A treebank-based study on the influence of Italian word order on parsing performance

Anita Alicante[1], Cristina Bosco[2], Anna Corazza[1], Alberto Lavelli[3]

(1) Dipartimento di Scienze Fisiche, Università "Federico II" di Napoli, Italy
(2) Dipartimento di Informatica, Università di Torino, Italy
(3) HLT Research Unit, Fondazione Bruno Kessler, Povo (Trento), Italy
anita.alicante@unina.it, bosco@di.unito.it, corazza@na.infn.it, lavelli@fbk.eu

Abstract

The aim of this paper is to contribute to the debate on the issues raised by Morphologically Rich Languages, and more precisely to investigate, in a cross-paradigm perspective, the influence of the constituent order on the data-driven parsing of one of such languages (i.e. Italian). It shows therefore new evidence from experiments on Italian, a language characterized by a rich verbal inflection, which leads to a widespread diffusion of the pro–drop phenomenon and to a relatively free word order. The experiments are performed by using state-of-the-art data-driven parsers (i.e. MaltParser and Berkeley parser) and are based on an Italian treebank available in formats that vary according to two dimensions, i.e. the paradigm of representation (dependency vs. constituency) and the level of detail of linguistic information.

Keywords: Word Order, Morphologically Rich Languages, Parsing

1. Introduction

In Morphologically Rich Languages (MRLs) morphological differences of word forms express information concerning the arrangement of words into syntactic units or cues to syntactic relations (Tsarfaty et al., 2010). This leads to a very large number of possible word forms, but also to free constituent order, discontinuity and pro–drop. On the one hand, where words are featured by a larger variety of inflected forms, they can more often freely change their position with respect to languages which rely on rigid phrase structure, like English and Chinese (Levy and Manning, 2003). On the other hand, rich morphological information in the Verbal head of clauses can predispose to omission of overt subjects, i.e. pro–drop.

A wide literature shows that most of the MRLs share scores of standard metrics for data-driven parsing significantly lower than English. In fact, the best reported F-score for English constituency parsing has surpassed 90% (McClosky et al., 2006)[1], while for several MRLs it remains around at least ten points lower. It should be moreover observed that the dependency paradigm has been demonstrated as more suitable for such kind of languages with respect to the constituency one.

When looking for the reasons for such lower performance, attention should be devoted to the fact that data-driven approaches focus on the number of occurrences of each pattern in the training set. Therefore, if a large number of different patterns is observed, each of them will occur a small number of times, so that the parameter estimation is less reliable and the approach is less effective in discriminating among the different solutions. This effect, known as *data sparseness*, is crucial for parsing performance.

The variation in word order can motivate at least in part both this degradation of results and the different performance in constituency versus dependency parsing since these paradigms deal with word order in different ways. Constituent approaches characterize syntactic structure in predominantly static terms paying minimal attention to word order variation (Weber and Müller, 2004); therefore they are in principle less adequate for representing orders like Verb-Subject-Object (VSO) or Object-Verb-Subject (OVS), where the Subject occurs after Verb, or where any number of adjuncts can be positioned between complements and Verb.

The consequences of such differences in the frameworks are even more evident in data-driven approaches, especially with sparse data. In fact, constituency approaches impose more constraints on word order and therefore consider a larger number of different patterns when a lot of variations are possible. This results in a lower number of occurrences for each pattern, and therefore in a larger impact of data sparseness. All in all, the constituency framework will be more seriously hampered by the changes in word order with respect to the dependency one.

In effect, initially methods for constituency parsing were mainly developed through experimentation on English data and especially the Penn Treebank (Green and Manning, 2010). On the contrary, dependency approaches do not explicitly constrain the word order, at least until structures are continuous and projective, giving a less specified representation of word order.

This paper aims at contributing in a cross-paradigm perspective to the investigation on the influence of the constituent order on the statistical parsing of MRLs. It focuses especially on Italian, a MRL characterized by a rich verbal inflection, which leads to a widespread diffusion of the pro–drop phenomenon and to a relatively free word order. The experiments presented in the paper consist in the application of two state-of-the-art data-driven parsers (i.e. MaltParser and Berkeley parser) to an Italian treebank,

[1] F-score 92.1%

i.e. the Turin University Treebank (TUT). This resource is available in formats that vary according to two dimensions, i.e. the paradigm of representation (dependency vs. constituency) and the level of detail of linguistic information, thus giving an adequate testbed for our investigation.

The paper is organized in the following way. Section 2. summarizes the main recent experiences in Italian parsing both for dependency and constituency. Then Section 3. is devoted to the description of the data used in our experiments and in particular to the description of the annotation formats of TUT. In Section 4., we present the experiments and a discussion of the results. Finally, we draw some conclusion and plans for future work.

2. Related work

In the last years, results for Italian parsing have been reported for both dependency and constituency paradigms mainly in the context of evaluation campaigns.

As far as dependency parsing is concerned, Italian was one of the languages on which parsers were tested during the multilingual track of the CoNLL Shared Task in 2007 (Nivre et al., 2007). The data set was taken from the Italian Syntactic-Semantic Treebank (ISST) (Montemagni et al., 2003), which has been semi-automatically converted to the CoNLL format using information from its two annotated levels, i.e. the constituency and the functional structure. Notwithstanding the relatively small size of the data set (71K tokens), the accuracy for Italian was among the highest (Labeled Accuracy Score 84.40) together with those for Catalan, Chinese and English (Labeled Accuracy Scores between 84.40 and 89.61). More recently, dependency parsers have been tested within the Evalita evaluation campaigns for Italian NLP tools, in 2007, 2009 and 2011 (Bosco et al., 2007; Bosco et al., 2009; Bosco and Mazzei, 2012b). The data sets were taken from the available releases of TUT, whose size progressively increased from 2,400 to more than 3,450 sentences (102,150 tokens in the current version). A version of this treebank in CoNLL format was created for the Evalita evaluation campaigns. The results reported in the Evalita contests improved constantly over the years. In 2007, the best reported Labeled Accuracy Score (LAS) was 86.94 by TULE (Lesmo, 2007), a rule-based system developed in parallel with TUT. In 2009, the best LAS was around 88.70 achieved by both TULE (Lesmo, 2009) and DeSR (Attardi et al., 2009), a statistical parser[2]. Finally in 2011, the best performance has been scored with LAS 91.23 and was achieved by the system described in Grella et al. (2012).

As far as constituency parsing is concerned, the only recently published results are those reported with reference to the Evalita campaigns, which have been constantly improved but still far from those for English. In fact, the best performance attested at the 2011 edition of Evalita was F_1 82.96, Bracketing Recall 82.97 and Bracketing Precision 82.94 (Bosco and Mazzei, 2012a).

The relevance of the comparison between different frameworks is also shown by a recent work (Tsarfaty et al., 2012) on the problem of a fair comparison of performance in different frameworks.

As far as word order (and, more specifically, constituent order) is concerned, it is usually included among the features that can motivate the degradation of results in parsing MRLs, and it has been mainly investigated in the perspective of tasks such as Machine Translation and Language Generation (Baldwin and Tanaka, 2000). It is acknowledged that a combination of several factors determines the order of words, e.g. semantic roles, topic, focus, theme/rheme and communicative events. In free word order languages, the order is used to structure the information being conveyed to the hearer, while in fixed word order languages the same role is played by intonation and stress (Hoffman, 1995). Nevertheless, a difficulty is related to the formal description and processing of free word order languages: instead of a complete lack of ordering rules, many subtle language specific restrictions apply to the order variation (Green and Manning, 2010). Therefore, a large amount of variations can be considered as grammatical in isolated sentences, but, depending on the context, different word orders are either required or more natural than others (Steinberger, 1994).

On the one hand, approaches to language description based on constituency characterize syntactic structure in predominantly static terms paying minimal attention to the communicative function that mainly motivates the word order variation (Weber and Müller, 2004). They are usually considered not adequate for representing orders like VSO (Verb-Subject-Object) or OVS (Object-Verb-Subject), where the Subject is after the Verb, or where any number of adjuncts can be positioned between complements. For instance, in order to represent such a kind of structures, the Penn format should be increased by new representational tools, like in TUT–Penn (see Section 3.2.). On the other hand, dependency approaches do not explicitly constrain the word order, at least until structures are continuous and projective, giving a less specified representation of word order. At least in part, this can explain the different performance in constituency versus dependency parsing for MRLs.

3. Data Sets: TUT

The experiments presented in this paper are based on TUT, i.e. the freely available Italian resource developed by the Natural Language Processing group of the University of Turin (Bosco et al., 2000)[3]. The data currently consist in 102,150 annotated tokens (among which 84,666 words, 10,056 punctuation marks and 7,428 null elements) in TUT native format, which correspond to around 3,500 sentences[4] extracted from texts varying from newspapers, to legal, to Wikipedia. In the rest of this section, we will describe in detail the formats available for TUT focusing in particular on the distinctive dimensions of variation which characterize the annotation of this treebank, namely the paradigm of representation (dependency vs. constituency) and the level of specification of linguistic information.

