The SPLeT–2012 Shared Task on Dependency Parsing of Legal Texts

Felice Dell'Orletta[*], **Simone Marchi**[*], **Simonetta Montemagni**[*], **Barbara Plank**[†], **Giulia Venturi**[◇]

[*] Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa, Italy
[†] DISI, University of Trento, Italy
[◇] Scuola Superiore Sant'Anna di Studi Universitari e di Perfezionamento, Pisa, Italy
{felice.dellorletta,simone.marchi,simonetta.montemagni,giulia.venturi}@ilc.cnr.it, barbara.plank@disi.unitn.it

Abstract

The 4th Workshop on "Semantic Processing of Legal Texts" (SPLeT–2012) presents the first multilingual shared task on Dependency Parsing of Legal Texts. In this paper, we define the general task and its internal organization into sub–tasks, describe the datasets and the domain–specific linguistic peculiarities characterizing them. We finally report the results achieved by the participating systems, describe the underlying approaches and provide a first analysis of the final test results.

Keywords: Domain Adaptation, Dependency Parsing, Legal Text Processing

1. Introduction and Motivation

As overtly claimed by McCarty (2007), "one of the main obstacles to progress in the field of artificial intelligence and law is the natural language barrier". This entails that it is of paramount importance to use Natural Language Processing (NLP) techniques and tools that automate and facilitate the process of knowledge extraction from legal texts. In particular, it appears that a number of different legal text processing tasks could benefit significantly from the existence of dependency parsers reliably dealing with legal domain texts, e.g. automated legal reasoning and argumentation, semantic and cross–language legal information retrieval, document classification, legal drafting, legal knowledge discovery and extraction, as well as the construction of legal ontologies and their application to the legal domain.

Dependency parsing thus represents a prerequisite for any advanced IE application. However, since Gildea (2001) it is a widely acknowledged fact that state–of–the–art dependency parsers suffer from a dramatic drop of accuracy when tested on domains outside of the data from which they were trained or developed on. In order to overcome this problem, the last few years have seen a growing interest in developing methods and techniques aiming at adapting current parsing systems to new domains. This is testified by several initiatives organized around this topic: see, for instance, the "Domain Adaptation Track" organized in the framework of the CoNLL 2007 Shared Task (Sagae and Tsujii, 2007a), or the ACL Workshop on "Domain Adaptation for Natural Language Processing" (DANLP, 2010). In this context, a particularly relevant initiative is represented by the "Domain Adaptation Track" (Dell'Orletta et al., 2012) organized in the framework of the third evaluation campaign of Natural Language Processing and Speech tools for Italian, Evalita–2011[1], where participants were asked to adapt their dependency parsing systems to the legal domain.

With the only exception of the Evalita–2011 "Domain Adaptation Track" whose results provided relevant feedback in this direction (unfortunately circumscribed to the Italian language), so far very few attempts have been carried out to quantify the performance of dependency parsers on legal texts (e.g. law or case law texts). Among the reasons behind this lack of attention is the unavailability of gold corpora of legal texts annotated with syntactic information with respect to which such an evaluation could be carried out. To our knowledge, exceptions exist only for German and Italian (as mentioned above). The first is the case of the corpus including 100 sentences taken from German court decisions and syntactically manually annotated, as described by Walter (2009). However, this corpus is currently encoded following the PReDS parser (Braun, 2003) native annotation format; its exploitation for the evaluation of dependency parsers would require the conversion of the native PReDS annotation into some kind of standard representation format (e.g. CoNLL).

For the Italian language two different annotated corpora exist: *i)* 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 tokens; 1,100 sentences) annotated with syntactic dependency information and *ii)* TEMIS (Venturi, 2012), a corpus of legislative texts (15,804 tokens; 504 sentences) enacted by three different releasing agencies (i.e. European Commission, Italian State and Piedmont Region) and regulating a variety of domains which is annotated with syntactic and semantic information. Interestingly, the two corpora represent two different sub–varieties of the Italian legal language. According to one of the main Italian scholars of legal language Garavelli (2001), the Civil Law Code articles are less representative of the much cited linguistic complexity of the so–called Italian *legalese* (i.e. the variety of Italian used in the legal domain) with respect to other kinds of legislative texts such as laws, decrees, regulations, etc. This is confirmed 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 the newspapers test corpus. Further evidence in the same

[1] http://evalita.fbk.eu/index.html

[2] http://www.di.unito.it/∼tutreeb/

direction emerged within the the Evalita–2011 "Domain Adaptation Track" (Dell'Orletta et al., 2012), where a subset of TEMIS was used: it turned out that parsing systems need to be adapted to reliably analyse legal texts such as laws, decrees, regulations, etc.

Following these premises, the shared task organised in the framework of the 4th Workshop on "Semantic Processing of Legal Texts" (SPLeT–2012) on dependency parsing of legal texts was aimed at: providing common and consistent task definitions and evaluation criteria in order to identify the specific challenges posed by the analysis of this type of texts across different languages; obtaining a clearer idea of the current performance of state–of–the–art parsing systems; and last but not least, developing and sharing multilingual domain–specific resources.

2. Definition of the Task

The shared task was organised into two different subtasks as described below:

1. **Dependency Parsing**: this represents the basic and mandatory subtask, focusing on dependency parsing of legal texts, aimed at testing the performance of general parsing systems on legal texts;
2. **Domain Adaptation**: this is a more challenging (and optional) subtask, focusing on the adaptation of general purpose dependency parsers to the legal domain, aimed at investigating methods and techniques for automatically extracting knowledge from large unlabelled target domain corpora to improve the performance of general parsing systems on legal texts.

The languages dealt with are English and Italian. Evaluation has been carried out in terms of standard accuracy dependency parsing measures, i.e. labeled attachment score (LAS) including punctuation, with respect to a test set of texts from the legal domain.

3. Datasets

For both languages, different datasets have been distributed. For the source domain, task participants have been provided with *i)* a training set exemplifying general language usage and consisting of articles from newspapers and *ii)* a manually annotated development set, also including labeled dependency relations. For the target domain, they have been supplied with *i)* a target corpus including automatically generated sentence splitting, tokenization, morpho–syntactic tagging and lemmatization, and *ii)* a development set, as for the source domain.

All distributed data adhere to the CoNLL 2007 tabular format used in the Shared Task on Dependency Parsing (Nivre et al., 2007) and they are described in detail in the following two sections.

Note that whereas for both English and Italian the final test set is represented by legislative texts enacted by the European Commission (namely, the English and Italian version of the same texts), the domain of development corpora is different for the two languages: for English the development corpora are represented by biomedical abstracts, for Italian they include legal texts belonging to a different sub–variety of the legal language. This is in line with the experimental setup defined for the "Domain Adaptation Track" of the CoNLL 2007 Shared task, where participants were provided with biomedical abstracts as development data, and chemical abstracts and parent–child dialogues as two separate sets of test data.

3.1. Italian Dataset

For the Italian language, the source domain data is drawn from a corpus of news, i.e. the ISST–TANL corpus jointly developed by the Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR) and the University of Pisa, exemplifying general language usage and consisting of articles from newspapers and periodicals, selected to cover a high variety of topics (politics, economy, culture, science, health, sport, leisure, etc.). This corpus has already been used in the Evalita–2011 "Domain Adaptation Track" (Dell'Orletta et al., 2012). Two different datasets have been distributed to participants: a training corpus (hereafter referred to as *it_isst_train*) of 71,568 tokens and 3,275 sentences and a test corpus (hereafter referred to as *it_isst_test*) of 5,175 tokens (231 sentences).

As target domain data, two different sets have been distributed:

1. a set used as development data drawn from an Italian legislative corpus, gathering laws enacted by Italian State and Regions and regulating a variety of domains (ranging from environment, human rights, disability rights to freedom of expression), articulated as follows:
   (a) a corpus of 13,095,574 tokens and 660,293 sentences automatically splitted, tokenized, morpho–syntactic tagged and lemmatized;
   (b) a manually annotated test set, also including labeled dependency relations, consisting of 5,194 tokens and 118 sentences (hereafter referred to as *it_NatRegLaw*);
2. a set used as test data drawn from an Italian legislative corpus, gathering laws enacted by European Commission and regulating a variety of domains (ranging from environment, human rights, disability rights to freedom of expression), articulated as follows:
   (a) a corpus of 28,263,250 tokens and 1,300,451 sentences automatically splitted, tokenized, morpho–syntactic tagged and lemmatized;
   (b) a manually annotated test set, i.e. sentence–splitted, tokenized, morpho–syntactically tagged and lemmatized, consisting of 5,662 tokens and 241 sentences (hereafter referred to as *it_gold_EULaw*).

The source and target domain data are annotated according to the morpho–syntactic[3] and dependency[4] tagsets 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[5].

3.1.1. Source vs Target Domain Corpora Annotation Criteria

Note that in order to properly handle legal language peculiarities, annotation criteria have been extended to cover domain–specific constructions. The specializations are concerned with both sentence splitting and dependency annotation.

For sentence splitting, in the target domain corpora sentence splitting is overtly meant to preserve the original structure of the law text. This entails that also punctuation marks such as ‘;’ and ‘:’, when followed by a carriage return, are treated as sentence boundary markers.

For what concerns dependency annotation, it should be considered that legal texts are characterized by syntactic constructions hardly or even never occurring in the source domain corpora. In order to successfully cope with such peculiarities of legal texts, dependency annotation criteria have been extended to cover the annotation of *a)* elliptical constructions, *b)* participial phrases as well as *c)* long distance dependencies resulting in non–projective links, to mention only a few. All these peculiar constructions have been explicitly represented in the development and final test sets.

3.2. English Dataset

For the English language, the source domain data is represented by the training and test data distributed in the CoNLL 2007 Shared Task. The two sets of data were extracted from the Penn Treebank (PTB)[6] which consists, according the description provided by the Linguistic Data Consortium[7], of 2,499 stories selected from a three year Wall Street Journal (WSJ) collection of 98,732 stories for syntactic annotation. In more detail, the distributed training set (hereafter referred to as *english_ptb_train*) includes sections 02–11 of the WSJ and is a corpus of 446,573 tokens and 18,577 sentences; the test set (hereafter referred to as *english_ptb_test*) is a subset of section 23 of the WSJ for a total amount of 5,003 tokens and 214 sentences.

As target domain data, two different sets have been distributed:

1. a development data set, including the files used for the final testing of the systems in the “Domain Adaptation Track” of the CoNLL 2007 Shared task, namely:
   (a) a corpus of chemical abstracts (CHEM corpus) of 10,482,247 tokens and 396,128 sentences automatically splitted, tokenized, morpho–syntactic tagged and lemmatized;
   (b) a manually annotated test set, also including labeled dependency relations, consisting of 5,001 tokens and 195 sentences (hereafter referred to as *english_pchemtb*);
2. a test data set, drawn from an English legislative corpus gathering laws enacted by the European Commission and regulating a variety of domains (ranging from environment, human rights, disability rights to freedom of expression), articulated as follows:
   (a) a corpus of 25,942,241 tokens and 1,260,621 sentences automatically splitted, tokenized, morpho–syntactically tagged and lemmatized;
   (b) a manually annotated test set, i.e. sentence–splitted, tokenized, morpho–syntactically tagged and lemmatized, consisting of 5,621 tokens and 214 sentences (hereafter referred to as *en_gold_EULaw*).

The source and target data are annotated according to the PTB[8] morpho–syntactic[9] and dependency tagsets.

3.2.1. Source vs Target Domain Corpora Annotation Criteria

The legal text contains pecularities regarding surface characteristics as well as dependency annotations that are hardly if at all present in the newspaper source domain data.

