Design and Development of TEMIS: a Syntactically and Semantically Annotated Corpus of Italian Legislative Texts

Giulia Venturi

Scuola Superiore Sant'Anna di Studi Universitari e di Perfezionamento - Pisa (Italy)
Piazza Martiri della Libertà, 33
giulia.venturi@ilc.cnr.it

Abstract

Methodological issues concerning the design and the development of TEMIS, a syntactically and semantically annotated corpus of Italian legislative texts, are presented and discussed in the paper. TEMIS is a heterogeneous collection of texts exemplifying different sub–varieties of Italian legal language, i.e. European, national and local texts. The whole corpus has been dependency annotated and a subset has been enriched with frame–based information by customizing the formalism of the FrameNet project. In both cases, a number of domain–specific extensions of the annotation criteria developed for the general language has been foreseen. The interest in building such a corpus stems from the increasing need for annotated collections of domain–specific texts recognized by both the Artificial Intelligence and Law (AI&Law) community and the Natural Language Processing (NLP) one. In two research communities the benefits of having a resource where both domain–specific content and its underlying linguistic structure are made explicit and aligned are widely acknowledged. To the author knowledge, this is the first annotated corpus of legal texts overtly devoted to be used for legal text processing applications based on NLP tools.

Keywords: Legal Text Processing, Syntactic and Semantic Annotation, Domain–specific Gold Corpora

1. Introduction

This paper presents issues and challenges encountered in designing and developing a corpus of Italian legislative texts enriched with two different layers of linguistic annotation, i.e. a syntactic dependency layer and a semantic one. The interest in building such a corpus stems from the increasing need for annotated collections of domain–specific texts recognized by both the Artificial Intelligence and Law (AI&Law) community and the Natural Language Processing (NLP) one.

On the one hand, in the last few years, a growing body of research and practice has been concerned with the use of Human Language Technology (HLT) for automating knowledge extraction from legal texts and for processing legal language. Such an interest is testified by several events organised on this topic, e.g. the Workshop "Applying Human Language Technology to the Law" held in 2011 (AHLTL 2011)[1] or past editions of the Workshop "Semantic Processing of Legal Texts" (SPLeT), focussed on issues and challenges concerning the use of Natural Language Processing tools in the legal domain.

However, researchers addressing these topics have to confront with the lack of large annotated legal text corpora to be used as reference domain–specific resources. As demonstrated by the promising results achieved in the bio–medical field, corpora annotated at different levels of analysis (e.g. syntactic and semantic levels) play a key role for a number of domain–specific NLP tasks (e.g. biological text mining, the construction of domain–specific ontological resources, event extraction) grounded on the automatic processing of domain–specific corpora.

On the other hand, in the NLP community, it is well known that annotated corpora are valuable resources for the automatic construction of statistical models which can be used in a number of different NLP tasks. However, currently available statistically trained NLP tools are mostly based on corpora made up of texts from the news domain. Applying these tools to out–of–domain corpora is known to be problematic (Gildea, 2001): when applied to domain–specific texts (e.g. bio–medical literature, law texts) their accuracy decreases significantly. Since available domain–specific resources are fundamental in supervised scenarios to adapt statistical NLP tools to new domains, some effort has been devoted to the construction of such resources. The most notable case is represented by the bio–medical domain where the GENIA corpus (Ohta et al., 2002), a collection of biomedical literature annotated with various levels of linguistic (e.g. morphological, syntactic) and semantic information (e.g. domain–specific entities, relational information), has been developed. The corpus is currently used for domain–specific semantic processing applications, e.g. for mining biomedical events from literature (Kim et al., 2008), as well as for supervised domain adaptation purposes, e.g. for improving the performance of statistical syntactic parsers by using bio–medical texts as additional training data (McClosky and Charniak, 2008).

Given these premises, the present article aims at illustrating the main methodological issues faced in syntactically and semantically annotating the TEMIS corpus (SynTactically and **SEM**antically Annotated **I**talian Legislative Corpu**S**). To the author knowledge, it is the first multi–level annotated corpus of legal texts specifically designed to contribute to the development of NLP–based legal text processing applications.

The paper is organized as follows. In Section 2., motivations for developing a multi–level annotated corpus of legal texts for semantic processing purposes are reported, together with related studies. In Section 3., a description

[1] http://wyner.info/research/Papers/AHLTL2011Papers.pdf

of the TEMIS corpus is presented, including some main legal language peculiarities characterizing the corpus with respect to a corpus of general Italian language. Sections 4. and 5. are respectively devoted to illustrate the syntactic and semantic approaches adopted for corpus annotation. Some of the ongoing research activities devoted to use TEMIS for parsing legal texts are illustrated in Section 6. Conclusions and future developments of this work are reported in Section 7.

2. Motivations and related work

The interest in developing a corpus enriched with syntactic and semantic information stems from the acknowledged benefits of having a domain–specific document collection where both domain–specific content and its underlying linguistic structure are made explicit and aligned. In other words, the present work was motivated by the fundamental role that a syntactically and semantically annotated corpus of legal texts could play for several NLP–based applications in the legal domain.

In what follows, recent related studies focussed on developing syntactically or semantically annotated corpora of legal texts are discussed.

2.1. Related work on syntactic annotation of legal texts

This work starts from the idea that any legal text semantic processing application "would be further supported with the creation of a large scale corpus of parsed legal documents" (Wyner and Peters, 2011). Similarly to open–domain NLP–based applications such as Information Extraction, Question Answering, Machine Translation, etc., it is broadly acknowledged that several domain–specific semantic processing applications, e.g. rule extraction from regulations (Wyner and Peters, 2011), formal representations of individual sentences occurring in legal provisions (de Maat and Winkels, 2011), automatic detection of arguments in legal texts (Palau and Moens, 2011), can benefit significantly from operating against the output of a syntactic parser.

However, to the author knowledge, no syntactically annotated corpora of legislative texts are available so far for any language. Accordingly, current statistical parsers are trained on corpora of newspapers, representative of *open–domain* texts. This affects their performances with respect to legal texts.

One exception is the portion of the Turin University Treebank (TUT)[2], developed at the University of Torino, including a section of the Italian Civil Law Code (28,048 word tokens, for a total of 1,100 sentences) annotated with syntactic dependency information. However, this corpus is representative of a legal language sub–variety acknowledged to be less complex with respect to other kinds of legislative texts such as laws, decrees, regulations, etc. According to one of the main scholar of legal language such as Garavelli (2001), the Civil Law Code articles are less representative of the much cited linguistic complexity of Italian *legalese* with respect to other kinds of legislative texts. From an applicative point of view, this is witnessed by the results achieved in the "Dependency Parsing" track of Evalita 2011 (Bosco and Mazzei, 2012) where all participant parsers have shown better performances when tested on the Italian Civil Law Code test set than when tested on newspapers test corpus. On the contrary, the "Domain Adaptation Track" organized at Evalita 2011 (Dell'Orletta et al., 2012), where a sub–set of the corpus presented in this paper has been used, revealed that parsing systems need to be further adaptated to reliably analyse legal texts such as laws, decrees, regulations, etc.

