation effort, we devised a new way to compute the inter-annotator agreement for the frame assignment task. To our knowledge, this is the first time that such a strategy has been exploited in a FrameNet-based annotation scenario.

Firstly, we computed the kappa coefficient with respect to the whole set of 106 frames. Then, we recalculated the agreement with respect to *super frames* which we built on the basis of the FrameNet hierarchy. This is to say that, in a first stage, we calculated the three annotators' consensus on the annotation of the MOTION, BODY_MOVEMENT, BRINGING, MOTION_DIRECTIONAL, PLACING, etc. single frames; in a second stage, we computed the agreement rate for a unique *super frame* MOTION linked by frame-to-frame relations to a number of different *child frames* such as BODY_MOVEMENT, BRINGING, MOTION_DIRECTIONAL, PLACING, etc. The list of the considered *child frames* which are included in the class of the *super frame* MOTION is reported in Table 1.

It should be noted that the construction of classes of frame types (so-called *super frames*) was carried out by climbing up the frame hierarchy up to 4 levels. This is to say that incrementally built *super frames*, initially including only directly linked frames, ended up with lumping together frames related via two intermediate frames: for example, the path linking the MAKING_FACES frame to the MOTION frame includes two intermediate frames, i.e. FACIAL_EXPRESSION and BODY_MOVEMENT.

As could be expected, the kappa coefficient significantly increased when it was calculated with respect to the *super frames* rather than to single frame types. We computed the kappa coefficient with respect to three configurations: *i*)

<i>Child frames</i>
BODY_MOVEMENT
BRINGING
MOTION_DIRECTIONAL
PLACING
SELF_MOTION
REDIRECTING
TRAVERSING
DEPARTING
ARRIVING
INGESTION
MAKING_FACES

Table 1: The list of *child frames* which are included in the class represented by the *super frame* MOTION.

single frame types (i.e. 106 frame types), *ii*) classes of *child frames* linked to the corresponding *super frame* through 1 level of frame-to-frame relations (i.e. 15 classes of *super frames*) and *iii*) classes of *child frames* linked to the corresponding *super frame* through up to 4 levels of frame-to-frame relations (i.e. 15 classes of *super frames*). Table 2 reports such an increment.

K value	No. of frame types	Level of FN hierarchy
0.45	106	–
0.60	15	1
0.61	15	3/4

Table 2: Increment of kappa coefficient according to the classes of frame types.

Interestingly, it can be noted that, according to the guidelines provided by Landis and Koch (1977), the agreement goes from *moderate* (0.45) to *substantial* (0.61) when we considered *super frame* classes also including frames linked through up to 4 levels of frame-to-frame relations. As mentioned above, this follows from the specific frame annotation task where the categories to be assigned are not mutually exclusive but are rather organized in a structured hierarchy of frames that can also be semantically very similar.

5. Extending the ISST-TANL treebank to the legislative domain: a case study

By starting from the suggestion expounded in Dolbey et al. (2006) that FrameNet can be seen "as a backbone of several domain-specific FrameNets", we are currently carrying on a case study aimed at building a FrameNet-based semantically annotated corpus of Italian legislative texts. Following the same annotation methodology adopted for the ISST-TANL corpus, the main aim is to show that a FrameNet approach can be suitable for insightful analyses of general language, as well as an important starting point to describe the syntactic and semantic combinatorial possibilities exhibited in a specialized language such as the legal one.

The interest in pursuing this goal stems from the need for making explicit domain-specific content starting from its

surface realization in unrestricted texts, as widely acknowledged within both Natural Language Processing (NLP) and Artificial Intelligence and Law (AI&Law) communities. In particular, such an approach is meant to ground domain-specific knowledge management tasks (such as Legal information extraction, Court decision structuring, Legal argumentation mining, etc.) on linguistically-driven semantic annotation of legal texts.

Similarly to the general language corpus, the annotation has been carried out on top of dependency-annotated legislative texts. Due to the lack of a gold dependency-annotated corpus for this domain, we used the output of the DeSR dependency parser (Attardi and Dell’Orletta, 2009), which was manually revised before the frame-semantic annotation phase, in order to correct erroneous annotations, mainly due to syntactic peculiarities of legislative texts. As described in Venturi (2011), a number of both legal language description issues and legal knowledge representation issues had to be tackled. In particular, even though the annotation methodology mostly consisted in maintaining and reusing the semantic frames and FEs already defined in FrameNet, several domain-specific customizations turned out to be needed. They mainly involved:

1. the introduction of one or more FEs within an existing frame. This happened when FrameNet did not foresee that an important piece of information was part of the background knowledge evoked by a predicative lexical unit. For example, FrameNet did not include a “Purpose” FE in the BEING_OBLIGATED frame, even though this piece of information is needed to fully describe the semantics conveyed by this frame, as shown in the following annotated sentence:

- [*Per la realizzazione delle opere previste nelle convenzioni già assentite alla data del 30 giugno 2002, ovvero rinnovate e prorogate ai sensi della legislazione vigente* *Purpose*] [*i concessionari Responsible_party*] sono [**tenuti** *TARGET*] [*ad appaltare a terzi una percentuale minima del 40 per cento dei lavori, Duty*] [*applicando le disposizioni della presente legge ad esclusione degli articoli 7, 14, 19, commi 2 e 2-bis, 27, 32, 33 Condition*]. (Lit. [For the realization of works planned in the conventions already assented on the date of the 30th June 2002, that is renewed and extended under the in force law *Purpose*] [the agents *Responsible_party*] are [**bound** *TARGET*] [to contract out to third party a percentage minimal of the 40% of works, *Duty*] [enforcing the provisions of the present law with the exception of articles 7, 14, 19, paragraphs 2 and 2-bis, 27, 32, 33 *Condition*].)

This sentence demonstrates that to fully characterize the BEING_OBLIGATED frame for the legal domain it is necessary to account for the particular scope that can be achieved if the “Responsible_party” performs a “Duty” (i.e. the “Purpose”);

2. the specification of domain-specific semantic types in order to classify FEs. This is done by adding seman-

tic types taken from an existing legal ontology, when no proper semantic type is available in FrameNet. For example, in the BEING_OBLIGATED frame neither the FE “Duty” nor “Responsible_party” were assigned any semantic type. Therefore, for these FEs the domain-specific customization included the typing with the semantic type ‘Duty’ and ‘Legal Subject’ respectively, two classes (i.e. two juridical concepts) which were taken from the Core Legal Ontology (CLO) (Gangemi et al., 2005);

3. the creation of new semantic frame(s). This represents the most controversial kind of customization. As (Dolbey et al., 2006) warns, on the one hand, the introduction of a new frame to specify domain-specific information would result in a richer representation of domain-specific semantics; on the other hand, there would be an increase in the complexity of the network of frames. For example, a new GRANT_LEGAL_PERMISSION frame was added in order to characterize a situation-type where an authority grants a permission to a grantee. In FrameNet there are two different frames that may evoke such a situation: PERMITTING and GRANT_PERMISSION. The first one describes a situation where a “State of Affairs is permitted by a Principle”; the second one represents a situation where “a Grantor (either a person or an institution) grants permission for a Grantee to perform an Action”. However, the latter frame, according to FrameNet’s definition, “does not include situations where there is a state of permission granted by authority or rule of law”. The new suggested frame inherits some of the FEs of the GRANT_PERMISSION frame with a number of domain-specific customizations³. Thanks to this newly introduced frame, it is thus possible to properly represent the legal content of the following sentence:

- [*Il Ministero della sanità Legal_grantor*], *per quanto riguarda gli aspetti ambientali d’intesa con il Ministero dell’ambiente*, [**autorizza** *TARGET*] [*ai sensi del presente decreto Circumstances*] [*l’immissione sul mercato e l’utilizzazione nel territorio italiano di un biocida Permitted_action*]. (Lit. [The Ministry of Health *Legal_grantor*], regarding the environmental aspects according to the Ministry of Environment [**authorizes** *TARGET*] [under this decree *Circumstances*] [the placing on the market and the usage in Italian territory of a biocidal *Permitted_action*].)

³The foreseen customizations are the following (on the left of the < the FEs of the new GRANT_LEGAL_PERMISSION frame, on the right the corresponding FEs of the GRANT_PERMISSION frame already existing in FrameNet):

- ‘Legal grantor’ < ‘Grantor’,
- ‘Grantee’ < ‘Grantee’,
- ‘Permitted_action’ < ‘Action’.

The case study carried out so far resulted in the annotation of 226 sentences of a sub-set of the TEMIS corpus (SynTactically and SEMantically Annotated Italian Legislative Corpus; Venturi (2012)), a corpus of Italian legislative texts enriched with two different layers of linguistic annotation, dependency syntax and frame semantics. In particular, the semantic annotation effort has been devoted to making explicit how the three main deontic modalities, i.e. *obligation*, *permission*, *prohibition*, are linguistically realized in the TEMIS corpus.

6. Conclusion

In this paper we described the design and the current results of an annotation methodology aimed at enriching the Italian ISST-TANL dependency-annotated corpus with frame information encoded according to the FrameNet paradigm. The creation of such a multi-layered corpus was conceived as the first step towards the development of semantic role labelling systems that could significantly contribute to the extension of the frame-annotated corpus as well as profitably be used in content-related natural language processing tasks. As discussed in the previous sections, many of the annotation choices were designed and motivated with this scenario in mind. First steps in this direction have already been taken: i.e. the ISST-TANL corpus enriched with semantic frame information was used as a training corpus for semantic-role labeling systems in the framework of the *Frame Labeling over Italian Texts* (FLaIT) task⁴ within EVALITA 2011, an initiative devoted to the evaluation of Natural Language Processing and Speech tools for Italian⁵. In the FLaIT task, systems were expected to label semantic frames and their arguments as evoked by input predicate words over plain text sentences: participant systems were based on a variety of learning techniques and turned out to achieve very good results, i.e. over 80% of accuracy in most defined subtasks (Basili et al., 2012).