[2]This contest included also a pilot task with training and testing data from ISST (best LAS 83.38 by Attardi et al. (2009)).

[3]http://www.di.unito.it/~tutreeb

[4]Average sentence length 23.90 words per sentence.

3.1. The dependency formats

The core of the TUT project is a treebank in an original dependency format, henceforth indicated as *native TUT*, which has been afterwards enriched by the converted versions in constituency (see 3.2.). Native TUT includes a specific format for representing Italian morphology and syntax.

For what concerns morphology, the tag set of the native TUT richly describes the features of a MRL like Italian, and includes 16 grammatical categories further specialized by 43 types which are associated with a large variety of features[5]. For the amalgamated words, which are frequently used in Italian as in other MRLs, it is assumed an explicit representation of each of their parts as separated morpho-syntactic items. For instance, Figure 1 shows instances for the Articled Prepositions "sulla" (on the[fem sing]) and "della" (of the[fem sing]) respectively composed by one item for the Preposition and one for the Article. In TUT the 7.25% of words are part of some amalgam, which are in the most of cases articled Prepositions, and in the rest of cases clitics.

Instead, for what concerns syntax, native TUT is a pure dependency representation centered upon the notion of argument structure and applying the major principles of the *Word Grammar* theoretical framework (Hudson, 1984). This is mirrored, for instance, in the annotation of Determiners and Prepositions which are represented in TUT trees as complementizers of Nouns or Verbs. See, for instance, in Figure 1 the Determiner "il" (the) which is the head for the Noun "vento" (wind). Contrary to most of dependency-based annotations, for pro–drop, flexible word order and discontinuity, the annotation strategy adopted in TUT consists in using null elements in order to avoid crossing branches in dependency trees and to allow an explicit representation of the argument structure of each Verb, also for omitted complements. Around the 7.25% of the annotated tokens of the treebank are null elements (around 7,500), and they are in the most of cases (61.07%) co–indexed with some other word of the sentence, to represent e.g. occurrences of gapping or equi phenomenon. Non co–indexed null elements are instead used e.g. for the representation of elliptical constructions, pro–drop subjects or other complements playing some role in argument structure of Verbs.

Moreover, the treebank exploits a rich set of grammatical relations designed to represent linguistic information according to three different perspectives, namely morpho–syntax, functional syntax and semantics. Since the information related to each perspective is annotated in specially designed part of the relation label of TUT, called *component* (i.e. *morpho–syntactic*, *functional–syntactic* or *syntactic–semantic component*), the amount of linguistic knowledge annotated in the treebank can be easily varied by assuming more or less detailed relations, i.e. including from one to three of the above mentioned perspectives (below referred as *1–Comp*, *2–Comp* and *3–Comp*). This means

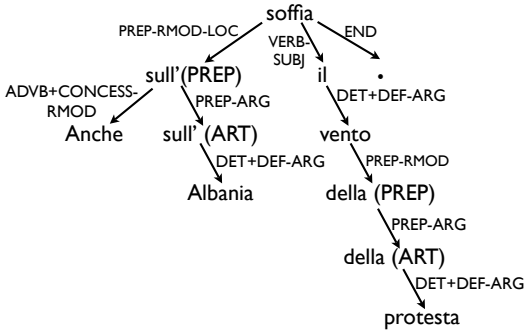

Figure 1: Sentence NEWS–549 in 3–Comp setting: "Anche sull'Albania soffia il vento della protesta." (Also on the Albania blows the wind of the revolt.).