2.2. Related work on semantic annotation of legal texts

Attention to issues and challenges posed by the semantic annotation of legal texts originates from the increasing interest in legal knowledge management tasks based on automatic text processing. Accordingly, several NLP–oriented works have appeared on this topic. Even though they differ in the approach, they aim at making legal texts structured and informative for different automatic semantic processing applications, such as legal argumentation mining (Palau and Moens, 2011), legal text summarization (Hachey and Grover, 2006), court decisions structuring (Kuhn, 2010), legal metadata extraction (see among others for the Italian case (Bartolini et al., 2004; Mazzei et al., 2009; Spinosa et al., 2009)), legal definitions extraction (Walter, 2009), legal case elements and case factors extraction (Wyner, 2010; Wyner and Peters, 2010b; Wyner and Peters, 2010a), legal information retrieval (Maxwell et al., 2009), rule extraction from regulations (Wyner and Peters, 2011), etc.

However, in spite of this widespread interest very little work has been devoted so far to developing a semantically annotated corpus to be used as reference corpus for some of the above mentioned legal text processing applications. To the author knowledge, two exceptions are the Vaccine/Injury Project Corpus (Walker et al., 2011) and the corpus of Brazilian court decisions and legislative texts semantically annotated according to Frame Semantics principles (Bertoldi and Chishman, 2012).

In the first case, Walker and colleagues have built a collection of legal decisions awarding or denying compensation for health injuries allegedly due to vaccinations and they have annotated it with models of the logical structure of the reasoning of the factfinders. The corpus is meant to provide "useful data for formal and informal logic theory, for natural–language research in linguistics, and for artificial intelligence research in those cases". In the second case, a corpus representative of the Brazilian legal language has been annotated with semantic frames information, i.e. by applying the *Frame Semantics* theory (Fillmore, 1985) and the FrameNet paradigm to the semantic annotation of legal texts. The Bertoldi and Chishman' initiative "is part of a larger project that researches how linguistic information could be used to improve legal information management and legal information retrieval in the Brazilian courts".

[2] http://www.di.unito.it/∼tutreeb/

3. TEMIS: a Syntactically and Semantically Annotated Italian Legislative Corpus

This section is intended to provide the overall description of the TEMIS resource and the principles which guided its design and construction.

The TEMIS corpus has been originally developed in the framework of the author's Ph.D thesis. Starting from a small set of sentences exemplifying legal language, the corpus has been further enlarged in the occasion of the Evalita 2011 campaign where a subset has been used in the "Domain Adaptation" track. As discussed in (Dell'Orletta et al., 2012) where the results of the track were reported, in that occasion the TEMIS subset was used as test corpus. This allowed quantifying the negative impact that the language used in legislative texts such as laws, decrees, regulations, etc. has on the performances of participant parsers trained or developed on newspaper language.

Three annotators, all with graduate training in linguistics, participated both in the syntactic and in the semantic annotation stage.

3.1. Corpus composition

TEMIS is a collection of legislative texts enacted by three different releasing agencies, i.e. European Commission, Italian State and Piedmont Region, and regulating a variety of domains, ranging from environment, human rights, disability rights to freedom of expression. It is a heterogeneous document collection including legal acts such as national and regional laws, European directives, legislative decrees, etc., as well as administrative acts, such as ministerial circulars, decision, etc.

This heterogeneous nature makes TEMIS a resource able to exemplify different sub–varieties of Italian legal language. Table 1 reports how the three different legal text types (i.e. European, national and local texts) are variously represented in the corpus.

Releasing agency	No. tokens	No. sentences
European Commission	6,683	275
Italian State	3,670	94
Piedmont Region	5,453	135
Total	15,804	504

Table 1: Distribution of different legal text types in TEMIS.

3.2. Corpus linguistic profile

In order to get evidence of the linguistic specificity of the legislative texts included in TEMIS, the corpus has been investigated with respect to a number of different parameters, which according to the literature on register variation (Biber and Conrad, 2009) are indicative of textual genre differences.

Different kinds of features have been taken here as representative of the linguistic profile of the considered legislative texts. They range from raw text features, such as sentence length, to more complex ones (e.g. parse tree depth) detected from the syntactic level of annotation. In what follows the most significant ones are illustrated and discussed. A comparison with the respective features for an Italian newswire corpus, chosen to be representative of general Italian language, helps to highlight the TEMIS's main linguistic characteristics. The ISST–TANL corpus, jointly developed by the Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR) and the University of Pisa in the framework of the TANL (Text Analytics and Natural Language processing) project, has been used. The corpus consists of articles from newspapers and periodicals, selected to cover a high variety of topics (politics, economy, culture, science, health, sport, leisure, etc.).

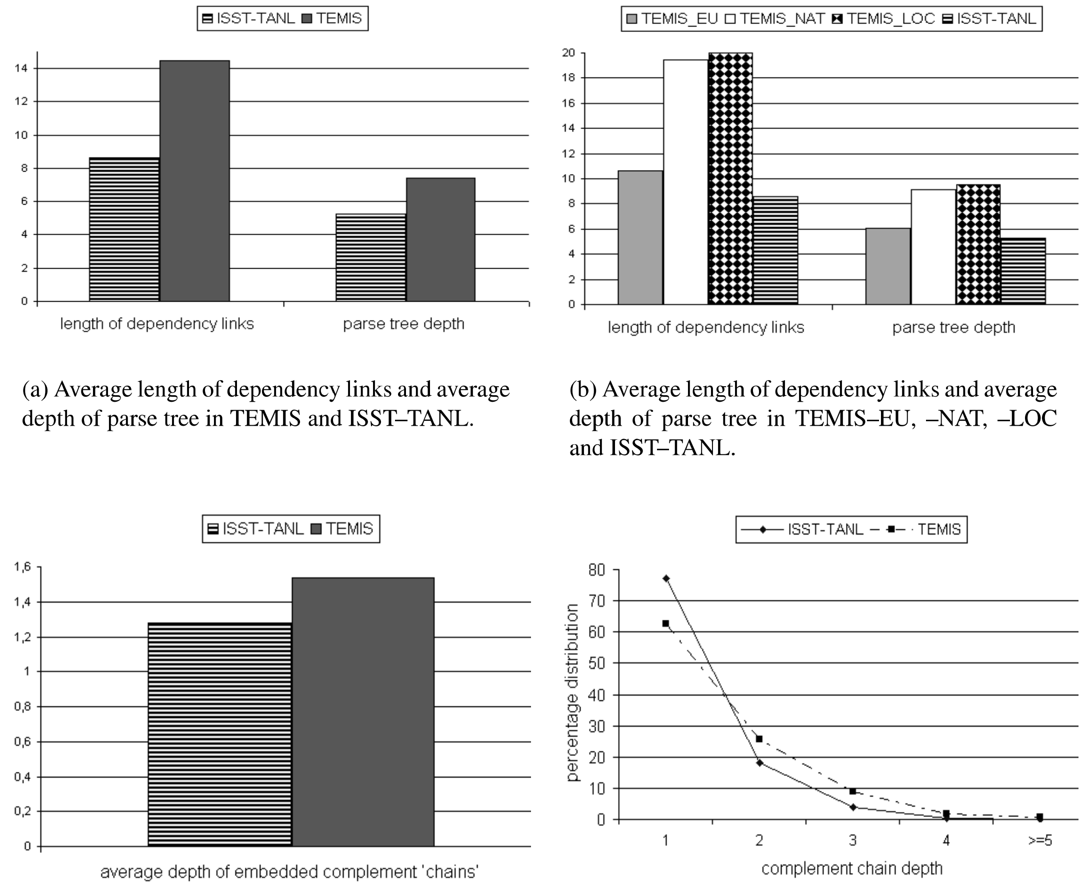
The TEMIS and ISST–TANL corpora differ significantly in many aspects starting from the average sentence length, calculated as the average number of words per sentence. As Table 2 shows, some differences can be also found amongst the three considered kinds of legal text sub–varieties. Notably, the legal texts enacted by the European Commission (TEMIS–EU, in the Table) show a behaviour which is more similar to ordinary language than the national (TEMIS–NAT) and local (TEMIS–LOC) legal texts.

Corpus	Avg sentence length (in tokens)
ISST–TANL	21.87
TEMIS	31.36
TEMIS–EU	24.56
TEMIS–NAT	39.04
TEMIS–LOC	41.95

Table 2: Average sentence length in *i)* TEMIS and ISST–TANL corpora and *ii)* the three TEMIS's sub–corpora.

Interestingly, several differences can be found between the two corpora with respect to the distribution of features typically correlated with text complexity, such as parse tree depth and length of dependency links. According to the approach to linguistic monitoring described in (Dell'Orletta et al., 2011), the TEMIS and ISST–TANL corpora have been compared with respect to *i)* the average length of dependency links, measured in terms of the words occurring between the syntactic head and the dependent, and *ii)* the average depth of the whole parse tree, calculated in terms of the longest path from the root of the dependency tree to some leaf. As it can be seen in Figure 1(a), *i)* legislative sentences contain dependency links much longer on average (14.5) than the ones of the general–Italian sentences (8.61) and *ii)* the average parse tree height of TEMIS (7.44) is higher than the one characterizing the ISST–TANL sentences (5.28). In addition, as it was previously pointed out, the Italian European legal texts have syntactic features which make them more similar to ordinary language than the national and local legal texts (see Figure 1(b)).

It is here worth noting (see Figure 1(c)) that TEMIS's sentences are characterized by an average depth of embedded complement 'chains' governed by a nominal head and including either prepositional complements or nominal and adjectival modifiers (1.54) higher than the one of ISST–TANL's sentences (1.28). However, the crucial distinguishing characteristic of legislative sentences appears to be the different percentage distributions of embedded complement



(a) Average length of dependency links and average depth of parse tree in TEMIS and ISST–TANL.

(b) Average length of dependency links and average depth of parse tree in TEMIS–EU, –NAT, –LOC and ISST–TANL.

(c) Average depth of embedded complement 'chains' in TEMIS and ISST–TANL.

(d) Percentage distributions of embedded complement 'chains' by depth in TEMIS and ISST–TANL.

Figure 1: Comparative syntactic behaviours in TEMIS and ISST–TANL.

'chains' by depth. As Figure 1(d) shows, legislative texts appear to have an higher percentage of deep complement 'chains' with respect to the Italian reference corpus.

4. The syntactic level of annotation

All the 504 sentences included in the TEMIS corpus have been enriched with a syntactic level of annotation. For this purpose, a semi–automatic strategy has been adopted. Firstly, the TEMIS corpus was automatically dependency-parsed by the DeSR parser (Attardi, 2006) using *i)* Support Vector Machine as learning algorithm and *ii)* the ISST–TANL corpus as training corpus.

Secondly, the result of the first stage was manually revised. The "Dependency Grammar Annotator" (DgAnnotator) tool[3] was used for such a manual revision step.

This semi–automatic annotation strategy is meant to reduce the annotation arbitrariness to a minimum. This allowed to keep consistent the manual annotation as well as to identify the main parsing errors due to the unique features of legal language.

4.1. Dependency Annotation Scheme and Data Format

The dependency syntactic annotation scheme developed for the ISST–TANL corpus labelling has been used for the TEMIS corpus annotation. That is, the dependency tagset[4] was maintained even though a number of extensions of the annotation criteria have been introduced in order to properly handle legal language syntactic peculiarities.

The dependency annotation format adheres to the standard CoNLL–2007 tabular format used in the "Shared Task on Dependency Parsing" (Nivre et al., 2007). Accordingly, each word–token is provided with information concerning the corresponding lemma, coarse- and fine–grained part–of–speech[5], morphological features[6], head of the dependency relation and the dependency relation type.

[3] It is an annotating and visualizing Dependency Graphs tool freely available at http://medialab.di.unipi.it/Project/QA/Parser/DgAnnotator/

[4] A description of the dependency tagset can be found at http://poesix1.ilc.cnr.it/ISST-TANL-DEPtagset-web.pdf

[5] A description of the part-of-speech (coarse- and fine-grained) tagsets can be found at http://poesix1.ilc.cnr.it/ISST-TANL-MStagset-web.pdf

[6] A description of the part-of-speech (coarse- and fine-grained) tagsets can be found at http://poesix1.ilc.cnr.it/ISST-TANL-MS_FEATStagset-web.pdf

For example, the following sentence is annotated as Table 3 reports:

- *Gli Stati membri provvedono affinché il gestore sia obbligato a trasmettere all'autorità competente una notifica entro i seguenti termini.* ('Member States shall require the operator to send the competent authority a notification within the following time–limits'.)