With regard to sentence splitting the same criteria were used as for Italian: in order to preserve the original structure of the law text, punctuation marks such as semicolon and colon that are followed by a carriage return are treated as sentence boundary markers. If no carriage return was present in the original text, the sentence was kept as is, thus resulting in some relatively long sentences. An example thereof is given in Figure 1. It will also serve as example to discuss some of the adopted annotation criteria. For instance, subsequent subordinate clauses without the main clause that are not present in the in–domain data, i.e. the subordinates introduced by *whereas* in our example. In this case, we chose to annotate the first instance of *whereas* as the *ROOT* node of the sentence and the second one as verbal modifier of the head of the preceding clause.

As will be shown further in Section 4., sentence length deviates considerably between the source and target domains. Another surface property of the target domain text that is different from the source domain is that the legal text contains a large amount of enumerations (lists, either hyphenated or enumberated with characters, numbers or roman numerals). In fact, one third (72 of 214) sentences in *en_gold_EULaw* are list items. Only very few of them (less

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

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

[5] http://medialab.di.unipi.it/wiki/SemaWiki

[6] http://www.cis.upenn.edu/~treebank/

[7] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99T42

[8] The head and dependency relation fields were converted using the algorithms described in (Johansson and Nugues, 2007).

[9] The fine grained part–of–speech are the gold standard part of speech tags from the WSJ, details of which can be found, http://bulba.sdsu.edu/jeanette/thesis/PennTags.html or http://www.cis.upenn.edu/~treebank/

(13) Whereas Member States should be able to require that a prior consultation be undertaken by the party that intends to bring an action for an injunction, in order to give the defendant an opportunity to bring the contested infringement to an end ; whereas Member States should be able to require that this prior consultation take place jointly with an independent public body designated by those Member States ;

Figure 1: Example sentence from *en_gold_EULaw*.

than half a percent of the sentences in the PTB) are present in the source domain training data. We tried to treat the enumeration part consistently: if followed or surrounded by hyphens or parenthesis (like *(13)* in Figure 1), the list item marker was considered the head of the punctuation marks and attached to the head of the following sentence or phrase (as VMOD or NMOD). Moreover, since the PTB part–of–speech tags contains a respective "list item marker" tag (LS), the POS tags were tagged as such, accordingly. If the list item ended with a semicolon followed by a single conjunction (e.g. *; or*), it was attached as DEP (unclassified relation) to the head of he preceding clause. Further pecularities of the target domain (like the depth of embedded complement chains) are discussed in more detail in the following section.

3.3. Linguistic Preprocessing of Datasets

Both English and Italian datasets used for development and final testing have been morpho–syntactically tagged and lemmatized by a customized version of the pos–tagger described in Dell'Orletta (2009).
The manually annotated test sets were initially parsed by the DeSR parser (Attardi and Dell'Orletta, 2009), a state–of–the-art linear–time Shift–Reduce dependency parser, and were then manually revised by expert annotators, also on the basis of the extended annotation criteria reported in Sections 3.1.1. and 3.2.1. for Italian and English respectively.

4. Source vs Target Domain Data: Linguistic Features

In order to get evidence of the differences among the source and target domain data, the Italian and English distributed gold datasets have been monitored with respect to a number of different linguistic parameters. This allowed us to empirically *i)* define what we mean by *domain* and *ii)* to explain the drop of accuracy of general parsers on domain–specific texts and thus to motivate the need for developing domain adaptation strategies for reliably parsing of legal texts. As demonstrated by the results of the linguistic monitoring reported in the following sections, the two different legal language sub–varieties as well as the chemical and newswire texts each represent different classes of texts, henceforth generically referred to as *domains*, each characterized by specific linguistic features. The typology of features selected to reconstruct the linguistic profile characterizing each class of texts is organised into four main categories: raw text features, lexical features, morpho-syntactic and syntactic features. In what follows, we report and discuss the monitoring results obtained with respect to these different textual classes or domains.

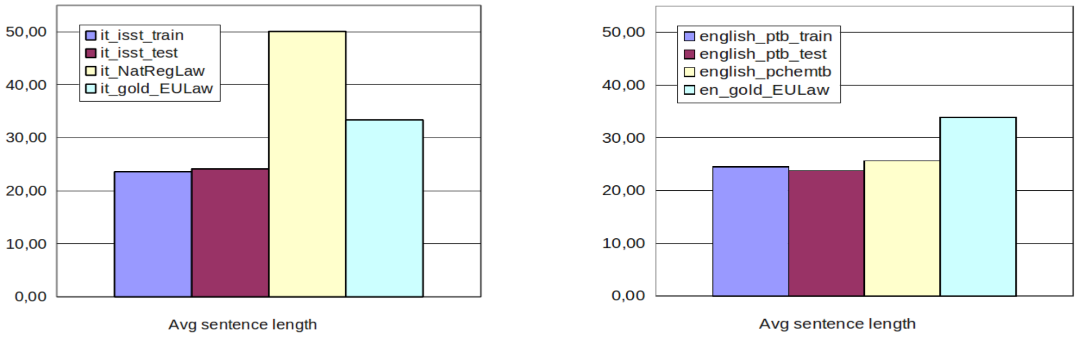
Raw Text Features

The source domain and legal datasets for both Italian and English differ significantly in many aspects starting from the **average sentence length**, calculated as the average number of words per sentence[10] (see Figure 2). As Figure 2(a) shows, *it_NatRegLaw* contains the longest sentences with respect to all the other datasets. Interestingly, the sentence lengths of *it_gold_EULaw* and *en_gold_EULaw* sets are very close, i.e. 33.38 and 33.86 word–tokens respectively. This is mainly due to the fact that the two sets contain aligned sentences as well as to the nature of European legal texts, i.e. their being translations of an original unique text. It is also worth noting that the length of the sentences contained in *english_pchemtb* is closer to *english_ptb_train* and *english_ptb_test* than to *en_gold_EULaw*. This supports the hypothesis that chemical texts represent a different domain with respect to English European legislative texts.
Since, as claimed in the literature on measures of syntactic complexity (see below), a longer sentence is grammatically more complex than a shorter one, it can be argued that sentence length affects parsing accuracy. This is typically the case when such a feature is associated with long dependency links, as demonstrated by McDonald and Nivre (2007).