7. Acknowledgments

The authors would like to thank Fabio di Basilio who contributed to the annotation process. Special thanks go to Eva Maria Vecchi who supervised the team of annotators.

8. References

- R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34:555–596.
- G. Attardi and F. Dell’Orletta. 2009. Reverse revision and linear tree combination for dependency parsing. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies short papers (NAACL HLT)*, Boulder, Colorado.
- R. Basili, D. De Cao, A. Lenci, A. Moschitti, and G. Venturi. 2012. Evalita 2011: the frame labeling over italian texts task. In *Working Notes of EVALITA 2011, 23-24th January 2012, ISSN 2240-5186*, Rome, Italy.
- A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Padò, and M. Pinkal. 2006. The salsa corpus: a german corpus resource for lexical semantics. In *Proceedings of the 5th international conference on Language Resources and Evaluation (LREC 2006)*, pages 969–974, Genova, Italy.
- B. Chen and P. Fung. 2004. Automatic construction of an english–chinese bilingual framenet. In *Proceedings of HLT/NAACL*, Boston, USA.
- A. Dolbey, M. Ellsworth, and J. Scheffczyk. 2006. Bioframenet: A domain-specific framenet extension with links to biomedical ontologies. In *Proceedings of the “Biomedical Ontology in Action” Workshop at KR-MED*, Baltimore, Maryland.
- A. Dolbey. 2009. *BioFrameNet: a FrameNet Extension to the Domain of Molecular Biology*. University of California, Berkeley doctoral dissertation.
- K. Erk and S. Pado. 2006. Shalmaneser - a flexible toolbox for semantic role assignment. In *Proceedings of LREC 2006*, Genoa, Italy.
- K. Erk, A. Kowalski, and S. Padò. 2003. The salsa annotation tool-demo description. In *Proceedings of the 6th Lorraine–Saarland Workshop*, pages 111–113, Nancy, France.
- C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- C.J. Fillmore. 1985. Frame and the semantics of understanding. *Quaderni di semantica*, 6(2):222–254.
- A. Gangemi, M.T. Sagri, and D. Tiscornia. 2005. A constructive framework for legal ontologies. In J. Breuker R. Benjamins, P. Casanovas and A. Gangemi, editors, *Law and the Semantic Web*, pages 97–124. Berlin: Springer verlag edition.
- R. Johannson and P. Nugues. 2006. A framenet-based semantic role labeler for swedish. In *Proceedings of Coling/ACL*.
- J.R. Landis and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.
- S. Montemagni and M. Simi, 2007. *The Italian dependency annotated corpus developed for the CoNLL-2007 Shared Task*. available online at http://www.ilc.cnr.it/tressi_prg/ISST@CoNLL2007/ISST@CoNLL2007.pdf.
- S. Montemagni, F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Lenci, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M.T. Paziienza, D. Saracino, F. Zanzotto, N. Mana, F. Pianesi, and R. Delmonte, 2003. *Building and the Italian Syntactic–Semantic Treebank*, pages 189–210. Kluwer, Dordrecht.
- K.H. Ohara. 2008. Lexicon, grammar and multilinguality in the japanese framenet. In *Proceedings of LREC 2008*, Marrakech, Morocco.
- M. Palmer, D. Gildea, and P. Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics Journal*, 31(1).
- J. Ruppenhofer, M. Ellsworth, M.R.L. Petruck, C.R. Johnson, and J. Scheffczyk, 2010. *FrameNet II: Extended Theory and Practice*. available online at http://framenet.icsi.berkeley.edu/fndrupal/the_book.

⁴http://www.evalita.it/2011/tasks/frame_labeling

⁵<http://www.evalita.it/>

- S. Siegel and N.J. Castellan. 1988. *Nonparametric statistics for the Behavioral Sciences*. McGraw-Hill, 2nd edition edition.
- C. Subirats. 2009. Spanish framenet: A frame-semantic analysis of the spanish lexicon. In H.C Boas, editor, *Multilingual FrameNets in Computational Lexicography : Methods and Applications*, pages 135–162. Mouton de gruyter edition.
- G. Venturi. 2011. Semantic annotation of italian legal texts: a framenet-based approach. In K. Ohara and K. Nikiforidou, editors, *Constructions and Frames 3:1*, pages 46–79. John benjamins publishing company edition.
- G. Venturi. 2012. Design and development of TEMIS: a syntactically and semantically annotated corpus of italian legislative texts. In *Proceedings of the 4th Workshop Semantic Processing of Legal Texts (SPLeT 2012), held in conjunction with the 8th international conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey.