Design, Construction and Use of an Italian Dependency Treebank: Methodological Issues and Empirical Results

Simonetta Montemagni
DyLan Lab

Istituto di Linguistica Computazionale “Antonio Zampolli”
ILC-CNR (Pisa, Italy)

Copenhagen Dependency Treebank Workshop - 23 August 2010
“Designing Treebanks”

Outline

- Treebanks: what for
- Basic requirements and annotation design
- Italian dependency Treebanks
  - ISST-* Italian dependency Treebank
  - the Turin University Treebank (TUT)
- Case studies
  - Head selection criteria
  - Argument/Adjunct distinction
  - Dependency tagset granularity
- Conclusions

Treebanks: what for

- Multiple uses of Treebanks
  - by linguists, to search for examples
  - by psycholinguists, to compute construction frequencies
  - by computational linguists, for tasks such as
    - training/tuning of statistical parsers
    - parser evaluation
    - induction of linguistic knowledge, e.g.
      - Syntactic subcategorisation frames
      - Predicate argument-structures
      - Selectional preferences
      - Word semantic classes
    - Information Extraction
      - Relation extraction

Basic requirements of the Treebank annotation scheme

- usability in both real applications and for research and/or educational purposes
- robustness and wide-coverage
- flexibility and customisability
- modularity
- reliability
- applicability in a coherent and replicable way
- portability to
  - different languages
  - different language varieties, e.g.
    - written vs spoken language
    - sublanguages
    - acquisitional data
- amenability to semi-automatic annotation

Basic requirements of the Treebank annotation scheme

- usability in both real applications and for research and/or educational purposes
- robustness and wide-coverage
- flexibility and customisability
- modularity
- reliability
- applicability in a coherent and replicable way
- portability to
  - different languages
  - different language varieties, e.g.
    - written vs spoken language

The design principles underlying the adopted annotation scheme heavily influence Treebank exploitation

DYNAMICS OF LANGUAGE

Focus on Dependency Treebanks

Increasing interest in **dependency-based representations** in recent years

- linguistically valuable
- more and more heavily used in NLP applications and tasks
- comparatively easy and "fair" to evaluate
- "lowest common ground" of a variety of different syntactic annotation schemes
- naturally multi-lingual
- abstract enough to deal with both spoken and written language, and any sort of non-canonical syntactic structure
- "lexical" enough in character to make provision for partial and focused annotation

Italian Dependency Treebanks

- Italian is:
  - a free constituent order language
  - a pro-drop language
- These two properties
  - make a constituency-based annotation of unrestricted texts difficult
  - can better be dealt within a dependency-based representation format, particularly suited for free word order languages

Existing Italian Dependency Treebanks for written language

1. **Italian Syntactic-Semantic Treebank (ISST)**
2. **Turin University Treebank (TUT)**

The **Italian Syntactic-Semantic Treebank (ISST)**: 1999-2001

- five-level structure covering orthographic, morpho-syntactic, syntactic and semantic levels of linguistic description
- syntactic annotation distributed over two different levels
  - constituent structure (89,941 tokens)
  - dependency structure (305,547 tokens)
- lexico-semantic annotation carried out in terms of sense tagging of lexical heads (69,972 identified semantic units)
  - reference resource: ItalWordNet

- corpus composition (305,547 tokens, time span: 1985-1995)
  1. a "balanced" corpus, testifying general language usage
     - 215,606 tokens
  2. a "specialised" corpus, with texts belonging to the financial domain
     - (89,941 tokens)

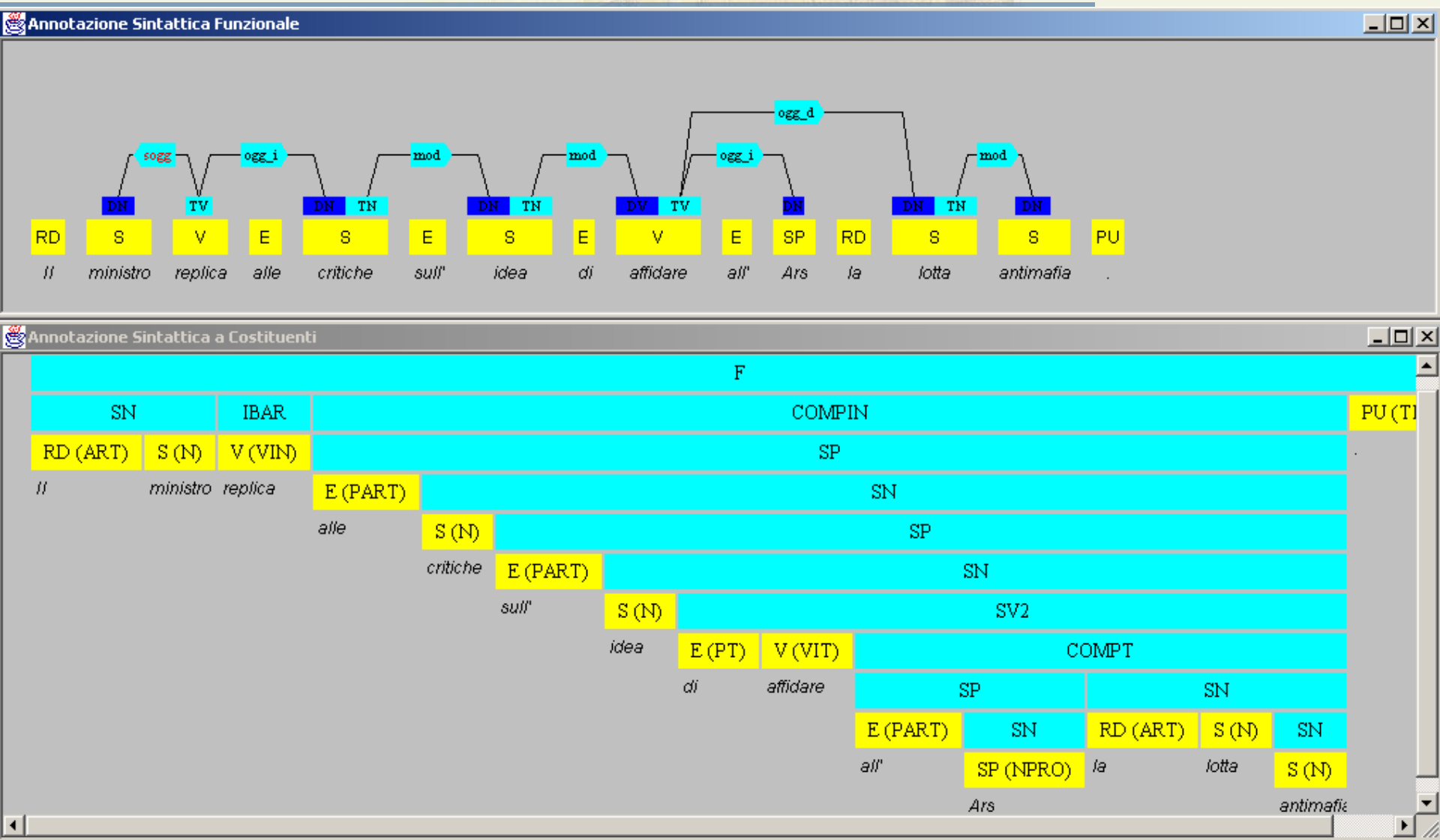
ISST-2001: the distributed approach to syntax (1)