In Table 3, it can be noted that each word form (in the column headed *FORM*) univocally marked by a numerical identifier (column *ID*) is associated with its corresponding lemma (column *LEMMA*), its coarse– (column *CPOSTAG*) and fine–grained (column *POSTAG*) part–of–speech and its morphologycal treats (column *FEATS*). Moreover, the annotation makes explicit the head of the dependency syntactic relation in which each word is involved (column *HEAD*) and the type of dependency relation (columnn *DEPREL*). For example, Table 3 shows that the word *notifica* ('notification') is the object (*obj*) of the verb *trasmettere* ('send').

4.2. Domain–specific extensions of the open–domain annotation criteria

The annotation criteria developed for the annotation of an *open–domain* corpus such as the ISST–TANL needed to be extended in order to properly handle specific syntactic peculiarieties specific to the legal language. Such extensions are concerned with different levels of text annotation ranging from the sentence splitting to the dependency annotation. The most significant cases are described in the following sections.

4.2.1. Sentence splitting

Differently from the criteria adopted for the open–domain case, here sentence splitting was overtly meant to preserve the original structure of the legal text. This entails that also punctuation marks such as ';' and ':', when followed by a carriage return, are treated as sentence boundary markers. Such an extension allowed to handle with specific cases frequently occurring in legislative texts, such as:

1. sentences that, occurring in a legislative preamble, start with phrases, such as *considerato che* ('Having regard to'), and end with a clause boundary punctuation mark, such as ';'

2. sentences that end with a clause boundary punctuation mark such as ':' and introduce an itemized list

3. sentences that, part of an itemized list, end with a clause boundary punctuation mark such as ';'.

4.2.2. Dependency annotation

In order to successfully cope with domain–specific syntactic constructions hardly or even never occurring in the ISST–TANL corpus, dependency annotation criteria have been extended to cover the annotation of the main following cases:

1. elliptical constructions frequently adopted in citations to whole legal texts or to specific partitions of legal texts (e.g. article, paragraph, etc.). This is the case, for example, of the sentence *i decreti legislativi di cui al comma 1 si conformano ...* ('legislative decrees referred to in paragraph 1 shall comply ...') which has been annotated as Figure 2 shows. Since a verb is missing in the relative clause *di cui al comma 1* ('referred to in paragraph 1'), a relative–modifier dependency relation (*mod_rel*) has been found between the antecedent *decreti* ('decrees'), the head token, and *al* ('to'), the dependent token;

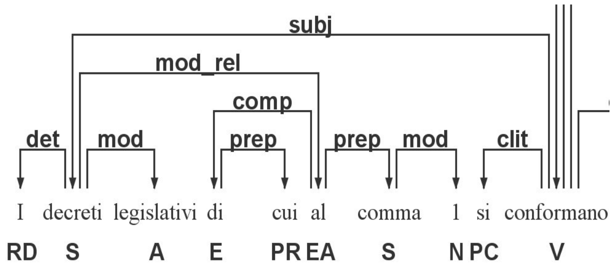


Figure 2: Example of annotation of elliptical construction.

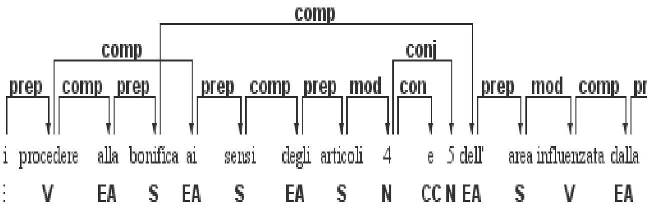


Figure 3: Example of non–projective link.

2. participial phrases, such as *fatto salvo* ('without any reserve'), used to express exceptions or limitations to main clauses. For example, as Figure 5 shows, in the sentence *il contravventore, fatti salvi ogni altro adempimento o comminatoria previsti dalle leggi vigenti, è tenuto al pagamento di una sanzione amministrativa ...* ('the infringer, without prejudice to any other obligation or on pain of applicable law, is required to pay an administrative penalty'), a modifier dependency relation (*mod*) has been found between the head of the participial phrase (i.e. *fatti*) and the syntactic head of the main clause (i.e. *tenuto* 'required');

3. non–projective links, often occurring in legal texts and mostly due a frequent exploitation of the free–word order nature of the Italian language. This is the case of the annotation excerpt of the sentence *E' fatto comunque salvo l'obbligo di procedere alla bonifica ai sensi degli articoli 4 e 5 dell'area influenzata dalla fonte inquinante* ('Is without prejudice to the obligation to carry out drenage in accordance with articles 4 and 5 of the area affected by pollution sources') reported in Figure 3 where not all the tokens part of the *comp* relation are dependent from the head token *bonifica* ('drenage'); on the contrary, the head of the token *ai* ('in') is *procedere* ('carry out');

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS	HEAD	DEPREL
1	Gli	il	R	RD	num=p|gen=m	2	det
2	Stati	Stati	S	SP	_	4	subj
3	membri	membro	S	S	num=p|gen=m	2	mod
4	provvedono	provvedere	V	V	num=p|per=3|mod=i|ten=p	0	ROOT
5	affinché	affinché	C	CS	_	4	mod
6	il	il	R	RD	num=s|gen=m	7	det
7	gestore	gestore	S	S	num=s|gen=m	9	subj_pass
8	sia	essere	V	VA	num=s|per=3|mod=c|ten=p	9	aux
9	obbligato	obbligare	V	V	num=s|mod=p|gen=m	5	sub
10	a	a	E	E	_	9	arg
11	trasmettere	trasmettere	V	V	mod=f	10	prep
12	all'	a	E	EA	num=s|gen=n	11	comp_ind
13	autorità	autorità	S	S	num=n|gen=f	12	prep
14	competente	competente	A	A	num=s|gen=n	13	mod
15	una	una	R	RI	num=s|gen=f	16	det
16	notifica	notifica	S	S	num=s|gen=f	11	obj
17	entro	entro	E	E	_	11	comp_temp
18	i	il	R	RD	num=p|gen=m	20	det
19	seguenti	seguente	A	A	num=p|gen=n	20	mod
20	termini	termine	S	S	num=p|gen=m	17	prep
21	.	.	F	FS	_	4	punc

Table 3: An example of an annotated sentence in CoNLL format extracted from TEMIS.