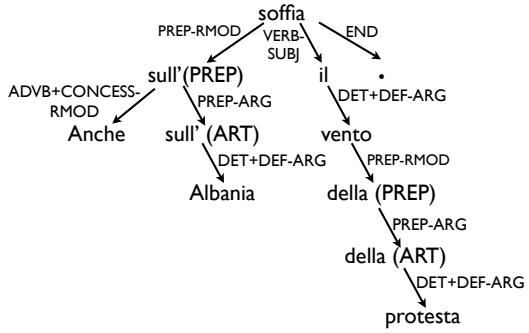

Figure 2: Sentence NEWS–549 in 2–Comp setting.

that each relation label can in principle include all the three components, but can be made more or less specialized, including information from only one (i.e. the functional–syntactic) or two of them. For instance, the relation used for the annotation of locative prepositional modifiers, i.e. PREP–RMOD–LOC (which includes all the three components, in Figure 1), can be reduced to PREP–RMOD (which includes only the morpho–syntactic and the functional–syntactic component, in Figure 2) or to RMOD (which includes only the functional-syntactic component, in Figure 3). This works as a means for the annotators to repre-

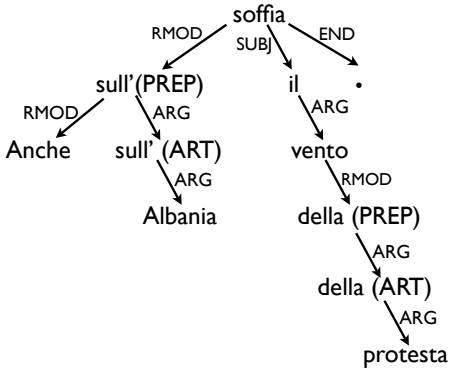

Figure 3: Sentence NEWS–549 in 1–Comp setting.

sent different layers of confidence in the annotation, but can also be applied to increase the comparability of TUT with

[5]The original PoS tag set of TUT is available at `http://www.di.unito.it/~tutreeb/syntcat-22-7-02.doc`.

other existing resources, by exploiting the amount of linguistic information more adequate for the comparison, e.g. in terms of number of relations. For instance, in the Evalita campaigns the 1–Comp setting of the treebank has been exploited. Since in more coarse-grained settings several relations can be merged into a single one (e.g. PREP–RMOD–TIME, used for temporal modifier, and PREP–RMOD–LOC are merged in RMOD), each setting includes a different number of relations: the setting based on the single functional–syntactic component (1–Comp) includes 72 relations, the one based on morpho–syntactic and functional–syntactic components (2–Comp) 140, and the one based on all the three components (3–Comp) 323.

3.2. Constituency formats

By applying conversion scripts to the treebank in native TUT format, the constituency version of TUT has been generated, which includes in particular the TUT–Penn and the Augmented–TUT–Penn (henceforth APE) formats.

TUT–Penn is an application of the English Penn Treebank (PTB) format to Italian, as happened for other languages, like Chinese[6] or Arabic[7], addressing the phenomena typical of these languages by new specific representational means. For what concerns morphology, the size of the PoS tag set of the TUT–Penn, if compared with that exploited in English PTB, clearly reflects the differences between MRLs and morphologically poorer languages. Nevertheless, even if the representation of morphology is more fine-grained with respect to the one adopted for English in PTB, it is reduced with respect to the PoS tag set used in native TUT in order to avoid serious sparse data problems (Collins et al., 1999). As said above, native TUT exploits a tag set including 16 grammatical categories, specialized by 43 types and a large variety of features. By contrast, TUT–Penn adopts a tag set of 68 tags only (versus 36 in the PTB). Beyond the information that the PTB tag set makes explicit[8], TUT–Penn takes into account a richer variety of features for Verbs, Adjectives and Pronouns. For the amalgamated words, as in native TUT, it is assumed an explicit representation of each of their parts as separated morpho-syntactic items, see e.g. the Articled Prepositions "sulla" (on the[fem sing]) and "della" (of the[fem sing]) in figure 4.

As far as syntax is concerned, the annotation in TUT–Penn is structurally the same as in PTB, but some difference can be observed with respect to the inventory of functional relations and the use of null elements. In fact, the (very limited set of) functional tags assumed in PTB is used also in TUT–Penn, but it is increased by some relations used for representing phenomena related to the flexible Italian word order. For instance, the label EXTPSBJ is used for the annotation of subjects in post-verbal position. Also the standard PTB inventory of null elements is adopted in TUT–Penn, but while for English null elements are mainly traces denoting constituent movements, in TUT–Penn they can play different roles: zero Pronouns, reduction of relative clauses, elliptical Verbs and also the duplication of Subjects which are positioned after Verbs (which occurs around 900 times in the corpus).

```
( (S
    (PP-LOC (ADVB Anche)
            (PREP sull')
            (NP (ART~DE sull') (NOU~PR Albania)))
    (NP-SBJ (-NONE- *-1))
    (VP (VMA~RE soffia)
        (NP-EXTPSBJ-1
            (NP (ART~DE il) (NOU~CS vento))
            (PP (PREP della)
                (NP (ART~DE della) (NOU~CS protesta))))))
    (..)) )
```

Figure 4: Sentence NEWS–549 in TUT–Penn.