- One of the main features of ISST wrt other Treebanks
  - monostratal view with two levels providing orthogonal views of the same surface syntax
  - independence and complementarity of the two annotation levels
    - none of them presupposes the other
    - combined views of the complementary information contained in them can be provided
- Motivations underlying this "double-track" approach
  - language-specific: optimal solution to tackle crucial issues of Italian syntax
  - usage-oriented: intended to be exploitable both in real applications and for research purposes

ISST-2001: the distributed approach to syntax (2)

- Constituent structure
  - Identification of phrase boundaries with labelling of constituents
  - Use of shallow structures
  - No use of empty elements (traces, pro-subjects) and/or coindexation
  - Annotation process carried out semi-automatically
- Dependency structure
  - Word-based
    - involving words belonging to major lexical classes only (i.e. non-auxiliary verbs, nouns, adjectives and adverbs)
    - information about grammatical words (e.g. determiners, prepositions, auxiliaries) encoded in terms of features
  - Dependency relations (e.g. subject, object), also involving displaced elements and null subjects
  - Annotation process carried out manually

ISST-2001: an example of the distributed approach to syntax



ISST-CoNLL: 2007

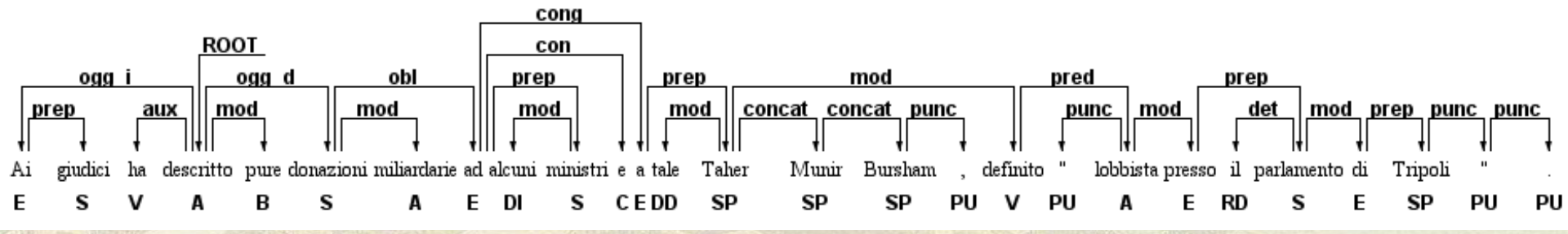
http://www.ilc.cnr.it/tressi_prg/ISST@CoNNL2007/ISST/ISST@CoNNL2007.pdf

- ISST was used as the starting point to build the Italian corpus for the CoNLL-2007 Shared Task
  - dependency parsing, multilingual track
- ISST-CoNLL was built through a semi-automatic conversion process combining information from the morpho-syntactic and dependency annotation levels (ILC-CNR, University of Pisa)
  - subset of the balanced ISST corpus of 79,654 word tokens (4,162 sentences) from the *Corriere della Sera* and *periodicals* partitions
- Conversion into the CoNLL-2007 tabular format had to
  - combine information coming from two different annotation levels
  - reconstruct dependency relations involving grammatical words from the ISST original annotation and revise accordingly the already existing dependency relations
  - other conversion issues concerned with
    - multi-headed tokens
    - empty tokens
    - identification of the sentence root
    - insertion of dependencies involving punctuation

ISST-CoNLL: 2007

http://www.ilc.cnr.it/tressi_prg/ISST@CoNNL2007/ISST/ISST@CoNNL2007.pdf

- ISST was used as the starting point to build the Italian corpus for the CoNLL-2007 Shared Task
  - dependency parsing, multilingual track
- ISST-CoNLL was built through a semi-automatic conversion process combining information from the morpho-syntactic and dependency annotation levels (ILC-CNR, University of Pisa)
  - subset of the balanced ISST corpus of 79,654 word tokens (4,162 sentences) from the *Corriere della Sera* and *periodicals* partitions
- Conversion into the CoNLL-2007 tabular format had to
  - combine information coming from two different annotation levels
  - reconstruct dependency relations involving grammatical words from the ISST original annotation and revise accordingly the already



ISST-TANL: 2007-2009

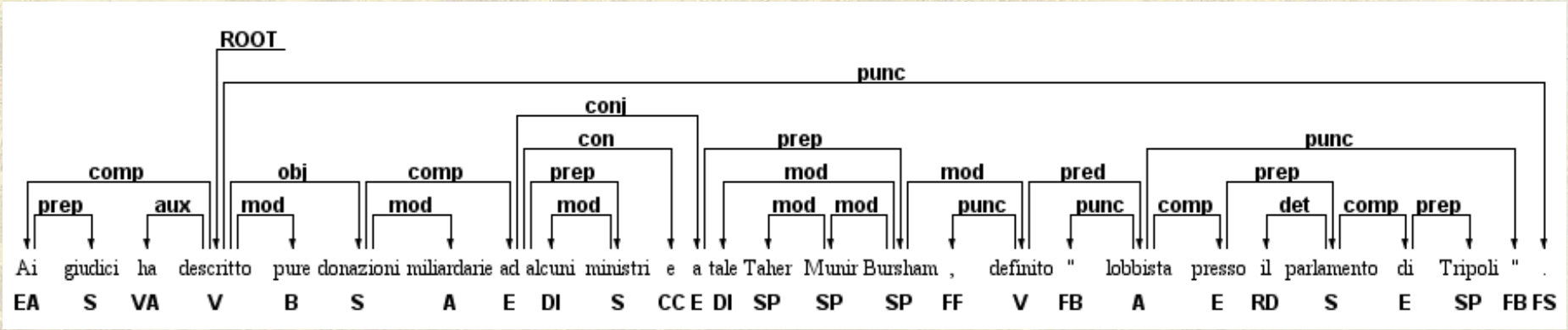
http://medialab.di.unipi.it/wiki/SemaWiki

- new release of the ISST-CoNLL corpus
  - ILC-CNR, University of Pisa
- revisions, all performed manually, mainly concerned with a reshaped dependency tagset and annotation criteria
  - neutralisation of the argument/adjunct distinction (restricted to prepositional complements)
  - linguistically-motivated treatment of punctuation
  - clitics
  - introduction of "semantically-oriented" distinctions
    - locative, temporal and indirect complements
    - locative and temporal modifiers
    - passive subject
    - "collapsed" version of the tagset neutralising such distinctions

ISST-TANL: 2007-2009

http://medialab.di.unipi.it/wiki/SemaWiki

- new release of the ISST-CoNLL corpus
  - ILC-CNR, University of Pisa
- revisions, all performed manually, mainly concerned with a reshaped dependency tagset and annotation criteria
  - neutralisation of the argument/adjunct distinction (restricted to prepositional complements)
  - linguistically-motivated treatment of punctuation
  - clitics
  - introduction of "semantically-oriented" distinctions



