Converting Italian Treebanks:
Towards an
Italian Stanford Dependency Treebank

Cristina Bosco - University of Turin
Simonetta Montemagni - Istituto di Linguistica Computazionale “Antonio Zampolli” (ILC–CNR)
Maria Simi - University of Pisa

The 7th Linguistic Annotation Workshop & Interoperability with Discourse
ACL Workshop, Sofia, August 8-9, 2013
Outline

• Building Italian Dependency Treebanks
  – a continuously evolving process
• *de jure* vs *de facto* standards in dependency annotation of corpora
• From MIDT to ISDT
  – The starting point
  – Comparison of the underlying annotation schemes
  – Conversion
  – Using ISDT for training a parser
• Open issues
• Current directions of research
Building Italian Dependency Treebanks
Building Italian Dependency Treebanks

Thanks to the conversion to the CoNLL format, TUT and ISST–exploited as reference resources in different evaluation campaigns:

- TUT in Evalita 2007, 2009 and 2011 Parsing Task
- ISST–CoNLL in CoNLL 2007 ST, Dependency Parsing, Multilingual Track
- ISST–TANL in
  - Evalita 2009 Parsing Task
  - Evalita 2011 Domain Adaptation Task
  - Shared task on Syntactic Parsing of Legal Text at LREC '12 (SPLeT '12)

Italian Syntactic-Semantic Treebank (ISST)

Turin University Treebank (TUT)

Merged Italian Dependency Treebank (MIDT)

Italian Stanford Dependency Treebank (ISDT)
Building Italian Dependency Treebanks

Thanks to the conversion to the CoNLL format, TUT and ISST–* exploited as reference resources in different evaluation campaigns:

- TUT in Evalita 2007, 2009 and 2011 Parsing Task
- ISST–CoNLL in CoNLL 2007 ST, Dependency Parsing, Multilingual Track
- ISST–TANL in
  - Evalita 2009 Parsing Task
  - Evalita 2011 Domain Adaptation Task
  - Shared task on Syntactic Parsing of Legal Text at LREC ’12 (SPLeT ’12)

- Small size resources usable in a restricted variety of tasks with an impact on the reliability of achieved results
- Italian NLP requires a larger resource
- Need to harmonize and merge existing resources
  - MIDT and ISDT as possible answers
de jure vs de facto standards

• Initiatives devoted to the definition of standards for the linguistic annotation of corpora
  – EAGLES (Expert Advisory Groups on Language Engineering Standards)
  – LAF/GrAF: ISO TC37/SC4 standardization activity aimed at defining a generic meta–model for linguistic annotation
    • SynAF: meta–model for syntactic annotation, including dependency structures
      – focus on the definition of a pivot format capable of representing diverse annotation types of varying complexity without providing specifications for the annotation of content categories
• ISOCat: the only initiative dealing with data categories
  – the set of ISOCat dependency categories: still basic and restricted
de jure vs de facto standards

• Initiatives devoted to the definition of standards for the linguistic annotation of corpora
  – EAGLES (Expert Advisory Groups on Language Engineering Standards)
  – LAF/GrAF: ISO TC37/SC4 standardization activity aimed at defining a generic meta–model for linguistic annotation
    • SynAF: meta–model for syntactic annotation, including dependency structures
  – focus on the definition of a pivot format capable of representing

For content categories
de facto standards are nowadays the only available option

STANFORD DEPENDENCY SCHEME
MIDT$^2$ ISDT: the starting point

• MIDT originates from a harmonization and merging effort applied to TUT and ISST-TANL (both CoNLL compliant)

• 3-steps methodology
  1. comparison of the underlying annotation schemes
     • TUT: 72 dependency types; ISST–TANL: 29 dependency types
     • substantial differences at the level of head selection criteria, annotation guidelines, treatment of specific constructions
  2. analysis of the performance of state–of–the–art dependency parsers using as training the original and target resources
     • the problem of semantically-oriented relations
  3. definition of the MIDT dependency tagset
     • “bridge” annotation scheme
     • automatically reconstructed by exploiting morpho–syntactic and dependency information contained in the original resources
MIDT\textsuperscript{2}ISDT: comparing the MIDT and SD annotation schemes

**Similarities**

Preference is given to relations which

- are semantically contentful and useful to applications
  - both neutralize the argument/adjunct distinction for what concerns prepositional complements, taken to be “largely useless in practice”

- link content words rather than being indirectly mediated via function words
  - both assign the head role to the content word in *determiner–noun* and *auxiliary–verb* constructions
MIDT2ISDT: comparing the MIDT and SD annotation schemes

Differences

Main dimensions of variation

- **inventory and granularity of dependency types**
  - 21 (MIDT) vs 48 (SD) dependency types
    - different types of mapping
    - relations *without a counterpart* in the other scheme
      - **MIDT-only**
        - Italian specific constructions, e.g. clitics in pronominal verbs
          - *la sedia si è rottata* lit. ‘the chair it is broken’ ‘the chair broke’
      - **SD-only** – absent in MIDT because of
        - the CoNLL format constraints (one-head-per-dependent)
        - their poor reproducibility in parsing (semantically oriented rels)

- **head selection criteria**
  - in sentential complements the head is the verb rather than the conj/prep
MIDT2ISDT: inventory and granularity of dependency types

Different types of mapping

MIDT

SD

mod

num

appos

1:n

nn

num

number

n:1

aux

abbrev

aux

lo ho mangiato
'I have eaten'

lo decisi di andare
'I decided to go'
MIDT2ISDT: comparing the MIDT and SD annotation schemes

Differences

Main dimensions of variation

- **inventory and granularity of dependency types**
  - 21 (MIDT) vs 48 (SD) dependency types
    - different types of mapping
    - relations **without a counterpart** in the other scheme
      - **MIDT-only**
        - Italian specific constructions, e.g. clitics in pronominal verbs
        - *la sedia si è rotta* lit. ‘the chair it is broken’ ‘the chair broke’
      - **SD-only** – absent in MIDT because of
        - the CoNLL format constraints (one-head-per-dependent)
        - their poor reproducibility in parsing (semantically oriented rels)

- **head selection criteria**
  - in sentential complements the head is the verb rather than the conj/prep
MIDT\textsuperscript{2} ISDT: conversion steps

**STEP 1** - recovering SD-relevant distinctions from MIDT parents

**STEP 2** – converting MIDT++ annotations into SD

**STEP 3** – using ISDT for training a parser

Steps 2 and 3 are iterated by interleaving conversion and automatic annotation
MIDT₂ ISDT: conversion step 2

Conversion patterns

• Structure-preserving mapping rules
  – dependency retyping only (tagset granularity differences)
    • 1:1 mapping, es. MIDT prep > SD pobj
    • 1:n mapping, es. MIDT mod > SD abbrev | amod | appos | nn | ...

• Tree restructuring mapping rules
  – head reassignment and dependency retyping (head selection criteria and tagset granularity differences)
    • 1:1 mapping, es. MIDT subj > SD csubj
    • 1:n mapping, es. MIDT arg > SD xcomp | ccomp
    • n:1 mapping, es. MIDT aux | prep > SD aux
Tree restructuring mapping rules - an example

Infinitival clause subcategorized for by the governing head

\[ \text{MIDT to ISDT: conversion step 2} \]

\[ \text{Infinitival clause subcategorized for by the governing head} \]

\[ \text{lit. ‘Giovanni told to-the judges to have paid the terrorists’} \]

\[ \text{‘Giovanni told the judges that he has paid the terrorists’} \]

\[ \text{MIDT: conversion step 2} \]

\[ \text{ISDT: conversion step 2} \]
MIDT2 ISDT: conversion step 2

- The conversion script: more than 100 rules

- Italian “localization” of the standard SD tagset
  - **Reductions**
    - prt: not relevant for Italian
    - xsubj: peculiarities of the original MIDT resource
  
  - **Extensions**
    - clit: used for dealing with clitics in pronominal verbs
    - nnp: proper noun modifiers

  - **Underspecification**
    - instead of resorting to the generic relation dep, the hierarchical organization of SD typed dependencies was exploited
      - comp and mod relations used when an appropriate relation could not be found among their dependency subtypes

- Revisions made were minimal to keep the scheme as close as possible to the original English version
MIDT2 ISDT: conversion step 2

- **nnp**: a new relation to specifically deal with proper noun modifiers as opposed to noun compound modifiers (nn)

Italian noun compounds: head-initial

Proper names: head-final (last name)
MIDT\textsuperscript{2} ISDT: conversion step 2

- The conversion script: more than 100 rules

- Italian “localization” of the standard SD tagset
  - Reductions
    - prt: not relevant for Italian
    - xsubj: peculiarities of the original MIDT resource
  - Extensions
    - clit: used for dealing with clitics in pronominal verbs
    - nnp: proper noun modifiers