Lexical and Morpho–syntactic Features

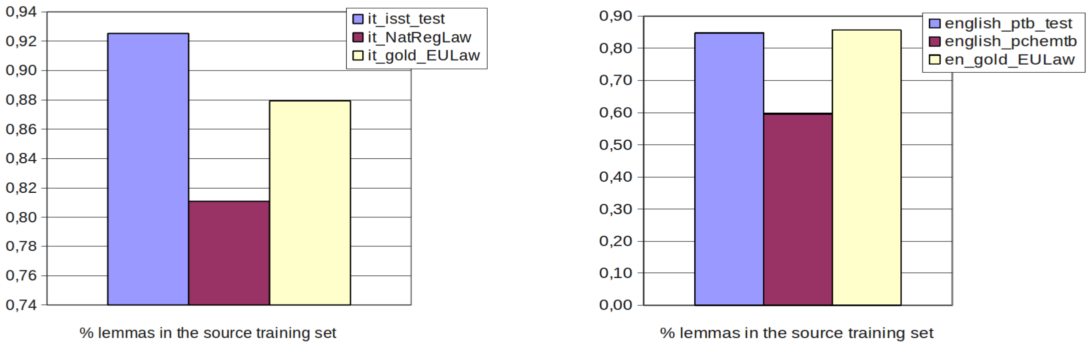
Figure 3 reports the lexical overlap of the different corpora, calculated as the percentage of lexical items of *it_isst_train* and *english_ptb_train* also contained in the target domain test sets. First of all, it is worth noting that as far as *english_pchemtb* is concerned the percentage of newswire lexicon (0.60%) is lower than in *en_gold_EULaw* (0.86%). This allows highlighting a peculiarity of legal domain texts which contain a higher percentage of newswire lexicon than other domains. This finding is in line with what observed by Lease and Charniak (2005), who report the unknown word rate (expressed in terms of tokens) for various technical domains (e.g. biomedical abstracts, abstracts in the field of aereodynamics, etc.) which has been computed with respect to sections 2–21 of the WSJ.
As it can be seen in Figures 3(a) and 3(b), the lexicon specific to the legal domain is not extremely different from the one of the newswire domain. Interestingly, this holds true both for the Italian and English legal language used in texts enacted by the European Commission. This suggests that the main differences between newswire and legal texts are mostly concerned with the underlying syntactic structure. Nevertheless, a difference between the two considered Italian legal language sub–varieties exists: the percentage of newswire lexicon contained in *it_NatRegLaw* (0.81%) is lower than the one observed in *it_gold_EULaw* (0.88%).
A last remark is in order here for what concerns the percentage of lexical items that the *it_isst_test* and *english_ptb_test* share with the corresponding training sets: the lexicon of the Italian test set turned out to be much more similar

[10] Note that sentence shorter than 5 word–tokens are excluded from the computation.



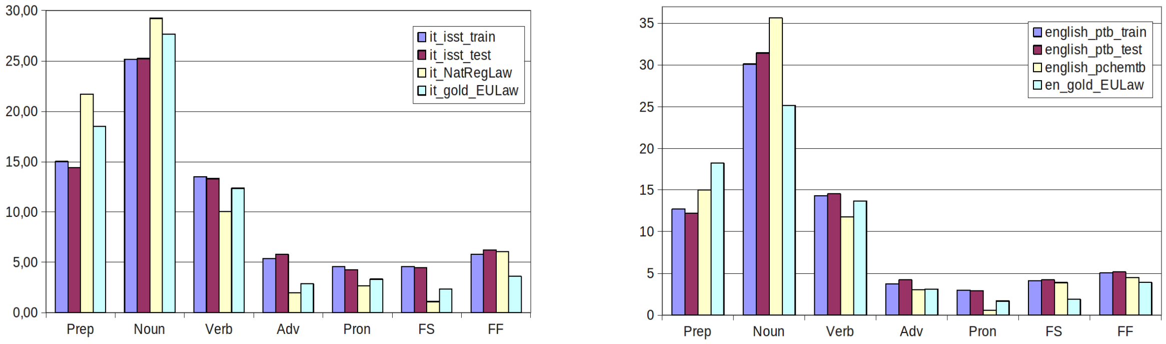
(a) Italian gold data (b) English gold data

Figure 2: Average sentence length in the Italian and English gold datasets.



(a) Italian gold data (b) English gold data

Figure 3: % of training set lemmas contained in the Italian and English gold datasets.



(a) Italian gold data (b) English gold data

Figure 4: Distribution of some of the main parts–of–speech in the Italian and English gold datasets.

(0.93%) to *it_isst_train* than the lexicon of *english_ptb_test* (0.85%) with respect to *english_ptb_train*. This follows from the strategy adopted for selecting the sentences contained in the test set: the sentences of *it_isst_test* have been randomly selected from the whole ISST–TANL corpus, while those in *english_ptb_test* have been taken from a section of the Penn Treebank different from the one included in *english_ptb_train*.

Let us focus now on the morpho–syntactic level. Figure 4 reports that different varieties of the legal language represented by *it_NatRegLaw*, *it_gold_EULaw* for Italian and *en_gold_EULaw* for English show a similar distribution of parts–of–speech: namely, they all have a higher percentage of prepositions (*Prep*) with respect to the ISST–TANL and PTB datasets, and a lower percentage of **verbs** (*Verb*), **adverbs** (*Adv*), **pronouns** (*Pron*), **punctuation marks**, i.e. full stops (*FS*) and commas (*FF*). These observed distributions can be taken as some of the main peculiar features of both Italian and English legal texts.