The **Turin University Treebank** (TUT):

http://www.di.unito.it/~tutreeb

- Dependency-based Treebank available in different formats
  - dependency representation
    - TUT native format
    - CoNLL format
  - constituency-based representation (Penn Treebank format)
  - more recently, also available in a format based on the Combinatory Categorial Grammar

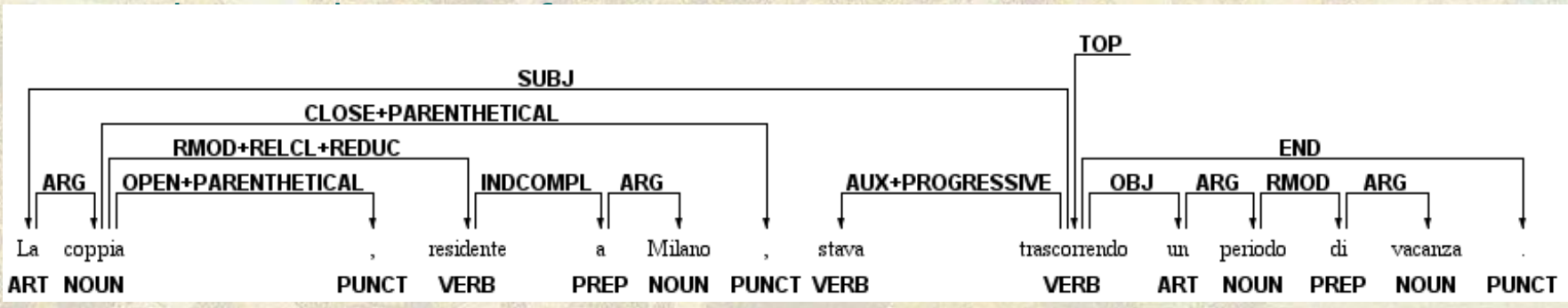
- Corpus composition
  - 2,400 sentences that correspond to 72,149 annotated tokens in TUT native format, and 66,055 tokens in CoNLL format
  - three subcorpora from
    - Italian newspapers (1,100 sentences and 30,561
    - the Italian Civil Law Code (1,100 sentences and 28,048 tokens)
    - the Italian section of the JRC-Acquis Multilingual Parallel Corpus, a collection of declarations of the European Community (200 sentences and 7,446 tokens)

The **Turin University Treebank** (TUT):

http://www.di.unito.it/~tutreeb

- Dependency-based Treebank available in different formats
  - dependency representation
    - TUT native format
    - CoNLL format
  - constituency-based representation (Penn Treebank format)
  - more recently, also available in a format based on the Combinatory Categorial Grammar

- Corpus composition
  - 2,400 sentences that correspond to 72,149 annotated tokens in TUT native format, and 66,055 tokens in CoNLL format



ISST-* vs TUT dependency annotation schemes

- Granularity and inventory of dependency types
  - ISST-*
    - ISST-2001: 10 dependency types augmented with features
    - ISST-CoNLL: 21 dependency types
    - ISST-TALN: 29 dependency types
  - TUT
    - 323 dependency types in the native TUT format
    - 72 dependency types in the TUT-CoNLL format
- Head selection
- Projectivity: TUT reduces the non-projective to projective structures, while ISST-* allows for non-projective representations
- Different annotation of various structures and phenomena, e.g.
  - coordination
  - punctuation
  - root constraint

DYNAMICS OF LANGUAGE

Case studies

- Head selection criteria
- Argument/Adjunct distinction
- Granularity of the dependency tagset

DYNAMICS OF LANGUAGE

Case study 1
Head selection criteria

- Criteria for a syntactic relation between a head H and a dependent D in a construction C [Zwicky 1985, Hudson 1990]
  1. H determines the syntactic category of C; H can replace C
  2. H determines the semantic category of C; D specifies H
  3. H is obligatory; D may be optional
  4. H selects D and determines whether D is obligatory
  5. The form of D depends on H (agreement or government)
  6. The linear position of D is specified with reference to H

- Issues:
  - Syntactic (and morphological) vs semantic criteria
  - Endocentric vs exocentric constructions
    - Economic news had little effect on [financial] markets
    - *Economic news had little effect on [markets]

(from Nivre-Kübler, Dependency Parsing Tutorial, COLING-ACL, 2006)

Head selection criteria: clear vs tricky cases (1)

Clear cases

- Exocentric constructions
  - **Verb**-subject
    - *Girls* **run**
  - **Verb**-object
    - The cat **drinks** *milk*
- Endocentric constructions
  - **Verb**-adverbial modifier
    - He **walked** *slowly*
  - **Noun**-adjectival modifier
    - *Economic* **news** affected *financial* **markets**

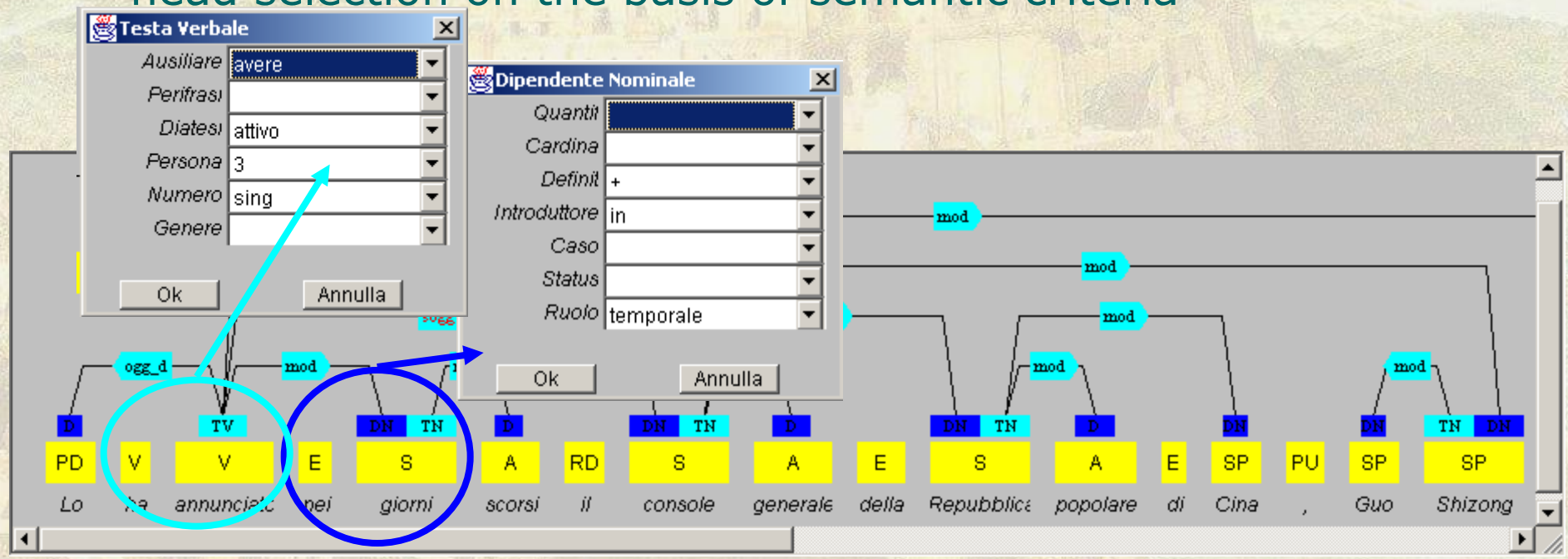
Head selection criteria: clear vs tricky cases (2)

Tricky cases

- Nominal phrases (determiner ↔ noun)
  - I **met** the girl
    - object: *the* or *girl*?
- Prepositional phrases (preposition ↔ noun)
  - I **arrived** to his house at 5pm
    - comp: *to* or *house*?
- Subordinate clauses (complementizer ↔ verb)
  - John **said** that he would have left soon
    - sentential comp: *that* or *left*?
- Complex verb groups (auxiliary ↔ main verb)
  - John has completed his job
    - sentence head: *has* or *completed*?

Head selection criteria in ISST-*, TUT and CDT

- **ISST-2001**
  - dependency relations between content words only
  - head selection on the basis of semantic criteria

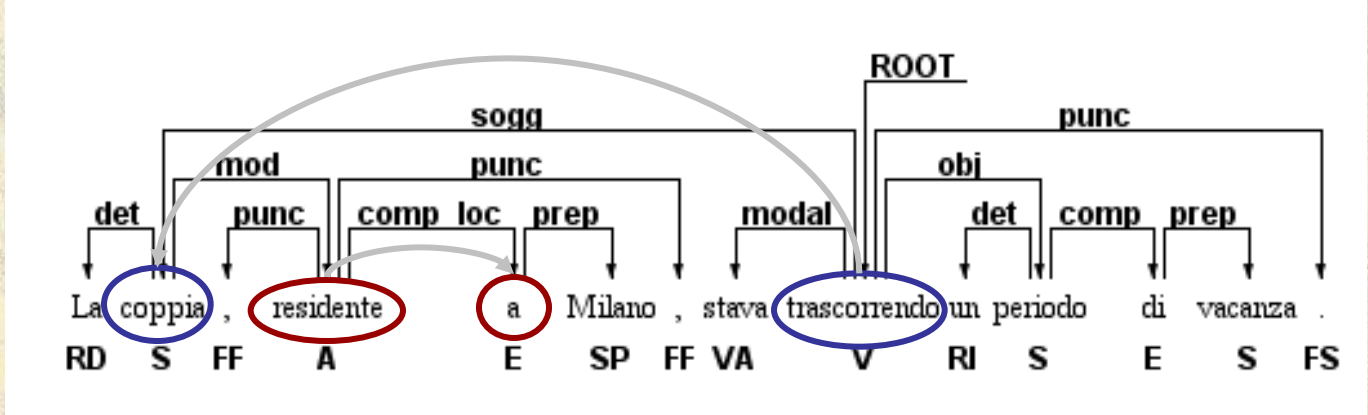


Lo ha annunciato nei giorni scorsi il console generale della Repubblica popolare di Cina, Guo Shizong

The general consul of the People's Republic of China, Guo Shizong, has announced it over the last days

Head selection criteria in ISST-*, TUT and CDT

- **ISST-CoNLL/TANL**
  - combination of syntactic and semantic criteria
    - in the determiner-noun and auxiliary-verb constructions the head role is assigned to the semantic head (noun/verb)
    - in preposition-noun and complementizer-verb constructions the head role is played by the element which is subcategorized by the governing head, i.e. the preposition and the complementizer

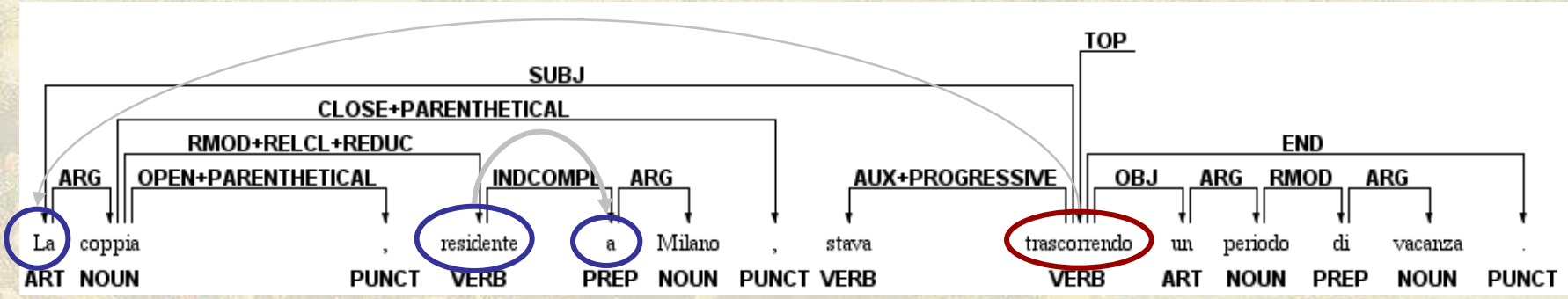


La coppia, residente a Milano, stava trascorrendo un periodo di vacanza

'The couple, living in Milan, was having a period of holiday'

Head selection criteria in ISST-*, TUT and CDT

- **TUT**
  - always assigns heads on the basis of syntactic criteria, i.e. in all constructions involving one function word and one content word (e.g. determiner-noun, preposition-noun, complementizer-verb) the head role is always played by the function word
  - only exception: in aux-main verb constructions the head role is played by the main verb