To expand the possibility of cross-framework and cross-paradigm comparison and assuming the importance of the representation of the predicate argument structure in constituency-based representations too, we developed also the APE, a format which extends TUT–Penn by inheriting, when possible, the functional-syntactic knowledge encoded in the native dependency TUT. Figure 5 shows an example where the tags of this format allow to draw distinctions among modifier and argument functions (e.g. PP–RMOD–LOC instead of PP–LOC in TUT–Penn, NP–ARG instead of NP to represent the arguments of Prepositions), and to annotate the function of the final punctuation mark (i.e. END). As in the native TUT, it is therefore possible to graduate the amount of linguistic knowledge annotated also in the constituency formats of TUT.

```
( (S
    (PP-RMOD-LOC (ADVB Anche)
                 (PREP sull')
                 (NP-ARG (ART~DE sull')
                         (NOU~PR Albania)))
    (NP-SBJ (-NONE- *-1))
    (VP (VMA~RE soffia)
        (NP-EXTPSBJ-1
            (NP (ART~DE il)
                (NOU~CS vento))
            (PP-RMOD (PREP della)
                     (NP-ARG (ART~DE della)
                             (NOU~CS protesta))))))
    (PUNCT-END .)) )
```

Figure 5: Sentence NEWS–549 in Augmented–TUT–Penn.

We conclude this section with some observation on the description of Italian that can be extracted from an analysis of TUT. This description mainly confirms that Italian has to be considered among MRLs (see e.g. (Tsarfaty et al., 2010)) since it shows quantitatively the features known in literature for this kind of languages: rich inflection, amalgamated words, pro–drop and a relatively free order of words. In particular, for word order, the analysis in Table 1(a) according to the Greenberg's six-ways typology (Greenberg, 1963), shows that all the six possible permutations of the

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

[7]See http://www.ircs.upenn.edu/arabic/.

[8]Apart from a few cases of English morphological features which do not exist (e.g. possessive ending) or do not correspond with Italian forms (e.g. comparative Adjective and Adverb).

order	**frequency**
SVO	79.09
SOV	7.20
OSV	6.61
OVS	5.13
VSO	1.08
VOS	0.89

(a)

order	**frequency**
SV	79.10
VS	20.90
OV	18.93
VO	81.07

(b)

Table 1: The frequency of permutations of (a) Subject, Verb and Object in Italian declarative clauses and of (b) the Subject and Object preceding and following the Verb in Italian declarative clauses.

main constituents can be found in declarative clauses[9] in TUT corpus, with the order SVO strongly prevailing on the others. But, since this analysis takes into account only transitive Verbs, with realized Object, and can be influenced by pro–drop, our observation has to be widened to the cases where the Verb precedes or follows Subject and Object (Dryer, 2007) (see Table 1(b)).

4. Experimental Assessment

The aim of the experimental assessment is to compare the robustness of dependency and constituency models with respect to a free constituent order language as Italian and with respect to the amount of annotated linguistic information. The experiments are performed on the different TUT formats (i.e. 1|2|3–Comp for dependency and APE and Penn for constituency) discussed in Section 3. by using two parsers, namely the Berkeley parser (Petrov and Klein, 2007) for the constituency model, and MaltParser (Nivre, 2003; Nivre et al., 2006) for the dependency one. Indeed, these two parsers have shown state-of-the-art performance during EVALITA 2009 and 2011.

The Berkeley parser is a constituency parser based on a hierarchical coarse-to-fine parsing, where a sequence of grammars is considered, each being the refinement, namely a partial splitting, of the preceding one. Its performance is at the state of the art for English and for other languages. An interesting characteristic is that porting the Berkeley parser to a new language requires no additional effort apart from the availability of a treebank. Constituency parser performance is evaluated as usual by labeled precision (LP) and recall (LR) and F_1.

MaltParser is a data-driven dependency parser showing topmost performance in the multilingual track of the CoNLL shared tasks on dependency parsing in 2006 and 2007 and in the EVALITA 2009 dependency parsing task for Italian. Dependency parser performance is evaluated in terms of Labeled Attachment Score (LAS).