While the different distribution of punctuation marks can support the hypothesis of a sentence structure specific to legal texts, the high occurrence of prepositions can be strongly connected with their presence within long sequences of complements (see below for more details).

Surprisingly enough, the percentage distribution of **nouns** (*Noun*) is quite different across languages, i.e. in *it_gold_EULaw* and *en_gold_EULaw*. Similarly to *it_NatRegLaw*, the Italian European legal texts contain a higher percentage of nouns with respect to the ISST–TANL datasets. On the contrary, the occurrences of nouns in *en_gold_EULaw* are fewer than in the PTB data.

Syntactic Features

Major differences hold at the level of considered syntactic features, for which we observe a peculiar distribution which characterizes legal texts with respect to the source domain as well as to the other target domain datasets.

The first monitored syntactic feature is concerned with the **average depth of embedded complement 'chains'** governed by a nominal head and including either prepositional complements or nominal and adjectival modifiers. Figures 5(a) and 5(b) show that both the Italian and English European legal texts are characterized by an average depth which is higher than the one observed in the ISST–TANL and PTB source domain datasets. This represents the syntactic counterpart of the peculiar distribution of prepositions observed in legal texts at the morpho–syntactic level (see above). Interestingly, the difference holding between the average depth of complement 'chains' occurring in *english_pchemtb* and the one observed with respect to the PTB dataset is less sharp than the difference between *en_gold_EULaw* and newswire PTB data. This demonstrates that the occurrence of deep embedded complement 'chains' appears to be a syntactic feature characterizing the legal domain with respect to newswire texts as well as to other domains. In Italian, this domain–specific feature appears to be more marked in the legal language sub–variety represented by *it_NatRegLaw*, which shows the deepest complement 'chains'.

A further distinguishing feature of legislative texts, still connected with the previous one, appears to be the different percentage distributions of embedded complement 'chains' by depth. As Figures 5(c) and 5(d) show, Italian and English legislative texts appear to have *i)* a lower occurrence of 'chains' including just one complement and *ii)* a higher percentage of deep complement 'chains' with respect to newswire data. Notably, *it_NatRegLaw* contains chains up to 9 embedded complements long.

It goes without saying that these two features can have a strong impact on the performances of parsers trained on the syntactic distributions of newswire texts.

The considered gold datasets have also 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 (with the exception of the punctuation marks), 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. It has been chosen to monitor these two features since they both can be indicative of the structural complexity of a dependency structure. If on the parsing side McDonald and Nivre (2007) report that statistical parsers have a drop in accuracy when analyzing long distance dependencies, on the other hand Lin (1996) and Gibson (1998) claim that the syntactic complexity of sentences can be predicted with measures based on the length of dependency links, given the memory overhead imposed by very long distance dependencies. Parse tree depth is another feature reflecting sentence complexity as stated by, to mention only a few, Yngve (1960), Frazier (1985) and Gibson (1998).

As it can be seen in Figure 6, *i)* Italian and English legislative texts contain much longer (on average) dependency links than newswire texts and *ii)* the average height of *it_gold_EULaw* and *en_gold_EULaw* parse trees is higher than in the case of ISST–TANL and PTB. In addition, as it was previously pointed out, *it_NatRegLaw* texts appear to be syntactically more distant from newswire texts than European legal texts (see Figure 6(a)).

Finally, we compared source and target domain data with respect to the **arity of verbal predicates**, calculated as the number of instantiated dependency links sharing the same verbal head (covering both arguments and modifiers). A low arity value seems to be a distinctive feature of both Italian and English legal texts in comparison with newswire texts (see Figure 7). As Figure 7(a) shows, *it_NatRegLaw* contains verbal predicates characterized by the lowest arity. As suggested by Venturi (2011), this distinguishing feature of legal texts can be due to the frequent occurrence of verbal participial forms and of elliptical constructions.

5. Participation Results

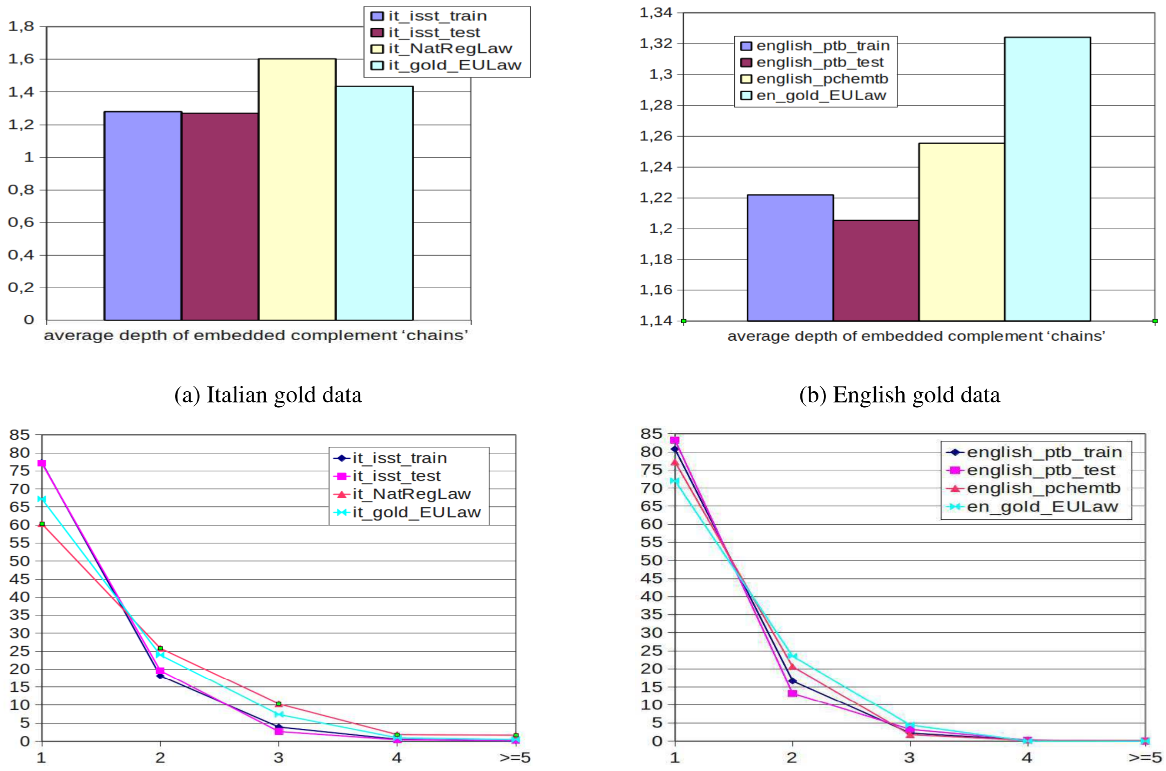
The participants to the shared task were three, namely **Attardi_et_al.** (University of Pisa, Italy), **Mazzei_Bosco** (University of Turin, Italy) and **Nisbeth_Søgaard** (University of Copenhagen, Denmark). Whereas the latter two teams participated only in the basic **Dependency Parsing** (DP) subtask for the Italian language, the first participant presented results for both languages and for both DP and **Domain Adaptation** (DA) subtasks.