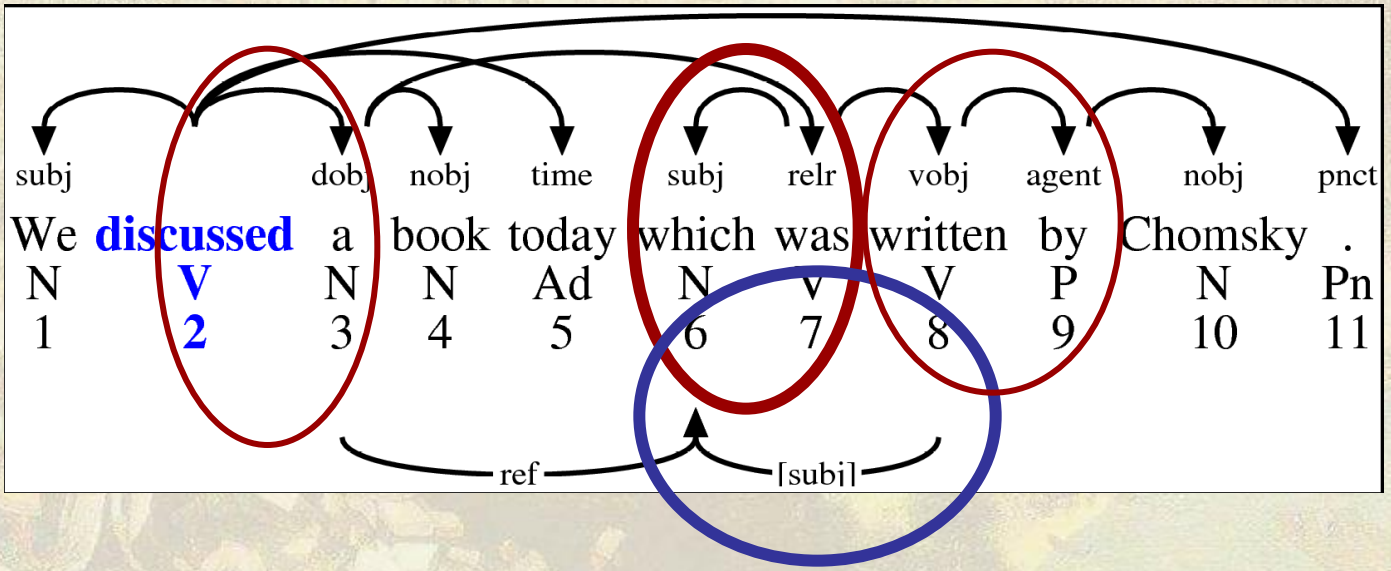


La coppia, residente a Milano, stava trascorrendo un periodo di vacanza

'The couple, living in Milan, was having a period of holiday'

Head selection criteria in ISST-*, TUT and CDT

- **CDT**
  - primary tree structure supplemented by an inventory of secondary relations
  - syntax-based head selection at the level of the primary tree structure



Head selection criteria: impact on Treebank usages (1)

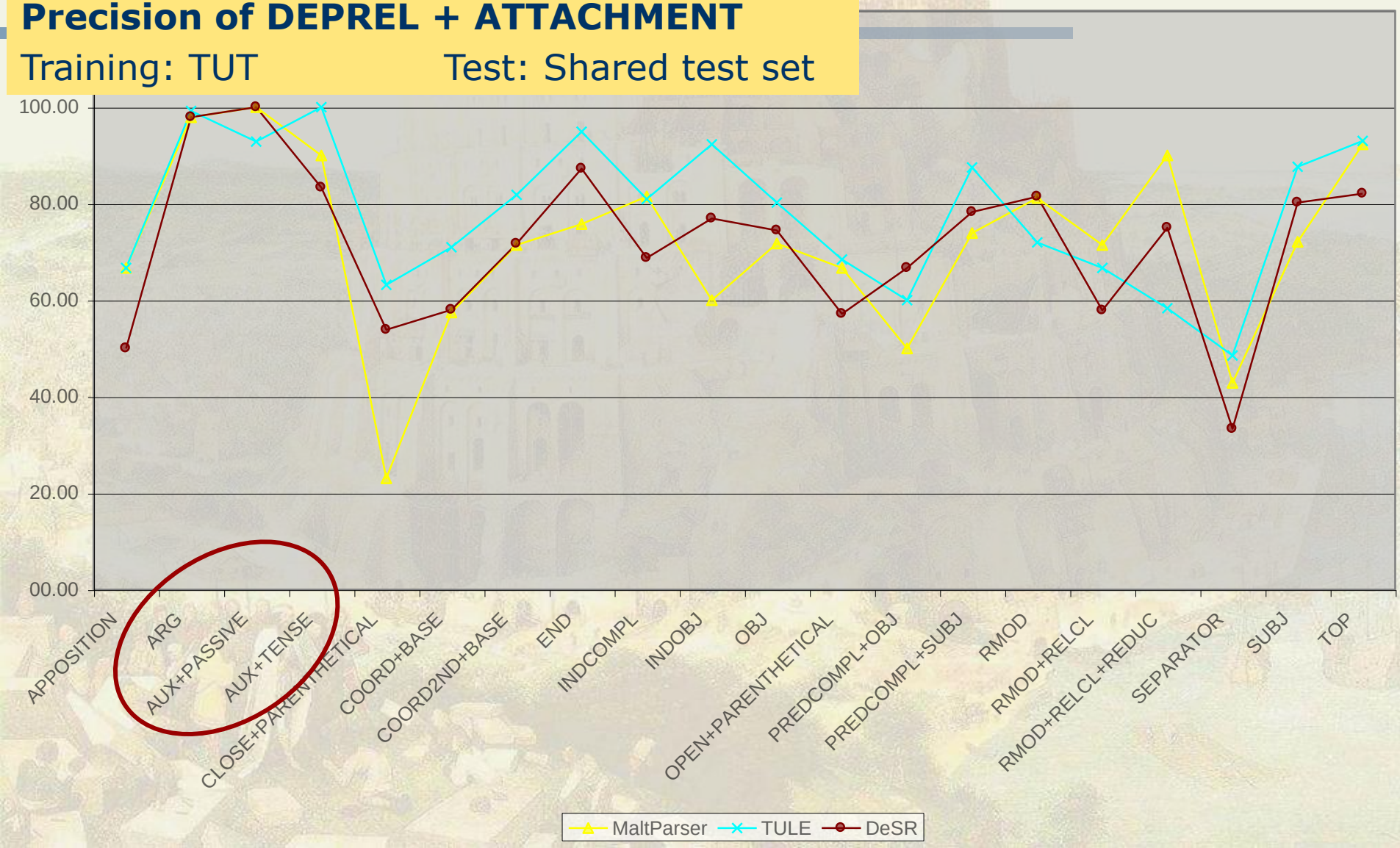
Usage 1: **Training of dependency parsers**

- EVALITA-2009 parsing evaluation campaign
  - dependency track
  - treebank-based
  - two subtasks based on two different treebanks: TUT (main subtask) and ISST-TANL (pilot subtask)
    - ideal testbed to start evaluating the influence of Treebanks on the performance of parsers
  - five parsing systems participated in both subtasks
    - 3 statistical parsers
    - 2 rule–based parsers
  - focus on
    - the results obtained by the three best performing systems
      - DeSR by Attardi et al.
      - MaltParser by Lavelli et al.
      - TULE by Lesmo (rule-based)
    - shared test set: 100 sentences
    - the dependency relations corresponding to the tricky cases