4. internal partitions of a legislative text (e.g. article, paragraph) that are hierarchically organized. They are treated as embedded modifier 'chains' governed by a nominal head, as exemplified by the annotation of the sentence *ai sensi dell'articolo 94, comma 3, lettera a) della l.r. 44/2000* ('under article 94, paragraph 3, letter a) of the local act 44/2000') reported in Figure 4. In this case, the internal partitions of the *l.r. 44/2000* ('local act 44/2000'), i.e. *articolo 94, comma 3, lettera a)* ('article 94, paragraph 3, letter a)'), has been annotated as a chain of nominal modifiers (*mod*).

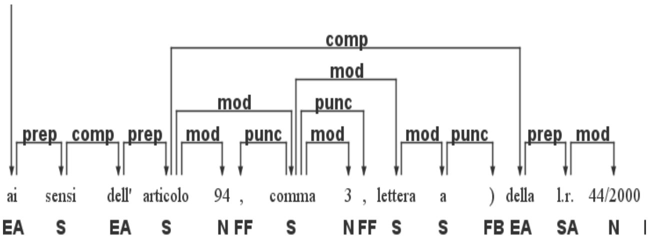

comp
mod
mod
punc
prep
comp
prep
mod
punc
mod
mod
punc
prep
mod
ai sensi dell' articolo 94 , comma 3 , lettera a) della l.r. 44/2000
EA S EA S N FF S N FF S S FB EA SA N I

Figure 4: Example of annotation of internal partitions of a legislative text.

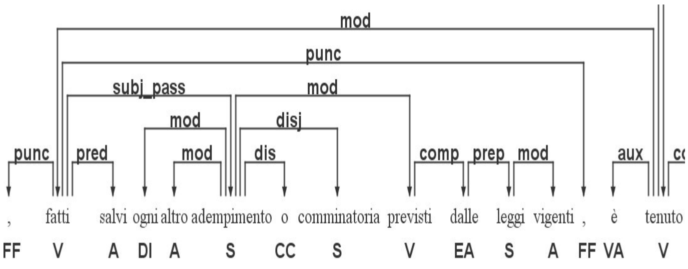

mod
punc
subj_pass
mod
mod
disj
punc
pred
mod
dis
comp
prep
mod
aux
cc
, fatti salvi ogni altro adempimento o comminatoria previsti dalle leggi vigenti , è tenuto
FF V A DI A S CC S V EA S A FF VA V

Figure 5: Example of annotation of participial phrase.

5. The semantic level of annotation: a FrameNet–based approach

A subset of the TEMIS corpus has been further enriched with a level of semantic annotation. The semantic annotation paradigm developed in the framework of the FrameNet project[7] has been adopted and specialized in order to properly describe the lexical semantic content of legislative texts included in TEMIS.

This initiative is part of the work that has been jointly done at the University of Pisa (Department of Linguistics) and at the Institute of Computational Linguistics "Antonio Zampolli" (ILC-CNR) where a frame-based annotation of the ISST-TANL corpus has been carried out in order to enrich the treebank with semantic frame information, as described in (Lenci et al., 2012).

By starting from the suggestion expounded by Dolbey et al. (2006a) that FrameNet can be seen "as a backbone of several domain–specific FrameNets", the annotation methodology adopted for the ISST–TANL corpus has been extendend and specialized. This aims at showing that a FrameNet–like approach to text annotation can be suitable for insightful analyses of general language, as well as an advantageous starting point for producing descriptions of syntactic and semantic combinatorial possibilities exhibited in a specialized language such as the legal language.

In particular, the annotation effort has been devoted to making explicit how the three main deontic modalities, i.e. *obligation, permission, prohibition*, are linguistic realized

[7] https://framenet.icsi.berkeley.edu/fndrupal/

in the TEMIS corpus. Accordingly, the annotation of the frames reported in Table 4 has been the main focus.

Formally represented in the existing legal ontologies, these three fundamental legal concepts are hardly associated with their actual lexical realization. Accordingly, the underlying idea here is that a FrameNet–based and linguistically–oriented representation of legal semantics would complement a domain–oriented one by providing a semantic description anchored to its corresponding textual realization. As Dolbey et al. (2006a) suggested, "FrameNet–style ontological descriptions of language can be integrated with information from" already existing domain–specific ontologies, such as bio–medical ontologies (in Dolbey's case) or legal ontology.

5.1. The FrameNet project

FrameNet 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 belongs to different 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", etc. , 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 written in subscript) are its FEs:

- [Matilde $_{Cook}$] **fried** [the catfish $_{Food}$] [in a heavy iron skillet $_{Heathing\_instrument}$].

The type of representation produced by FrameNet is a network of "situation–types" (frames) organized across inheritance relations between frames, as opposed to a network of meaning nodes, as in the case of WordNet. In FrameNet, FE can be also specified with Semantic Types (i.e. ontological categories) employed to indicate the basic typing of fillers that are expected in the FE. 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.

Despite the fact that FrameNet annotations are triples where each FE realization is coupled with its phrase type (e.g. NP, PP, etc.) and its grammatical function (e.g. object, subject, etc.) played in the annotated sentence, as overtly claimed by (Dolbey, 2009), "FrameNet annotations are not linked to syntactic parse trees"; consequently, it is often the case that frame elements instantiations do not "correspond to syntactic constituents provided by a syntactic parse of the whole sentence".

5.2. Annotation methodology

The TEMIS's semantic annotation has been focussed on a sub–set of the whole corpus. Namely, a set of 9,273 tokens (for a total of 226 sentences) extracted from the three considered sub–varieties of legislative texts has been semantically annotated.

The annotation methodology which has been followed intends to continue the one described in (Venturi, 2011b) where a FrameNet–based approach to the semantic annotation of Italian legislative texts has been presented. The strategy devised for the annotation of the ISST-TANL corpus has been mostly adopted, even though a number of domain–specific customizations have been considered.

Similarly to the ISST–TANL case, the annotation was carried out manually with the SALTO tool (Erk et al., 2003). The syntactic annotation level of the TEMIS corpus 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 FrameNet, by following the methodology fully described in (Lenci et al., 2012).

The annotation has been carried out on top of dependency syntactic representation. This entails that during the annotation phase frames and FEs have been anchored to the dependency annotation. Such an approach mostly resembles the strategy to frame–semantic corpus annotation that the SALSA project (Burchardt et al., 2006) adopted, which included manually annotating a large corpus of German newspapers with semantic role information starting from a syntactically annotated corpus.

FrameNet–style annotations are not linked to syntactic parse trees: the choice to ground the level of semantic annotation on the fully parsed texts is intended to overcome this state of affairs. Indeed, as Dolbey (2009) emphasizes, the annotation strategy adopted in the FrameNet project may cause "difficulties for end users who want to perform automatic processing that includes information from FrameNet's annotation collection".