Statistical significance has been evaluated by using Dan Bikel's Randomized Parsing Evaluation Comparator[10]. This test checks whether the following null hypothesis can be rejected:

H_0: the difference in performance between the two experiments is not statistically significant.

To do so, the performance scores for the single sentences are shuffled between the two models, and then precision and recall are recomputed. The shuffling is repeated a large number n_t of times (up to $10,000$), and the number n_c of times where shuffling induced a variation in performance larger than the difference between the two models is counted. Eventually the probability p that the null hypothesis is incorrectly rejected is estimated by $p = \frac{n_c+1}{n_t+1}$. In other words, the difference in performance between the two models gets more statistically significant as long as the value of p gets smaller.

In order to study performance variations on sentences with different constituent order, the data set has been split in two parts: the former (SVO) includes all the sentences where the SVO constituent order is represented at least once; the latter (noSVO) includes all the other sentences, where all

Data set	pattern	size
training set	SVO	646
	noSVO	2,379
	all	3,025
test set	SVO	110
	noSVO	390
	all	500

Table 2: Data set dimensions

the other constituent orders are represented, but not the SVO. As shown in Table 2, the split of data between the two patterns is strongly unbalanced in favor of the noSVO, corresponding to nearly four times the number of sentences of the other pattern, both in the training and in the test sets. To overcome the difficulty of such unbalance, we therefore decided to randomly subsample the noSVO data sets to obtain a training and a test set with exactly the same dimensions of the SVO case. To avoid the risk of biased results, we repeated each experiment 20 times and averaged the corresponding outputs, obtaining the performance reported in Tables 3 and 4. A statistical significance test is applied to each iteration. The TUT data set is not divided in training and test set. Therefore assessment is performed by following the 10-fold cross validation protocol.

4.1. Constituency Parser

For constituency, parsing performance for Penn and APE formats is depicted in Table 3. The five macro columns correspond to the five different models, obtained by training the parser on: (i) all the training set (All); (ii) only the SVO and (iii) the noSVO parts of the training data (SVO and noSVO respectively); (iv) by averaging performance on 20 runs made by subsampling the noSVO training set (sub-noSVO); and (v) by considering for training the union of the SVO training set and each of the sets in sub noSVO, and again averaging performance (balanced). For all the models, performance in terms of LP, LR and F_1 is reported. In all the cases, the null hypothesis can be rejected with values of p lower than 0.05 and then the comparisons between performance of all pairs of models result to be statistically

[9] The term declarative clause refers to clauses where Verb is in tensed form and not playing the role of relative.

[10] The tool is freely available from `http://www.cis.upenn.edu/~dbikel/software.html#comparator`

	Training Set														
	All			SVO			noSVO			sub-noSVO			balanced		
Test Set															
Penn	LR	LP	F_1	LR	LP	F_1	LR	LP	F_1	LR	LP	F_1	LR	LP	F_1
SVO	81.75	81.37	81.56	72.34	71.49	71.91	79.39	78.10	78.74	69.73	67.95	68.83	76.87	76.41	76.64
noSVO	80.03	80.19	80.11	71.04	70.09	70.56	77.90	77.37	77.64	70.46	69.56	70.01	76.50	76.46	76.48
all	80.51	80.53	80.52	71.42	70.50	70.95	78.32	77.58	77.95	70.26	69.10	69.68	76.60	76.45	76.52
APE															
SVO	77.11	76.96	77.04	69.56	70.21	69.88	78.50	78.90	78.70	67.03	65.36	66.18	74.90	74.03	74.46
noSVO	79.26	79.47	79.36	72.02	72.12	72.07	79.18	78.92	79.05	70.12	69.06	69.59	75.71	75.42	75.57
all	78.69	78.88	78.78	71.57	71.47	71.52	79.02	78.92	78.97	69.34	68.12	68.73	75.51	75.08	75.29

Table 3: Constituency parser performance: comparisons between all pairs of models are statistically significant as $p \leq 0.05$.

significant. Also the standard deviation has been computed for all averaged cases (sub-noSVO and balanced) and its values are always lower than 3. The values have not been reported for providing more compact and readable tables.