Head selection criteria: impact on training of dependency parsers

Precision of DEPREL + ATTACHMENT

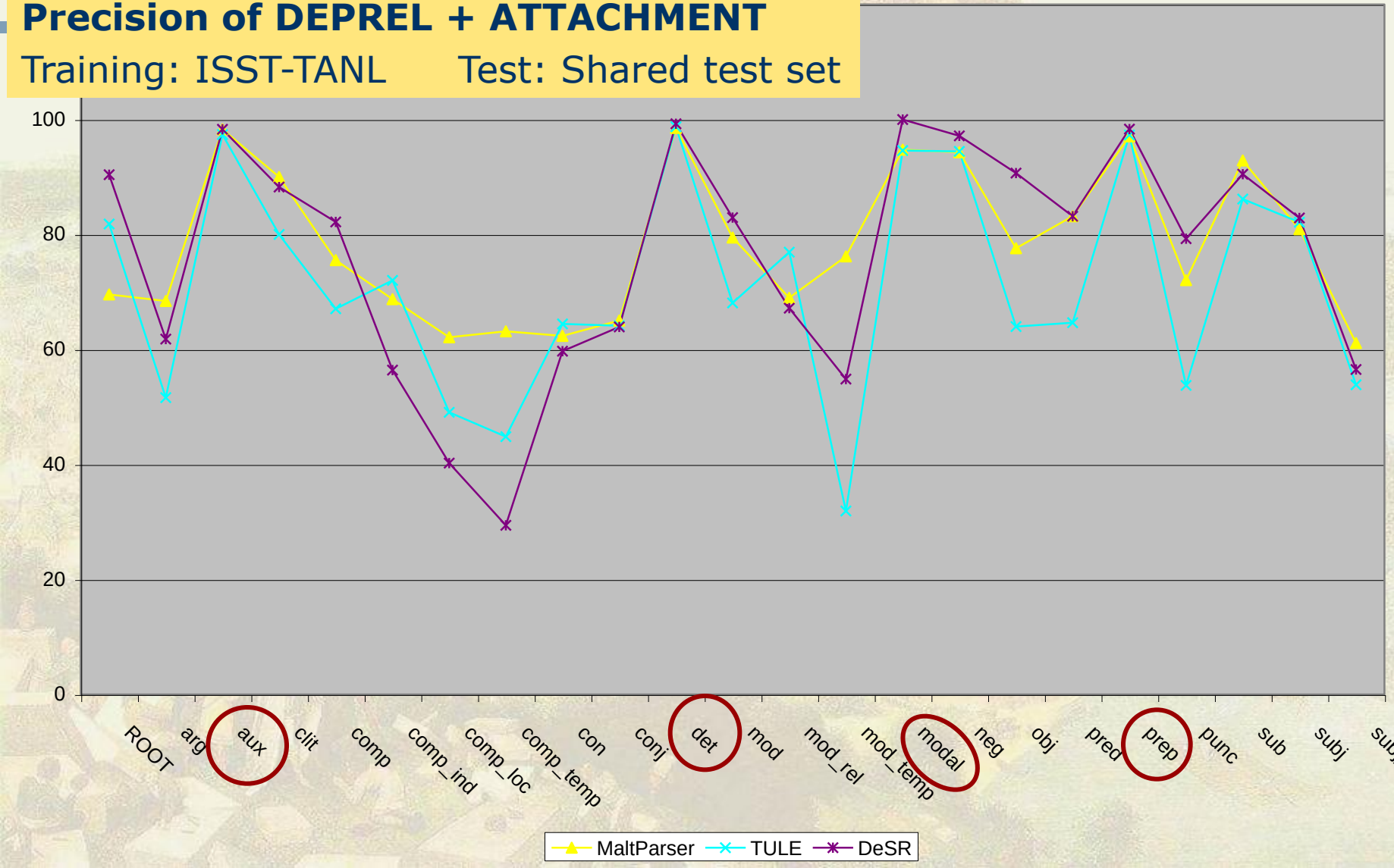
Training: TUT Test: Shared test set



Head selection criteria: impact on training of dependency parsers

Precision of DEPREL + ATTACHMENT

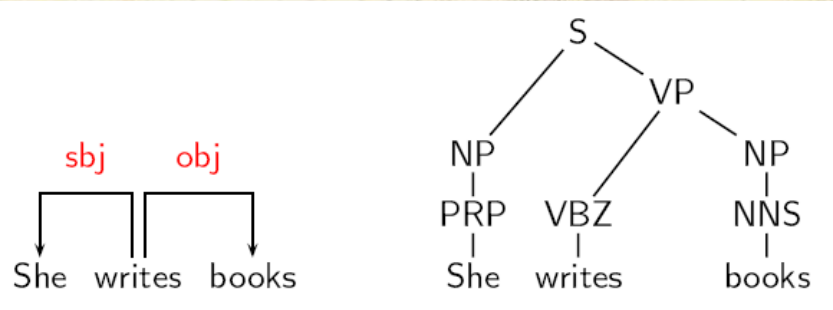
Training: ISST-TANL Test: Shared test set



Head selection criteria: impact on Treebank usages (2)

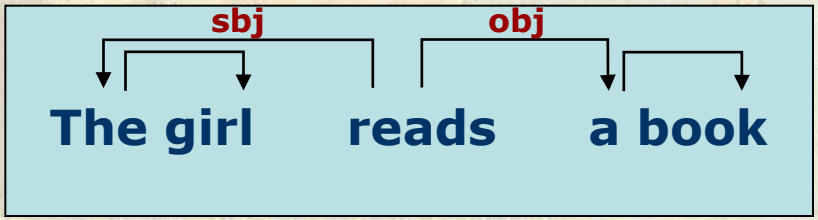
Usage 2: **Information Extraction**

- Dependency-based syntactic representations give a **transparent** encoding of predicate-argument structure



- Transparency is partially obscured when syntactic criteria are adopted
  - the extraction of relational information becomes trickier

Who is reading what?



Head selection criteria: impact on Treebank usages (2)

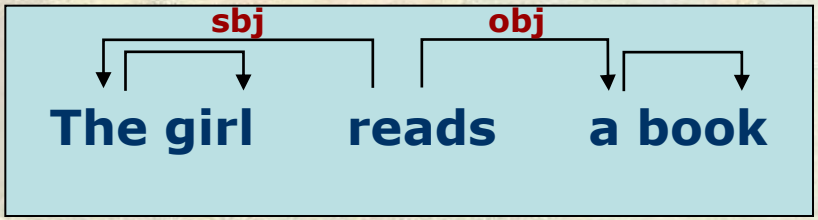
Usage 2: **Information Extraction**

- Dependency-based syntactic representations give a **transparent** encoding of predicate-argument structure

In the representation of semantically contentful relations suitable for relation extraction, dependency relations between content words should be preferred

(de Marneffe, Manning, COLING-2008)

Who is reading what?



Case study 2
argument/adjunct distinction (1)

- The distinction between arguments and adjuncts is typically accounted for in terms of lexical specification
  - arguments are lexically specified
  - adjuncts are not

Chris | ***gave*** | *Kim* | *some candy* | on Tuesday | in the park

- Some theories make a structural distinction as well
  - arguments are dependents of the predicate
  - adjuncts are dependents of the predication
  - such a distinction can be nicely accounted for in constituency-based syntactic representations
    - arguments are sisters to the head
    - adjuncts are sisters to a phrasal node
  - impossible to encode this distinction in the dependency structure
    - possible solution: encoding it at the level of the dependency label

Case study 2
argument/adjunct distinction (2)

- Central tenet in theoretical and computational linguistics as well as in lexicography
  - despite its importance, the exact nature of the distinction is difficult to characterize
  - very large gray area in which it is difficult to discriminate between arguments and adjuncts
    - *Kim changed the tire with a monkey wrench*
  - battery of linguistic criteria proposed in the literature aimed at identifying arguments vs adjuncts [Somers 1984, Grishman et al. 1994]
    - obligatoriness of arguments vs adjuncts
      - John put the book ***on the table*** *yesterday*
      - John put the book ***on the table***
      - *John put the book
    - Implied meaning: when omitted, the meaning of arguments is implied
      - I came to your house (***from x***) (*yesterday*)
      - I ate (***the pudding***) (*in the garden*)