5.1. Base Parsing Models

All participants adopted ensemble–based systems in which several base parsers produce dependency trees, which are then combined using different weighting functions (to weigth each dependency arc) and different combination algorithms.

Attardi_et_al. used a combination strategy exploiting the approximate linear time combination algorithm described by Attardi and Dell'Orletta (2009). The combined parsers are three different configurations of *DeSR* (Attardi, 2006), which is a Shift/Reduce deterministic transition–based parser that by using special rules is able to handle non–projective dependencies in linear time complexity. The configurations are: two versions differing with respect to the used learning algorithm (MultiLayer Perceptron (MLP) vs Support Vector Machine (SVM)) of the two stage Reverse Revision parser (i.e. a stacked righ-to-left parser that uses hints produced by a first pass left–to–right parser, Attardi and Dell'Orletta (2009)), and a right–to–left parser using an MLP classifier.

Mazzei_Bosco used a combination strategy based on a simple voting approach: for each word of the sentence the algorithm assigns the dependency head and dependency label more voted from the combined parsers and in the case



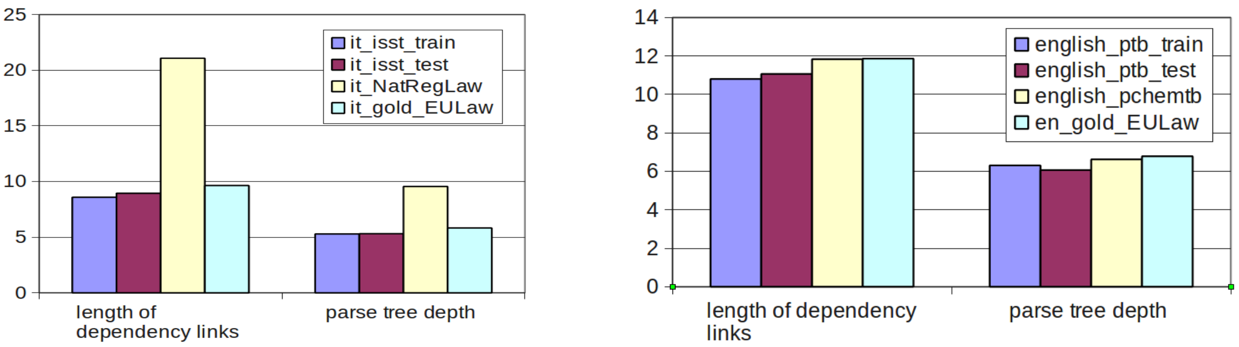
(a) Italian gold data

(b) English gold data

(c) Italian gold data

(d) English gold data

Figure 5: Average depth of embedded complement 'chains' (first row) and their distribution by depth (second row) in the Italian and English gold datasets.



(a) Italian gold data

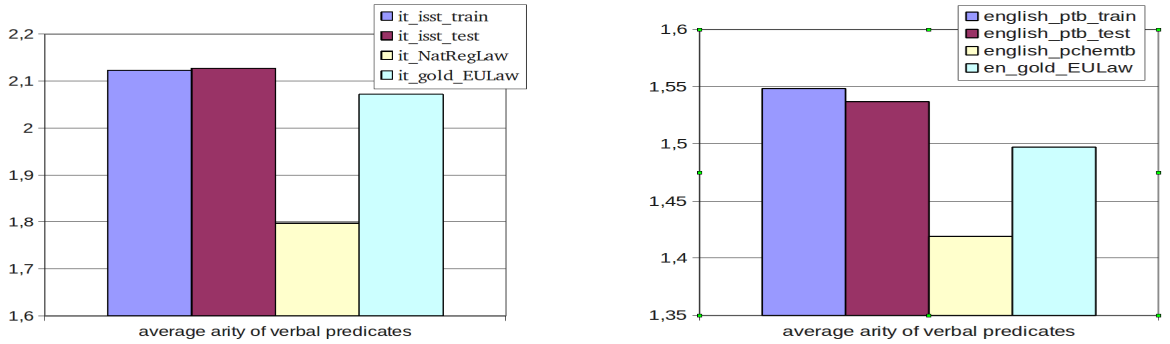
(b) English gold data

Figure 6: Length of dependency links and parse tree depth in the Italian and English gold datasets.

that each parser assigns a different dependency, the algorithm selects the dependency assigned by the best parser. Whenever in the resulted dependency structure there are cycles, the algorithm selects the tree produced by the best parser. Three different parsers are combined: *i)* left–to–right DeSR, using MLP as learning algorithm; *ii)* MaltParser (Nivre et al., 2006), a Shift/Reduce transition–based parser composed by a nondeterministic transition system for mapping sentences to dependency trees and a classifier that predicts the next transition for every possible system configuration (SVM was used as learning algorithm); *iii)* MateParser (Bohnet, 2010), an efficient implementation of the second order maximum spanning tree dependency parsing algorithm of Carreras (Carreras, 2007). The parser is trained using the margin infused relaxed algorithm (MIRA) (McDonald et al., 2005) and combined with a hash kernel (Shi et al., 2009).