An example of annotation is reported in Figure 6, where the following two frames have been annotated on top of the syntactic dependency tree of the sentence *Obbligati al pagamento della tassa sono gli esercenti i grandi impianti di combustione di cui all'articolo 1* (lit. 'Obligated to tax payment are tradespeople of big combustion plants mentioned by Article 1'):

1. BEING_OBLIGATED frame, evoked by the verb *obbligare* ('to obligate') in the passive form (i.e. *essere obbligato*, 'be obligated'), with the FEs "Responsible party" and "Duty", instantiated as passive subject (*subj_pass*) and as complement (*comp*) of the verb, respectively;

2. COMMERCE_PAY frame, evoked by the deverbal noun *pagamento* ('payment'), with the constructionally null instantiated FE "Buyer", omitted as the passive subject of the main verb *obbligare* ('to obligate'), and the "Money" FE, syntactically instantiated as a complement (*comp*) of the deverbal noun.

In order to properly represent legal semantics, the foreseen customizations of the annotation methodology adopted for the ISST–TANL corpus mainly concern the *i)* the kind of legal content one would like to make explicit and *ii)* the

Deontic modality: **obligation**		
FrameNet frame	FrameNet definition	Frame–evoking LUs annotated in TEMIS
OBLIGATION_SCENARIO (Non–Lexical Frame)	Under some, usually implicit, Condition a Duty needs to be fulfilled by a Responsible_party. If the Duty is not performed, there may be some undesirable social Consequence for the Responsible_party. This Consequence may or may not be stated overtly.	–
BEING_OBLIGATED	Under some Condition, usually left implicit, a Responsible_party is required to perform some Duty. If they do not perform the Duty, there may be some undesirable Consequence, which may or may not be stated overtly.	*tenuto* 'required', *obbligato* 'obligated', *chiamato* 'called', *obbligo* 'obligation', *costretto* 'forced', *sottoposto* 'subjected', *(essere)soggetto* '(to be)subject', *(avere)obbligo* '(have)obligation'
BEING_OBLIGATORY	Under some Condition, usually left implicit, a Duty needs to be fulfilled by a Responsible_party. If the Duty is not performed, there may be some undesirable Consequence for the Responsible_party, which may or may not be stated overtly. Compare this frame to the Being_obligated frame.	*obbligatorio* 'obligatory', *spettare* 'to be due', *dovuto* 'due', *incombere* 'to be incumbent'
IMPOSING_OBLIGATION	A Duty is imposed on a Responsible_party according to a Principle which regulates how the Responsible_party should respond to a Situation. The Situation may be expressed metonymically by reference to an Obligator, whose action invokes the Principle. It is only rarely the case that the Principle and the Situation/Obligator are both expressed overtly.	*irrogato* 'imposed', *irrogare* 'to impose', *disporre* 'to decide', *prevedere* 'to provide', *imposto* 'imposed', *predisporre* 'to establish', *definire* 'to fix', *stabilire* 'to establish', *istituire* 'to introduce', *prescrizione* 'prescription', *obbligare* 'to obligate', *disposto* 'provided', *determinare* 'to fix', *(fare)obbligo* '(make)obligation'
Deontic modality: **permission**		
PERMITTING	In this frame a State_of_affairs is permitted by a Principle. Raising constructions are common in this frame. In this frame the Principle which sanctions the State_of_affairs is not an agent who grants permission to a specific individual or group of individuals, and thus differs from the Grantor in the Grant_permission frame.	*autorizzato* 'authorized', *autorizzare* 'to authorize', *ammesso* 'permitted', *accordato* 'granted', *consentire* 'to allow', *consentito* 'allowed', *concessione* 'permission', *concesso* 'granted', permesso 'permission'
Deontic modality: **prohibition**		
PROHIBITING	In this frame a State_of_affairs is prohibited by a Principle. Raising constructions are common in this frame. In this frame the Principle which prohibits the State_of_affairs is not an agent who denies permission to a specific individual or group of individuals, and thus differs from the Authority in the Deny_permission frame.	*interdizione* 'disability', *divieto* 'prohibition', *vietato* 'prohibited', *(fare)divieto* '(make)prohibition'
DENY_PERMISSION	In this frame, an Authority orders a Protagonist not to engage in an Action.	*divieto* 'prohibition', *interdizione* 'disability', *negare* 'to deny', *proibire* 'to prohibit', *(fare)divieto* '(make)prohibition'

Table 4: FrameNet frames describing the *obligation, permission, prohibition* deontic modalities and the corresponding evoking lexical units annotated in the TEMIS corpus.

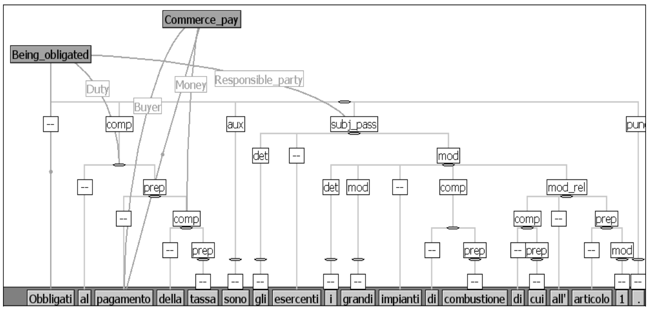


Figure 6: An example of frame–semantic annotation using the SALSA tool.

kind of domain–specific customizations and extensions of the general FrameNet resource required by the legal domain.

5.2.1. Between lexicographic and full–text annotation

The first issue to address for a legal semantic annotation task concerns the kind of text content to make explicit. It is well–known that the law simultaneously *describes* objects and events and *regulates* them. Thus, *legal domain knowledge* is mixed with knowledge of domain of interest to be regulated (i.e. *world knowledge*).

To successfully cope with this mixture, a special annotation mode has been adopted, which is meant to be between the two annotation modes that FrameNet has used, i.e. *lexicographic* and *full–text* annotation. Here, two different annotation strategies have been followed to annotate frame–semantic information evoked by lexical units that convey *legal* knowledge and by those lexical units that express *world* knowledge. To be more specific, the annotation of lexical units that convey fundamental legal concepts, e.g. those units expressing deontic modalities (e.g. *proibire* 'to prohibit', *divieto* 'prohibition', *obbligare* 'to obligate', etc.), followed the lexicographic mode. That is, the annotation started from a list of lexical units belonging to the legal domain. In addition, a full–text annotation mode has been followed in the annotation of frame–semantic information conveyed by those lexical units in the regulated domain. This annotation has been done only when the lexical units were already part of a situation–type (i.e. a semantic frame) belonging to the legal domain. In fact, frame–evoking lexical units that express domain–specific world knowledge have been annotated only when they served as lexical fillers of a FE part of a legal frame.