Case study 2
argument/adjunct distinction (3)

- arguments can participate in diathesis alternations
  - John loaded ***the truck with hay*** – John loaded ***hay on the truck***
  - Ann threw ***the ball to me*** – Ann threw ***me the ball***
- "do so" test (Somers 1984)
  - John ran ***around the block*** *in the winter* and so did Alice
  - John ran ***around the block*** *in the winter* and so did Alice *in the summer*
  - *John ran ***around the block*** *in the winter* and so did Alice **around the reservoir**
- high relative frequency of arguments wrt predicates
  - I came ***from home***
  - I heard it *from you*
- …

- Unfortunately, it may occur that
  - there are not applicable criteria for a given context
  - different criteria can converge on different interpretations
    - corpus evidence from SketchEngine (ItWac Corpus – 2 GB tokens)
      - L'ho letto *sul* giornale 'I read it *on* the newspaper' (the location of the information is the most frequent co-occurring complement): argument?

The argument/adjunct distinction and Treebanks

The question is whether and how the argument/adjunct distinction should be accounted for in a Treebank

- "Fuzzy" distinction from the theoretical point of view
- Possible strategies in Treebank annotation
  1. annotators deal with it on the basis of
     - their intuition as native speakers
     - a battery of identification/discrimination criteria provided in the annotation guidelines
     - frequency in reference corpora
     - **potential problem: inconsistency of resulting annotation**
  2. annotators resort to a reference lexicon containing predicate-argument structure information
     - limited coverage of existing lexical resources
     - domain-specific distinction
     - **vicious circle: a Treebank is typically used to acquire information about the arguments selected by a predicate**

Argument/adjunct distinction and dependency Treebank annotation schemes

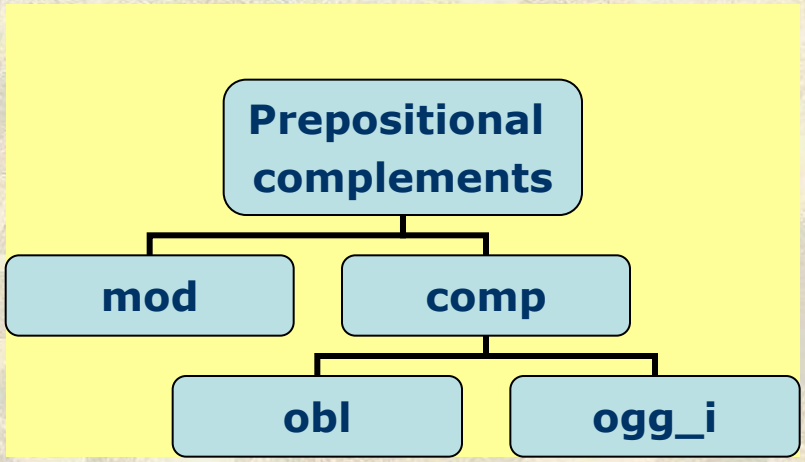
Different options are available to deal with this distinction in the design of a Treebank annotation scheme

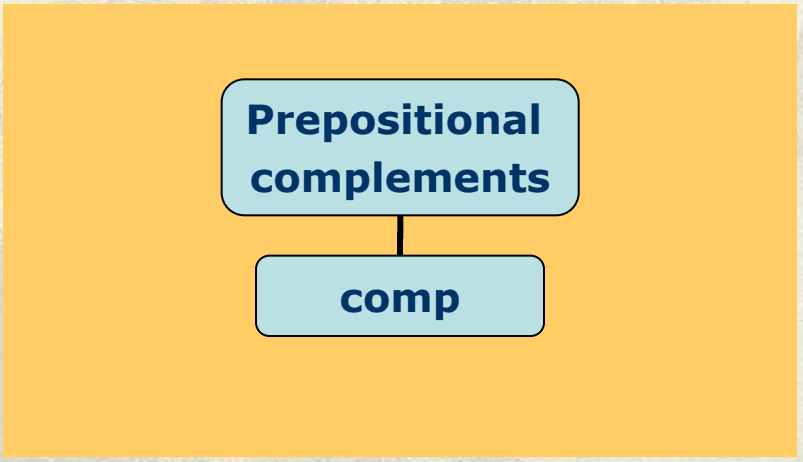
1. Key distinction of the annotation scheme
   - **TUT**
     - ARG(s): SUBJ, OBJ, INDCOMPL, INDOBJ, PREDCOMPL, EXTRAOBJ
     - MODIFIER(s): RMOD, APPOSITION
   - Similar distinction in **CDT**
2. Middle-ground solution
   - the distinction is explicitly represented in clear cut cases
   - underspecified representation to deal with problematic cases
     - **ISST-2001**, **ISST-CoNLL**
       - mod, comp, ogg_i, obl vs comp
3. Distinction is neutralised
   - **ISST-TANL**, where COMP covers all relations between a head and a prepositional complement, whether a modifier or a subcategorized argument

Argument/adjunct distinction: impact on Treebank usages (1)

Usage 1: **Training of dependency parsers**

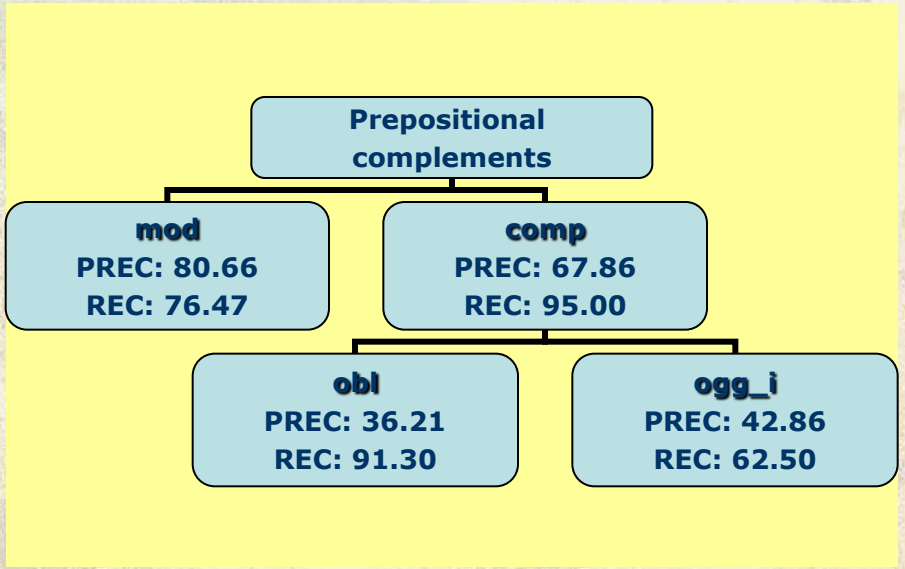
- Used parser
  - DeSR (Attardi, Dell'Orletta 2009)
    - transition-based statistical parser
    - trained using SVM and Multilayer Perceptron
    - state-of-the-art technology for Italian dependency parsing
- Training corpora
  - **ISST-CoNLL** vs **ISST-TANL**