First of all, note that the first column (All) represents a sort of baseline, where all available data are exploited. We can see how the addition of more detailed information, in APE with respect to Penn format, does not help parsing (except when the training set is composed by noSVO parse trees and in the two test sets there are noSVO and All examples), probably because of the increased data sparsity. In fact, we would need a bigger treebank to accurately train the more precise APE labels. Furthermore, when comparing parsing performance on the SVO and noSVO data sets, we note that the Penn format favors the SVO pattern, while the APE favors the noSVO. This property is maintained also when training is performed either on SVO or on noSVO data alone, and this is quite surprising, but it probably still depends on the influence of the annotation and on the inclusion in the SVO data set of some noSVO pattern (SVO data set contains all and only the sentences containing *at least* one SVO pattern). On the other hand, when we consider the two models obtained by subsampling, namely sub-noSVO and balanced-train, the former always performs better on the corresponding noSVO test set. We can therefore conclude that the better performance of the noSVO model is also related to the fact that the training set is much larger than in the SVO case.

In general, we can conclude that the best choice is to include all the data available in the training set: indeed, this is the case with the best performance on both SVO and noSVO test sets. As a second choice, when the training sets are balanced, the best performance is obtained, as could have been expected, by training the parser on sentences as similar as possible to the ones composing the test set.

4.2. Dependency Parser

Also for the dependency paradigm, performance deteriorates when the information in the annotation augments, particularly for 3-Comp. Moreover, 3-Comp performance is much less stable than the other two cases, suggesting that we are in a data sparsity condition. We therefore decided to focus our analysis on 1-Comp and 2-Comp.

In general, in the dependency case performance remains more or less the same even when training is performed on sentences with a different constituent order with respect to the test set. In fact, in no comparison the value of p is small enough to guarantee the statistical significance of the differences, with the only exception of the difference between the models trained on noSVO and on SVO and tested on the SVO test set. In this case there is no statistical significance both with and without subsampling for the noSVO training set.

The fact that performance is only slightly sensitive to the different patterns suggests that the dependency paradigm is more robust than the constituency one with respect to variability in the constituent order and therefore more suitable to MRLs with such feature.

5. Conclusion and Future Work

The comparison between the preliminary results obtained with the constituency and with the dependency approaches suggests that the latter is more effective with respect to the free order of constituents than the former. The results should be considered as preliminary because of the limited size of the data set. Indeed, data sparseness hampers the reliability of results, especially for the most detailed annotation formats. As soon as more annotated data are available, we will be able to carry on new experiments that exploit more accurate annotation schemata, such as APE for constituency and 3-Comp for the dependency paradigm.

While we can expect that more annotated data will result in more reliable performance estimation, we do not think that the difference between constituency and dependency will substantially reduce. In fact, with more data, the number of different patterns is likely to grow more rapidly for the constituency paradigm, where different patterns are produced by different word orders, than for the dependency approach.

Another aspect that we plan to investigate is related to null elements. Usually they are removed before parsing, both for constituency and dependency[11]. Given that in Italian null elements occur quite frequently, it would be interesting to apply to Italian what was done for Korean in Chung et al. (2010), for investigating the effects of taking into account null elements in parsing.

Eventually, when enough data will be available, we could also consider the effect of variations between the different textual genres. Indeed, the TUT data set even now contains legal texts which are substantially different from, for examples, the kind of texts extracted from Wikipedia.

[11] See e.g. the standard CoNLL format, where null elements are not allowed.

	Training Set				
	All	**SVO**	**noSVO**	**sub-noSVO**	**balanced**
Test Set					
1-Comp					
SVO	88.44	86.13	83.63	83.49	87.34
noSVO	87.62	86.87	82.43	83.95	85.57
all	87.86	86.65	83.63	83.81	86.08
2-Comp					
SVO	88.84	86.33	86.25	82.62	86.84
noSVO	86.72	86.45	81.11	82.91	84.77
all	87.34	86.41	82.62	82.82	85.38
3-Comp					
SVO	84.60	81.92	86.53	78.09	82.71
noSVO	83.10	82.55	76.87	78.72	80.83
all	83.54	82.86	81.98	78.53	81.38

Table 4: Dependency parser performance: Labeled Accuracy Score.

Acknowledgements

We would like to thank Enrico Principe for his support in the experiments. Moreover, we would like to thank Joakim Nivre for making available MaltParser and Slav Petrov for making available the Berkeley parser. The work reported in this paper was partly carried out in the framework of the project PARLI ("Portal for the Access to the Linguistic Resources for Italian") benefiting from a research funding PRIN-2008 from the Italian Ministry of Public Instruction and University.

6. References

G. Attardi, F. Dell'Orletta, M. Simi, and J. Turian. 2009. Accurate dependency parsing with a stacked multilayer perceptron. In *Proceedings of Evalita'09*, Reggio Emilia.