- Underspecification
  - instead of resorting to the generic relation dep, the hierarchical organization of SD typed dependencies was exploited
    - comp and mod relations used when an appropriate relation could not be found among their dependency subtypes

- Revisions made were minimal to keep the scheme as close as possible to the original English version
MIDT ISDT: conversion step 3

Using ISDT for training a parser

• Data sets
  – SD conversion of the native resources
    • ISST–TANL-SD, TUT-SD
  – The merged resource or ISDT
    • ISST-TANL-SD.train + TUT-SD.train; ISST-TANL-SD.test + TUT-SD.test

• Parser
  – DeSR (Dependency Shift Reduce), a transition–based statistical parser (Attardi, 2006)
    • for these experiments we used a basic and fast variant of the DeSR parser based on Multi-Layer Perceptron (MLP)
**MIDT² ISDT: conversion step 3**

Using ISDT for training a parser

- Results

<table>
<thead>
<tr>
<th>Training</th>
<th>Test</th>
<th>LAS</th>
<th>LAS nopunct</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISST-TANL native</td>
<td>ISST-TANL native</td>
<td>82.09</td>
<td>Not available</td>
</tr>
<tr>
<td>TUT native</td>
<td>TUT native</td>
<td>89.88</td>
<td>Not available</td>
</tr>
<tr>
<td>ISST-TANL-SD</td>
<td>ISST-TANL-SD</td>
<td>80.55</td>
<td>82.11</td>
</tr>
<tr>
<td>TUT-SD</td>
<td>TUT-SD</td>
<td>84.14</td>
<td>85.57</td>
</tr>
<tr>
<td>TUT+ISST-TANL-SD</td>
<td>TUT+ISST-TANL-SD</td>
<td>83.34</td>
<td>84.16</td>
</tr>
<tr>
<td>TUT+ISST-TANL-SD</td>
<td>ISST-TANL-SD</td>
<td>79.94</td>
<td>81.86</td>
</tr>
<tr>
<td>TUT+ISST-TANL-SD</td>
<td>TUT-SD</td>
<td>84.14</td>
<td>85.79</td>
</tr>
</tbody>
</table>

Training with the larger combined resource did not appear to provide a substantial advantage so far.
MIDT\textsuperscript{2} ISDT: open issues (1)

Conversion of infinitival modifiers

- Current solution in ISDT in line with the SD standard scheme
  - problem: different internal structure of infinitival modifiers depending on the governing head
- Alternative solution to be explored
  - use prep-pcomp for all infinitival modifiers
  - different structure assigned to infinitival modifiers wrt to xcomp
    - the preposition makes it explicit the role of the infinitival modifier wrt the governing head

\textbf{MIDT}

same dependency (mod) regardless of the PoS of the governing head

\textbf{SD}

governing head=noun $> infmod$ (infinitival head=verb)
prep-pcomp for other clausal modifiers (clause head=preposition)
MIDT2 ISDT: open issues (2)

• The grey area of annotation guidelines
  – the current resource may still include inconsistent annotations due to “implicit” annotation guidelines, especially across the different sources
MIDT2 ISDT: conclusions

• The *Italian Stanford Dependency Treebank* (ISDT) is the newest achievement in the development of Italian dependency treebanks
  – Version 1.0 available for download from
    • [http://medialab.di.unipi.it/wiki/ISDT](http://medialab.di.unipi.it/wiki/ISDT)
  – Currently being used by Google in different applications (MT and Google voice)
  – Potentially a new module within the *Universal Stanford Dependency Treebank* [McDonald et al. ACL2013]

• Among the results
  – A methodology for harmonizing and converting treebanks
    • from 21 to 48 dependency types
  – An Italian “localization” of the Stanford Dependency scheme for what concerns
    • inventory of dependency types
    • annotation guidelines
MIDT2 ISDT: current developments

• An application-oriented ISDT version with collapsed dependencies and propagation of conjunct dependencies is being developed

• Semi-automatic validation of treebank annotations for the identification of problematic areas for human revision
  – errors as well as inconsistencies in annotation (deriving from different annotation “styles” for the same construction)

• Semi-automatic extension of the treebank to other textual genres and/or language varieties
  – starting from the output of dependency parsers
  – using algorithms for automatically identifying (un)reliable parses to minimize human annotators’ efforts in treebank construction
Building a Treebank is like Cathedral Building
Thanks!