Nisbeth_Søgaard adopted the combination strategy introduced by Sagae and Lavie (2009): using the analyses generated by the component parsers and a weighting function, a weighted directed graph is created where each word in the sentence is a node; finally, a maximum spanning tree algorithm is used to select the final analysis. To produce this combination they used MaltBlender software[11]. The ensemble system is based on several unoptimised parsers: *i)*

[11] w3.msi.vxu.se/users/jni/blend/



(a) Italian gold data

(b) English gold data

Figure 7: Average arity of verbal predicates in the Italian and English gold datasets.

ten instances of the MaltParser, one for each of the learning algorithms it provides; *ii)* one istance of the MateParser; *iii)* two istances (projective and non–projective) of MST-parser (McDonald et al., 2006), i.e. a graph–based parser which uses a maximum spanning tree algorithm for finding the highest scoring tree.

5.2. Results of the Dependency Parsing Subtask

Table 1 reports the results achieved by the participating systems on both the development in–domain test set (*it_isst_test*) and the out–domain test set (*it_gold_EULaw*) for the Dependency Parsing subtask for the Italian language. Unexpectedly, two out three participant parsing systems do not show a drop of accuracy when tested on the European legal texts. Interestingly, Mazzei_Bosco has an increment of 0.72 percentage points when their system was tested on the legal texts with respect to the newswire test.
This can be due to two main reasons. On the one hand, as already demonstrated by the results reported in (Sagae and Tsujii, 2007b), ensemble parsing systems are less affected by a drop of accuracy when tested on out–domain data in a domain adaptation scenario than single parsing systems, in particular when the types of parsing algorithms involved in the combination are different. On the other hand, as shown in Section 4., European legal texts are characterised by lexical, morpho–syntactic and syntactic features which make them not so distant from in–domain data.
On the contrary, the peculiar statistical distribution of monitored linguistic features in *national and regional Italian legal texts* (see Section 4.) can be seen as underlying the drop of accuracy of participant systems when tested on the out–domain development data provided (i.e. *it_NatRegLaw*), as reported in Table 2.

System	*it_isst_test*	*it_gold_EULaw*
Mazzei_Bosco	82.36	83.08
Attardi_et_al.	82.90	81.93
Nisbeth_Søgaard	81.43	81.58

Table 1: LAS for Dependency Parsing subtask for the Italian language.

System	*it_NatRegLaw*
Mazzei_Bosco	75.88
Attardi_et_al.	74.03
Nisbeth_Søgaard	75.55

Table 2: LAS of participants on national and regional Italian legal texts.

Table 3 reports the Dependency Parsing results by Attardi_et_al. on both the development in–domain test set (*english_ptb_test*) and the out–domain test set (*en_gold_EULaw*) for the English language. Differently from Italian, for English we observe a noticeable drop of accuracy, of nearly 10 LAS percentage points. Different reasons can be seen as underlying this state of affairs. Among them, it is worth mentioning the occurence of syntactic structures specific to European legal texts and never occurring in the PTB test set for which new annotation criteria had to be defined (see Section 3.2.1.) and which can hardly be learned by a statistical parser trained on PTB. Moreover, the freer word order of Italian with respect to English can help explaining why statistical variations between the in– and out–domain texts might have a deeper impact on parser performances for English than for Italian: this is just an initial intuition which should be explored in more detail.

System	*english_ptb_test*	*en_gold_EULaw*
Attardi_et_al.	88.81	78.90

Table 3: LAS for Dependency Parsing subtask for the English language.

5.3. Results of the Domain Adaptation Subtask

For this subtask, Attardi_et_al. used a method based on active learning. They followed a two–step incremental process where each step generates a new training corpus including manually revised dependency–annotated sentences from the out–domain test unlabelled corpus. Each step can be summarised as follows: a) DeSR with MLP (Multi Layer Perceptron Algorithm) as learning algorithm is used to parse the unlabeled target corpus; b) perplexity measures based on the overall likelihood of the analysis of each sentence provided by DeSR are exploited to identify 100 sentences with the highest perplexity (*Lowest Likelihood*, LLK); and c) sentences selected during the previous step are manually revised and used to extend the training corpus in order to build a new parser model.

The new parser model was used to parse the target domain test set. For the last run they used the parser system described in section 5.1..

System	*it_isst_test*	*it_gold_EULaw*
Attardi_et_al.–run1		82.78
Attardi_et_al.–run2	82.05	83.52

Table 4: LAS for Domain Adaptation subtask for the Italian language.

System	*english_ptb_test*	*en_gold_EULaw*
Attardi_et_al.–run1	87.17	78.38

Table 5: LAS for Domain Adaptation subtask for the English language.

Tables 4 and 5 report the results achieved within the Domain Adaptation subtask for Italian and English respectively. *Attardi_et_al.–run1* and *Attardi_et_al.–run2* refer to the first and second step of the active learning process. For Italian, we can observe that the adopted domain adaptation strategy shows a significant parsing improvement: the parser shows a LAS improvement of 0.85 percentage points after the first added 100 sentences, and of 1.59 points after the second step. For English, the same DA strategy does not produce the same effect. After the active learning process, the parser has a drop of accuracy of 0.52 LAS percentage points. Among the reasons behind this drop there may be disalignments between gold annotations based on the new annotation criteria defined for dealing with legal texts (as discussed in Section 5.2.) and annotations performed by the annotators involved in the active learning process.

