EVALITA 2009
The Parsing Task

C. Bosco*, S. Montemagni#, A. Mazzei*,
V. Lombardo*, F. Dell’Orletta#, A. Lenci#

* Interaction Models Group
  Dipartimento di Informatica
  Università di Torino

# DyLan Laboratory
  ILC – CNR
  Dipartimento di Linguistica – Univ. di Pisa
• Introduction

• Structure of the Task:
  – Constituency Track
  – Dependency Track (Main and Pilot Subtask)

• Evaluation and results in:
  – Constituency Track
  – Dependency Main Subtask
  – Dependency Pilot Subtask
The Parsing Task

aims to define and extend the current state of the art in parsing Italian

by encouraging the application of different models and approaches

and accounting for different annotation paradigms
The Parsing Task consists in the automatic annotation of the syntactic structure of PoS tagged sentences.

The Parsing Task is treebank-based: a syntactically annotated corpus of Italian sentences is given for the development and training of parsing systems.
Structure of the task

Two treebanks developed for Italian
- Turin University Treebank (TUT) in both dependency and constituency format
- ISST–TANL, a revision of the ISST–CoNLL corpus, in its turn derived from the Italian Syntactic–Semantic Treebank (ISST, SITAL project)

TUT and ISST–TANL differ at the level of
- corpus composition
- dependency annotation schemes

They are a testbed to start evaluating the influence of treebanks on the performance of parsers
The evaluation is based on standard metrics

- For constituency parsing
  - Precision: percentage of found brackets which are correct
  - Recall: percentage of brackets correct which are found
  - F measure: \(2 \times (P \times R) / (P + R)\)

- For dependency parsing
  - Labeled Attachment Score: percentage of tokens with correct head and relation label
The Task is articulated in tracks and subtasks according to different annotation paradigms and formats.

- Constituency Track
- Dependency Track:
  - Main Task
  - Pilot Task
( (S (NP-SBJ (ART~DE La) (NOU~CS coppia))
    (PRN (, ,))
    (VP (VMA~PA residente))
    (NP (-NONE- *))
    (PP (PREP a))
    (NP (NOU~PR Milano)))
   (ADVB anche) (CONJ se)
    (VP (-NONE- *))
    (PP-PRD (PREP di))
    (NP (NOU~CP origini) (ADJ~QU siciliane))))
  (, ,))))
  (VP (VAU~IM stava))
    (VP (VMA~GE trascorrendo))
    (NP
      (NP (ART~IN un) (NOU~CS periodo))
      (PP (PREP di))
      (NP (NOU~CS vacanza)))))))
  (, .)))
Data:

- Annotation: TUT-Penn format
- Training set: 2,200 sentences (average length 24.98 words)
- Test set: 200 sentences (shorter than 40 words) shared with Dependency Main subtask
- Balance: 1,100 sentences from newspapers and 1,100 from Civil Law code

### Evaluation and results

#### Constituency

<table>
<thead>
<tr>
<th>Participant</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
</tr>
<tr>
<td>Lavelli_FBKTrento</td>
<td>78.73</td>
</tr>
<tr>
<td>Sangati_uniAmsterdam</td>
<td>75.79</td>
</tr>
</tbody>
</table>

- Balance: 1,100 sentences from newspapers and 1,100 from Civil Law code
The results positively compare with the previous Evalita edition with the following improvement:

- 10.76% for F-measure
- 12.12% for Precision
- 9.21% for Recall
1 La IL ART ART DEF—F—SING
2 coppia COPPIA NOUN NOUN COMMON—F—SING
3 , PUNCT PUNCT -
4 residente RISIEDERE VERB VERB MAIN—PARTICIPLE—PAST—
INTRANS—SING—ALLVAL
5 a A PREP PREP MONO
6 Milano MILANO NOUN NOUN PROPER—F—SING—CITY
7 anche ANCHE ADV ADV CONCESS
8 se SE CONJ CONJ SUBORD—COND
9 di DI PREP PREP MONO
10 origini ORIGINE NOUN NOUN COMMON—F—PL
11 siciliane SICILIANO ADJ ADJ QUALIF—F—PL
12 , PUNCT PUNCT -
13 stava STARE VERB VERB AUX—IND—IMPERF—
INTRANS—3—SING
14 trascorrendo TRASCORRERE VERB VERB MAIN—GERUND—PRES—
TRANS—SING
15 un UN ART ART INDEF—M—SING
16 periodo PERIODO NOUN NOUN COMMON—M—SING
17 di DI PREP PREP MONO
18 vacanza VACANZA NOUN NOUN COMMON—F—SING
19 , PUNCT PUNCT -
14 END - -
| 1 | La | lo | R | RD | num=s—gen=f | 2 | det |
| 2 | coppia | coppia | S | S | num=s—gen=f | 13 | subj |
| 3 | , | , | F | FF | - | 4 | punc |
| 4 | residente | residente | A | A | num=s—gen=n | 2 | mod |
| 5 | a | a | E | E | - | 4 | comp_loc |
| 6 | Milano | milano | S | SP | - | 5 | prep |
| 7 | anche_se | anche_se | C | CS | - | 4 | con |
| 8 | di | di | E | E | - | 4 | conj |
| 9 | origini | origine | S | S | num=p—gen=f | 8 | prep |
| 10 | siciliane | siciliano | A | A | num=p—gen=f | 9 | mod |
| 11 | , | , | F | FF | - | 4 | punc |
| 12 | stava | stare | V | VA | num=s—per=3—mod=i—ten=i | 13 | modal |
| 13 | trascorrendo | trascorrere | V | V | mod=g | 0 | ROOT |
| 14 | un | un | R | RI | num=s—gen=m | 15 | det |
| 15 | periodo | periodo | S | S | num=s—gen=m | 13 | obj |
| 16 | di | di | E | E | - | 15 | comp |
| 17 | vacanza | vacanza | S | S | num=s—gen=f | 16 | prep |
| 18 | . | . | F | FS | - | 13 | punc |
TUT vs ISST–TANL format

Granularity and inventory of dependency types

Head selection, e.g. determiner is head of noun in TUT and dependent in ISST–TANL

Projectivity
TUT reduces the non-projective to projective structures, while ISST–TANL allows for non-projective representations

Annotation of various structures and phenomena, e.g. coordination, punctuation, root constraint (single in TUT but not in ISST–TANL)
Evaluation and results

Data:

- Annotation: TUT schema in CoNLL format

<table>
<thead>
<tr>
<th>Participant</th>
<th>All testset</th>
<th>Newspaper SharedTS</th>
<th>Civil law</th>
<th>Passage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesmo_uniTorino</td>
<td>88.73</td>
<td>84.68</td>
<td>92.63</td>
<td>89.63</td>
</tr>
<tr>
<td>Attardi_uniPisa</td>
<td>88.67</td>
<td>82.60</td>
<td>91.54</td>
<td>90.10</td>
</tr>
<tr>
<td>Lavelli_FBKTrento</td>
<td>86.50</td>
<td>79.91</td>
<td>90.23</td>
<td>89.11</td>
</tr>
<tr>
<td>Sangati_uniAmsterdam</td>
<td>84.98</td>
<td>76.66</td>
<td>89.93</td>
<td>87.87</td>
</tr>
<tr>
<td>Soegaard_uniCopenhagen</td>
<td>80.42</td>
<td>72.84</td>
<td>86.04</td>
<td>80.94</td>
</tr>
<tr>
<td>Dini_CELITorino</td>
<td>68.00</td>
<td>63.86</td>
<td>70.74</td>
<td>68.89</td>
</tr>
</tbody>
</table>
Data:

- Training and development set: 2,868 + 241 sentences (66,528 + 4,745 tokens)