For example, the sentence given below, where the verb *ha vietato* ('prohibited') evokes a PROHIBITING frame conveying *legal* information, has been annotated as follows:

- [*La decisione 90/200* $_{Principle}$] *ha* [***vietato*** $_{TARGET}$] [*l'esportazione dal Regno Unito di taluni tessuti e organi bovini* $_{State\_of\_affairs}$] [*solo dopo il 9 aprile 1990* $_{Time}$].

  ('[The decision 90/200 $_{Principle}$] [prohibited $_{TARGET}$] [the exportation from the United Kingdom of certain bovine tissues and organs $_{State\_of\_affairs}$] [only after the 9th of April 1990 $_{Time}$]').

The deverbal noun *esportazione* ('exportation') evokes an EXPORTING frame, conveying *world* knowledge, and was included in the FE "State of affairs" which belongs to the PROHIBITING frame. Therefore, a second annotation has been provided for that sentence as follows:

- [*La decisione 90/200* $_{Principle}$] *ha* [*vietato* $_{TARGET}$] [[***l'esportazione*** $_{TARGET}$] [*dal Regno Unito* $_{Exporting\_area}$] [*di taluni tessuti e organi bovini* $_{Goods}$] $_{State\_of\_affairs}$] [*solo dopo il 9 aprile 1990* $_{Time}$].

  ('[The decision 90/200 $_{Principle}$] [prohibited $_{TARGET}$] [[the exportation$_{TARGET}$] [from the United Kingdom$_{Exporting\_area}$] [of certain bovine tissues and organs $_{Goods}$] $_{State\_of\_affairs}$] [only after the 9th of April 1990 $_{Time}$]').

5.2.2. Domain–specific customizations issues

The annotation methodology mostly consists of maintaining and reusing the semantic frames and FEs already defined in FrameNet. However, several domain–specific customizations were needed. Three customization strategies, fully described in (Venturi, 2011a), have been followed. They differ in their increasing degree of modification to the FrameNet resource and they concern:

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 convenctions 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 semantic 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. (2006b) 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 customizations[8]. 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 $_{Permitted\_action}$].)

6. Using TEMIS

As mentioned in Section 3., a subset of TEMIS was used as test corpus in the occasion of the "Domain Adaptation" track at Evalita 2011 to test the performances of statistical parsers trained or developed on newspaper corpora. Similarly, a subset of the syntactically annotated resource described here is currently been used in the "First Shared Task on Dependency Parsing of Legal Texts" at SPLeT 2012.

In this section, the author intends to illustrate the results of first experiments devoted to use TEMIS to parse legal texts. The main aim is to quantify the impact that the language of the legislative texts included in TEMIS has on the accuracy of the parser exploited here.

Results are reported in Table 5. The DeSR parser (Attardi, 2006) has been exploited using *i)* Support Vector Machine as learning algorithm and *ii)* the ISST–TANL corpus as training corpus. The parser has been tested on a test set of 5,165 tokens extracted from the ISST–TANL corpus (ISST–TANL–test) and on a test set of 5,866 tokens extracted from

[8] The 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 TEMIS corpus (TEMIS–test). Evaluation have been carried out in terms of the standard accuracy dependency parsing measure, i.e. labeled attachment score (LAS): the percentage of tokens for which it has predicted the correct head and dependency relation.

Test set	LAS
ISST–TANL–test	79.71
TEMIS–test	74.72
TEMIS–EU–test	79.30
TEMIS–NAT–test	72.75
TEMIS–LOC–test	72.19

Table 5: Performance of the DeSR parser trained on the ISST–TANL corpus and tested on TEMIS.

As it was expected, the parser trained on the newswire domain has lower performance when tested on the legislative texts of TEMIS. The DeSR performances have a drop of 4.99 percentage points passing from a LAS of 79.71% obtained on the ISST–TANL–test to a LAS of 74.72% on the TEMIS–test.

In order to test the parser accuracy with respect to the three considered legal text sub–varieties, a further experiment has been carried out. Thus, DeSR has been tested on *i)* a test of 1,932 tokens taken from the TEMIS's sub–corpus made up of European Italian legal texts (TEMIS–EU–test), *ii)* a test of 1,971 tokens from the sub–corpus of national legal texts and *iii)* a test of 1,963 tokens from the sub–corpus of local legal texts. Interestingly, the parser has the highest performance when tested on the European legal texts. This is in line with the results of the linguistic monitoring reported in Section 3.2., where it has been demonstrated that this latter legal text sub–variety has a linguistic behaviour which is more similar to newswire texts than the national and local legal texts.

In addition, it is suggested here that the TEMIS corpus can be helpful for domain adaptation purposes. By embracing a supervised approach, it has been used to improve the performance of DeSR as domain–specific additional training data. A pilot experiment has been carried out adding a set of 9,940 tokens from TEMIS to the parser training data, i.e. the training set portion of ISST–TANL. Interestingly, the parser has an improvement of 6.66 percentage points passing from a LAS of 74.72 to a LAS of 81.38.

7. Conclusion and future work

Methodological issues concerning the design and the development of TEMIS, a syntactically and semantically annotated corpus of Italian legislative texts, have been presented and discussed. To the author knowledge, this is the first initiative aiming at building an annotated corpus of legal texts which is overtly devoted to be used for legal text processing applications based on NLP tools. Accordingly, a number of future directions of research can be foreseen.

As illustrated in Section 6., the syntactically annotated resource can be used to parse legal texts as training data of a statistical parser. In addition, it can be exploited in a supervised domain adaptation scenario to improve the performances of a parser originally trained on a different domain. Currently, it has been planning to increase the amount of sentences semantically annotated in TEMIS by *i)* annotating additional textual instances of the deontic modalities considered so far and *ii)* making explicit further information relevant for the legal domain. The latter direction of research is related to the number of foreseeable legal uses of the corpus. For example, the organization principles of the semantic annotation methodology adopted in the present work could be used to linguistically ground the logical structure of the reasoning of the factfinders in a corpus such as the Vaccine/Injury Project Corpus (Walker et al., 2011). Accordingly, an adequate ontology of frames and frame elements could be devised aiming at associating (semantic) information concerning, for example, under– or over–compensation for health injuries with their (textual) linguistic realization.