Tables 4 and 5 also report the results obtained for the in–domain development sets after the domain adaptation process: a small drop of accuracy can be observed. This is in line with what observed by McClosky et al. (2010) and Plank and van Noord (2011) who proved that parsers trained on the union of more than one different gold corpora taken from different domains achieve lower accuracy with respect to the same parsers trained on data belonging to a single target domain.

6. Conclusion

The SPLet 2012 shared task was the first competition on dependency parsing of legal texts. In this context, different parsing systems – all based on ensemble methods – have been tested against Italian and English legal data sets.

Different results have been achieved for the two languages. A significant drop in accuracy has been observed with respect to the English test set. Differently, for Italian two out of three participant systems showed no drop in accuracy against the final test set represented by European legal texts; however, the performance of all participant systems appear to significantly decrease when tested against texts belonging to the language sub–variety represented by national and regional legislative texts. This asymmetric behaviour of parsers can be explained by comparing the statistical distribution of linguistic features within in–domain training corpora and out–domain test sets. All participants used statistical parsers based on machine learning algorithms: this fact can help explaining why their performance decreases when parsing sentences characterized by features hardly or never occurring in the training set.

This prompts the need for domain adaptation strategies. In this shared task, only one system participated in the Domain Adaptation subtask by exploiting an active learning method which achieved good results for the European Italian legal texts. On the contrary, no improvement has been obtained for what concerns European English legal texts: this is very likely due to both language–specific peculiarities and annotation choices adopted to handle domain–specific syntactic constructions.

The SPLET 2012 Shared Task was successful in defining and analysing the stat–of–the–art performance of dependency parsing in the legal domain. The evaluation results of the final submissions for both subtasks from the participants are both promising and encouraging for the future of legal Information Extraction applications. Developed domain–specific annotated corpora together with descriptions of participant systems represent rich resources for finding directions for improvements. Last but not least, the experience of the shared task provides valuable input for facing further challenges specific to the domain.

7. Acknowledgments

The Shared Task Organizers would like to thank to the CoNLL 2007 Shared Task Organizers and the Linguistic Data Consortium for preparing and providing the in–domain and development English datasets. We also wish to acknowledge Tommaso Petrolito and Francesco Asaro for their contribution in annotating the Italian target domain data.

8. References

G. Attardi and F. Dell'Orletta. 2009. Reverse revision and linear tree combination for dependency parsing. In *Proceedings of NAACL-HLT*.

G. Attardi. 2006. Experiments with a multilanguage non-projective dependency parser. In *Proceedings of the Shared task CoNLL-X*, pages 166–170, New York City.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China.

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

C. Braun. 2003. Parsing german text for syntactico–semantic structures. In *Prospects and Advances in the Syntax/Semantics Interface, Proceedings of the Lorraine–Saarland Workshop*, Nancy, France.

Xavier Carreras. 2007. Experiments with a higher-order projective dependency parser. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, page 957961.

DANLP. 2010. *Proceedings of the Workshop on Domain Adaptation for Natural Language Processing (2010)*. available at http://aclweb.org/anthology/W/W10/W10-2600.pdf.

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.

F. Dell'Orletta. 2009. Ensemble system for part-of-speech tagging. In *Proceedings of Evalita'09*, Reggio Emilia.

L. Frazier, 1985. *Syntactic complexity*. Cambridge University Press, Cambridge, UK.

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

E. Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

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

Richard Johansson and Pierre Nugues. 2007. *Extended Constituent-to-Dependency Conversion for English*. available at http://dspace.utlib.ee/dspace/bitstream/handle/10062/2560/reg-Johansson-10.pdf.

M. Lease and E. Charniak. 2005. Parsing biomedical literature. In *Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP-05)*, pages 58–69, Jeju Island, Korea.

D. Lin. 1996. On the structural complexity of natural language sentences. In *Proceedings of COLING 1996*, pages 729–733.

L.T. McCarty. 2007. Deep semantic interpretations of legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law (ICAIL2007)*, pages 217–224, Stanford, California.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28–36, Los Angeles, California.

R. McDonald and J. Nivre. 2007. Characterizing the errors of data-driven dependency parsing models. In *Proceedings of the EMNLP–CoNLL*, pages 122–131.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Association for Computational Linguistics*.

R. McDonald, K. Lerman, and F. Pereira. 2006. Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of CoNLL*, New York City.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. Maltparser: a data-driven parser-generator for dependency parsing. In *Proceedings of LREC-200*, pages 2216–2219.

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.

Barbara Plank and Gertjan van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1566–1576, Portland, Oregon.

Kenji Sagae and Alon Lavie. 2009. Parser combination by reparsing. In *Proceedings of HLT-NAACL*.

K. Sagae and J. Tsujii. 2007a. Dependency parsing and domain adaptation with lr models and parser ensembles. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL*, pages 1044–1050, Prague.

Kenji Sagae and Junichi Tsujii. 2007b. Dependency parsing and domain adaptation with lr models and parser ensemble. In *Proceedings of the EMNLP–CoNLL 2007*, pages 1044–1050.

Qinfeng Shi, JamesPetterson Gideon Dror, John Langford, Alex Smola, and S.V.N. Vishwanathan. 2009. Hash kernels for structured data. *Journal of Machine Learning*, 15(1):143–172.

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

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)*, Istanbul, Turkey.

S. Walter. 2009. Definition extraction from court decisions using computational linguistic technology. In G. Grewendorf and M. Rathert, editors, *Formal Linguistics and Law*, pages 183–224. Mouton de Gruyter.

V. H.A. Yngve. 1960. A model and an hypothesis for language structure. In *Proceedings of the American Philosophical Society*, pages 444–466.