- Test set: 260 sentences (5,011 tokens), 100 shared with the dependency main subtask

- Balance: newspapers and periodicals (no internal organisation in subcorpora)

- Annotation: ISST–TANL
Evaluation and results

<table>
<thead>
<tr>
<th>Participant</th>
<th>Test set</th>
<th>SharedTS</th>
<th>Rest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attardi_uniPisa</td>
<td>83.38</td>
<td>84.67</td>
<td>82.70</td>
</tr>
<tr>
<td>Lavelli_FBKTrento</td>
<td>80.54</td>
<td>81.12</td>
<td>80.24</td>
</tr>
<tr>
<td>Soegaard_uniCopenhagen</td>
<td>78.51</td>
<td>78.61</td>
<td>78.45</td>
</tr>
<tr>
<td>Lesmo_uniTorino</td>
<td>73.44</td>
<td>75.12</td>
<td>72.56</td>
</tr>
<tr>
<td>Dini_CELITorino</td>
<td>57.81</td>
<td>60.78</td>
<td>56.27</td>
</tr>
</tbody>
</table>

**SharedTS**: built by enforcing a constraint on the sentence length which could not exceed 40 tokens (average sentence length 17.16)  
**Rest**: average sentence length 20.59
Evaluation and results
MAIN vs PILOT Subtasks

Overall results
- the best results refer to the Main subtask
  - 5.35: difference between the best LAS in Main and the best LAS in Pilot
  - 8.15: difference between the average LAS scores in Main and Pilot

Shared Test Set
- No significant difference between the best LAS scores obtained in the two subtasks (84.68 vs 84.67)
- 0.718: difference between the average LAS scores obtained in Main and Pilot

performance of individual parsing systems across subtasks
- stochastic parsers show higher LAS scores in Pilot
- rule–based parsers achieve best results in the Main
Dependency track
• obtained results are promising and positively compare with other evaluation campaigns

EVALITA–07 (comparison based on the MAIN subtask)
• improvement, not only for the best scores
  – Evalita–07 best LAS: 86.94
  – Evalita–09 best LAS: 88.73
• average LAS passed from 72.48 to 82.88

Multi–lingual track of CoNLL 2007 Shared Task on Dependency Parsing (comparison based on the PILOT subtask)
• results in line with the state–of–the–art dependency parsing
• results for Italian at CoNLL—2007 and EVALITA–09 show the same range of variation
  – CoNLL–2007: from 84.40 to 59.75
  – Evalita–09: from 83.38 to 57.81
By testing the participant systems on two different annotated corpora we confirmed that:

- statistical approaches show more flexibility in adapting themselves to new texts and domains, while the adaptation of rule-based systems need more efforts

- the comparison of the results of the parsing systems on the SharedTestSet across subtasks raises the issue of the role of training corpora in the parsers performance. Differences lie at the level of
  - size and composition (in text genres) of the training corpora
  - adopted annotation schemes
Annotation scheme and parsers performance: MAIN vs PILOT

Main subtask: LAS scores on the Shared test set by dependency type

PILOT Subtask: LAS scores on the Shared Test Set by dependency type
Annotation scheme and parsers performance: PILOT subtask (1)

PILOT Subtask: LAS scores on the Shared Test Set by dependency type

AndersSoegaard_PREC
CELI_PREC
Lavelli_PREC
Lesmo_PREC
ATTARDI_PREC
Annotation scheme and parsers performance: PILOT subtask (2)

PILOT Subtask: LAS scores on the Shared Test Set by underspecified dependency type
Conclusions

- Results achieved in the EVALITA’09 Parsing Task are promising
  - in line with state-of-the-art for dependency parsing
  - significantly improved for constituency parsing

- The task create the prerequisites to contribute to the current debate about the influence of paradigms and models with respect to Italian

- The comparative analysis of TUT and ISST–TANL can be the basis for the design and development of a bigger resource for Italian dependency parsing