T. Baldwin and H. Tanaka. 2000. The effects of word order and segmentation on translation retrieval performance. In *COLING*, pages 35–41.

C. Bosco and A. Mazzei. 2012a. The Evalita 2011 parsing task: the constituency track. In *Evalita 2011 Working Notes*.

C. Bosco and A. Mazzei. 2012b. The Evalita 2011 parsing task: the dependency track. In *Evalita 2011 Working Notes*.

C. Bosco, V. Lombardo, L. Lesmo, and D. Vassallo. 2000. Building a treebank for Italian: a data-driven annotation schema. In *Proceedings of LREC'00*, Athens, Greece.

C. Bosco, A. Mazzei, and V. Lombardo. 2007. Evalita parsing task: an analysis of the first parsing system contest for Italian. *Intelligenza artificiale*, 2(IV).

C. Bosco, S. Montemagni, A. Mazzei, V. Lombardo, F. Dell'Orletta, and A. Lenci. 2009. Evalita'09 parsing task: comparing dependency parsers and treebanks. In *Proceedings of Evalita'09*, Reggio Emilia.

T. Chung, M. Post, and D. Gildea. 2010. Factors affecting the accuracy of Korean parsing. In *Proceedings of SPMRL 2010*.

M. Collins, J. Hajic, L. Ramshaw, and C. Tillmann. 1999. A statistical parser of Czech. In *Proceedings of the ACL'99*.

M. S. Dryer. 2007. Word order. In Timothy Shopen, editor, *Clause Structure, Language Typology and Syntactic Description*, volume vol. 1. Cambridge University Press.

S. Green and C. D. Manning. 2010. Better Arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of COLING 2010*.

J. H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Language*, pages 73–113. MIT Press, London.

M. Grella, M. Nicola, and D. Christen. 2012. Experiments with a constraint-based dependency parser. In *Evalita 2011 Working Notes*.

B. Hoffman. 1995. Integrating "free" word order syntax and information structure. In *Proceedings of EACL'95*.

R. Hudson. 1984. *Word grammar*. Basil Blackwell, Oxford and New York.

L. Lesmo. 2007. The rule-based parser of the NLP group of the University of Torino. *Intelligenza artificiale*, 2(IV).

L. Lesmo. 2009. The Turin University Parser at Evalita 2009. In *Proceedings of Evalita'09*, Reggio Emilia.

R. Levy and C. D. Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? In *Proceedings of ACL'03*, pages 439–466.

D. McClosky, E. Charniak, and M. Johnson. 2006. Effective selt-training for parsing. In *Proceedings of NAACL'06*.

S. Montemagni, F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Lenci, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M. T. Pazienza, D. Saracino, F. Zanzotto, N. Nana, F. Pianesi, and R. Delmonte. 2003. Building the Italian syntactic- semantic treebank. In A. Abeillè, editor, *Treebanks: Building and Using Parsed Corpora*, volume Treebanks: Building and Using Parsed Corpora, pages 189–210. Kluwer.

J. Nivre, J. Hall, and J. Nilsson. 2006. MaltParser: a data-driven parser-generator for dependency parsing. In *Proceedings of LREC-2006*.

J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson,

S. Riedel, and D. Yuret. 2007. The CoNLL 2007 Shared Task on dependency parsing. In *Proceedings of the CoNLL 2007 Shared Task, EMNLPCoNLL*, pages 915–932, Prague.

J. Nivre. 2003. An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)*, pages 149–160.

S. Petrov and D. Klein. 2007. Learning and inference for hierarchically split PCFGs. In *Proceedings of AAAI (Nectar Track)*.

R. Steinberger. 1994. Treating 'free word order' in Machine Translation. In *Proceedings of COLING '94*.

R. Tsarfaty, D. Seddah, Y. Goldberg, S. Kübler, Y. Versley, M. Candito, J. Foster, I. Rehbein, and L. Tounsi. 2010. Statistical parsing of morphologically rich languages (SPMRL) what, how and whither. In *Proceedings of SPMRL 2010*.

R. Tsarfaty, J. Nivre, and E. Andersson. 2012. Cross-framework evaluation for statistical parsing. In *Proceedings of the 13th Conference of the European Chapter of the ACL (EACL 2012)*, Avignon, France, April.

A. Weber and K. Müller. 2004. Word order variation in German main clauses: a corpus analysis. In *Proceedings of the 20th International Conference on Computational Linguistics*.