Finally, the TEMIS corpus semantically annotated can be a useful resource for several semantic processing tasks, such as Semantic Role Labeling (SRL) of legal texts. Following the strategy adopted for the ISST–TANL corpus which has been recently used as training corpus in the framework of the "Frame Labeling over Italian Texts" (FLaIT) task of Evalita 2011 (Basili et al., 2012), the frame–based information annotated in TEMIS can be exploited to train a domain–specific semantic role labeler.

8. Acknowledgements

The author would like to thank Francesco Asaro and Tommaso Petrolito who have contributed to both the syntactic and the semantic annotation process.

9. References

G. Attardi. 2006. Experiments with a multilanguage non–projective dependency parser. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X '06)*, pages 166–170, New York City, New York.

R. Bartolini, A. Lenci, S. Montemagni, V. Pirrelli, and C. Soria, 2004. *Automatic classification and analysis of provisions in legal texts: a case study*, pages 593–604. Lecture Notes in Computer Science, 3292/2004, Springer-Verlag.

R. Basili, A. Lenci, D. De Cao, and G. Venturi. 2012. Frame labeling over italian texts. In *Working Notes of EVALITA 2011*, Rome, Italy.

A. Bertoldi and R. Chishman. 2012. Frame semantics and legal corpora annotation: theoretical and applied challenges. *Linguistic Issues in Language Technology*, 7(9).

D. Biber and S. Conrad. 2009. *Register, genre, and style*. Cambridge, Cambridge University Press.

C. Bosco and A. Mazzei. 2012. The evalita 2011 parsing task: the dependency track. In *Working Notes of EVALITA 2011*, 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.

E. de Maat and R. Winkels. 2011. Formal models of sentences in dutch law. In *Proceedings of the Workshop Applying Human Language Technology to the Law (AHLT 2011)*, pages 28–40, Pittsburgh, Pennsylvania.

F. Dell'Orletta, S. Montemagni, and G. Venturi. 2011. Read-it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, Scotland.

F. Dell'Orletta, S. Marchi, S. Montemagni, G. Venturi, T. Agnoloni, and E. Francesconi. 2012. Domain adaptation for dependency parsing at evalita 2011. In *Working Notes of EVALITA 2011*, Rome, Italy.

A. Dolbey, M. Ellsworth, and J. Scheffczyk. 2006a. Bioframenet: A domain-specific framenet extension with links to biomedical ontologies. In *Proceedings of the Biomedical Ontology in Action, Workshop at KR–MED*, pages 87–94, Baltimore, Maryland.

A. Dolbey, M. Ellsworth, and J. Scheffczyk. 2006b. 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*. Ph.D. thesis, University of California, Berkeley.

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.J. Fillmore. 1985. Frame and the semantics of understanding. *Quaderni di semantica*, IV(2), dicembre: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.

B. Mortara Garavelli. 2001. *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani*. Torino, Einaudi.

D. Gildea. 2001. Corpus variation and parser performance. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2001)*, pages 167–202, Pittsburgh, PA.

B. Hachey and C. Grover. 2006. Extractive summarisation of legal texts. *Artificial Intelligence and Law*, 14(4):305–345.

J.-D. Kim, T. Ohta, and J. Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(10).

F. Kuhn. 2010. A description language for content zones of german court decisions. In *Proceedings of the 3rd Workshop on Semantic Processing of Legal Texts (SPLeT 2010)*, pages 1–7, La Valletta, Malta.

A. Lenci, S. Montemagni, G. Venturi, and M.R. Cutrullà. 2012. Enriching the isst-tanl corpus with semantic frames. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC 2012)*, page Forthcoming, Istanbul, Turkey.

K. T. Maxwell, J. Oberlander, and V. Lavrenko. 2009. Evaluation of semantic events for legal case retrieval. In *Proceedings of the Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR 2009)*, pages 39–41, Barcelona, Spain.

A. Mazzei, D. P. Radicioni, and R. Brighi. 2009. Nlp–based extraction of modificatory provisions semantics. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009)*, pages 50–57, Barcelona, Spain.

D. McClosky and E. Charniak. 2008. Self–training for biomedical parsing. In *Proceedings of the Association for Computational Linguistics (ACL 2008)*, Columbus, Ohio.

J. Nivre, J. Hall, S. Kubler, R. McDonald, J. Nilsson, S. Riedel S., and D. Yuret. 2007. The conll 2007 shared task on dependency parsing. In *Proceedings of the EMNLP–CoNLL*, pages 915–932.

T. Ohta, Y. Tateisi, and J.-D. Kim. 2002. Genia corpus: an annotated research abstract corpus in molecular biology domain. In *Proceedings of the Human Language Technology Conference (HLT 2002)*, San Diego, CA.

R. Mochales Palau and M.F. Moens. 2011. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009)*, pages 98–107, Pittsburgh, Pennsylvania.

J. Ruppenhofer, M. Ellsworth, M.R.L. Petruck, C.R. Johnson, and J. Scheffczyk. 2010. *FrameNet II: Extended Theory and Practice*. available at https://framenet.icsi.berkeley.edu/fndrupal/the_book.

P. L. Spinosa, G. Giardiello, M. Cherubini, S. Marchi, G. Venturi, and S. Montemagni. 2009. Nlp–based metadata extraction for legal text consolidation. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009)*, pages 40–49, Barcelona, Spain.

G. Venturi. 2011a. *Lingua e diritto: una prospettiva linguistico–computazionale*. Ph.D. Thesis, Università di Torino.

G. Venturi. 2011b. 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.

V. Walker, N. Carie, C. DeWitt, and E. Lesh. 2011. A framework for the extraction and modeling of fact–finding reasoning from legal decisions: Lessons from the vaccine/injury project corpus. *Artificial Intelligence and Law*, 19(4):291–331.

S. Walter, 2009. *Definition extraction from court decisions using computational linguistic technology*, pages 183–224. Berlin–New York: Mouton de Gruyter.

A. Wyner and W. Peters. 2010a. Lexical semantics and expert legal knowledge towards the identification of legal case factors. In *Proceedings of the Legal Knowledge and*

Information Systems Conference (JURIX 2010), pages 127–136, Liverpool, United Kingdom.

A. Wyner and W. Peters. 2010b. Towards annotating and extracting textual legal case factors. In *Proceedings of the 3rd Workshop on Semantic Processing of Legal Texts (SPLeT 2010)*, pages 36–45, La Valletta, Malta.

A. Wyner and W. Peters. 2011. On rule extraction from regulations. In *Proceedings of the 24th International Conference on Legal Knowledge and Information Systems (JURIX 2011)*, University of Vienna.

A. Wyner. 2010. Towards annotating and extracting textual legal case elements. In *Proceedings of the 4th Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT 2010)*, pages 9–18, Fiesole, Italy.