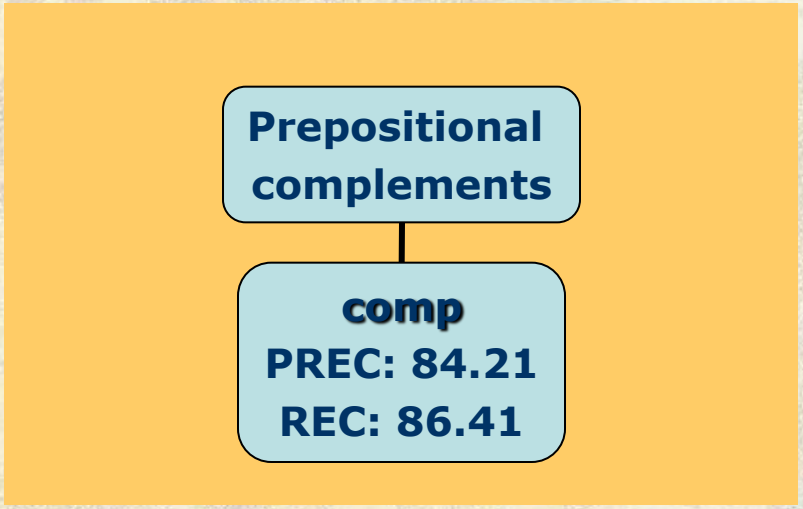


Argument/adjunct distinction: impact on training of dependency parsers

Training: **ISST-CoNLL**



Training: **ISST-TANL**



Significantly more reliable parsing results are achieved by neutralising such a distinction

Argument/adjunct distinction: impact on Treebank usages (2)

Usage 2: **Information Extraction**

The Stanford typed Dependency annotation scheme is not concerned with the argument/adjunct distinction which is largely useless in practice

(de Marneffe, Manning, COLING-2008)

- **Bio-text mining**
  - **particular requirements for Subcategorization Frames (SCFs) in biomedical language**
    - **SCFs should not be restricted to arguments but should also include strongly selected modifiers (such as location, manner and timing), as these are deemed to be essential for the correct interpretation of texts**
  - **a looser notion of SCF is required, which includes typical verb modifiers in addition to strongly selected arguments**
    - **no a priori knowledge about the set of possible SCFs**

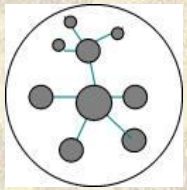
Argument/adjunct distinction and Treebanks

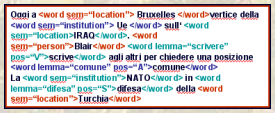
The question is whether the argument/adjunct distinction should be handled in a Treebank

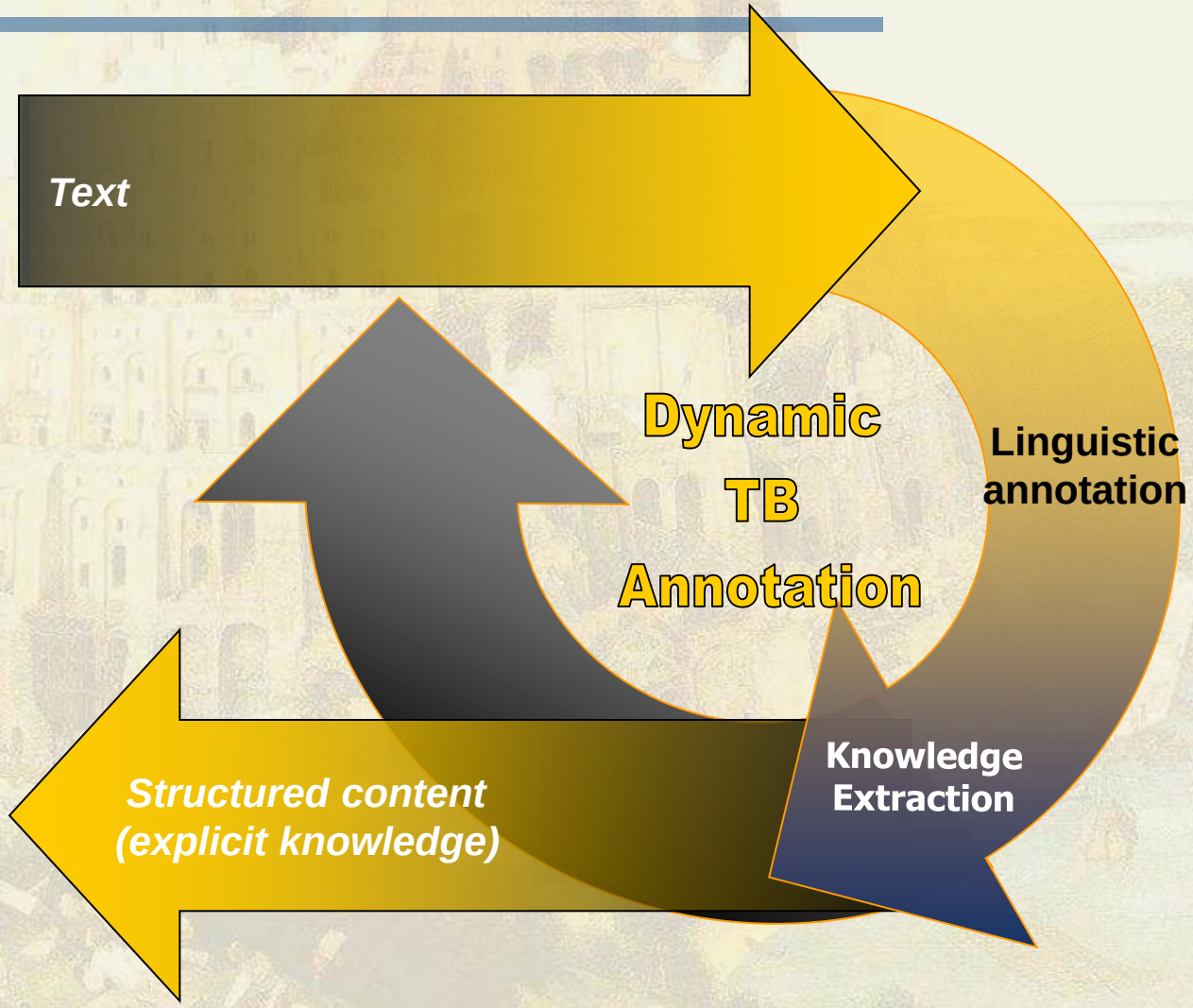
NO!

- This does not question the validity of the distinction, motivated linguistically and psycho-linguistically
- Simply, due to
  - the absence of reliable classification criteria
  - the domain-specificity of the notion
  - the fact that Treebanks are typically exploited to bootstrap linguistic knowledge
- Such a distinction should be dealt with in a postprocessing stage in which **reliably produced underspecified representations** are exploited to derive such knowledge which could be reprojected back to the corpus

Argument/adjunct distinction and Treebanks





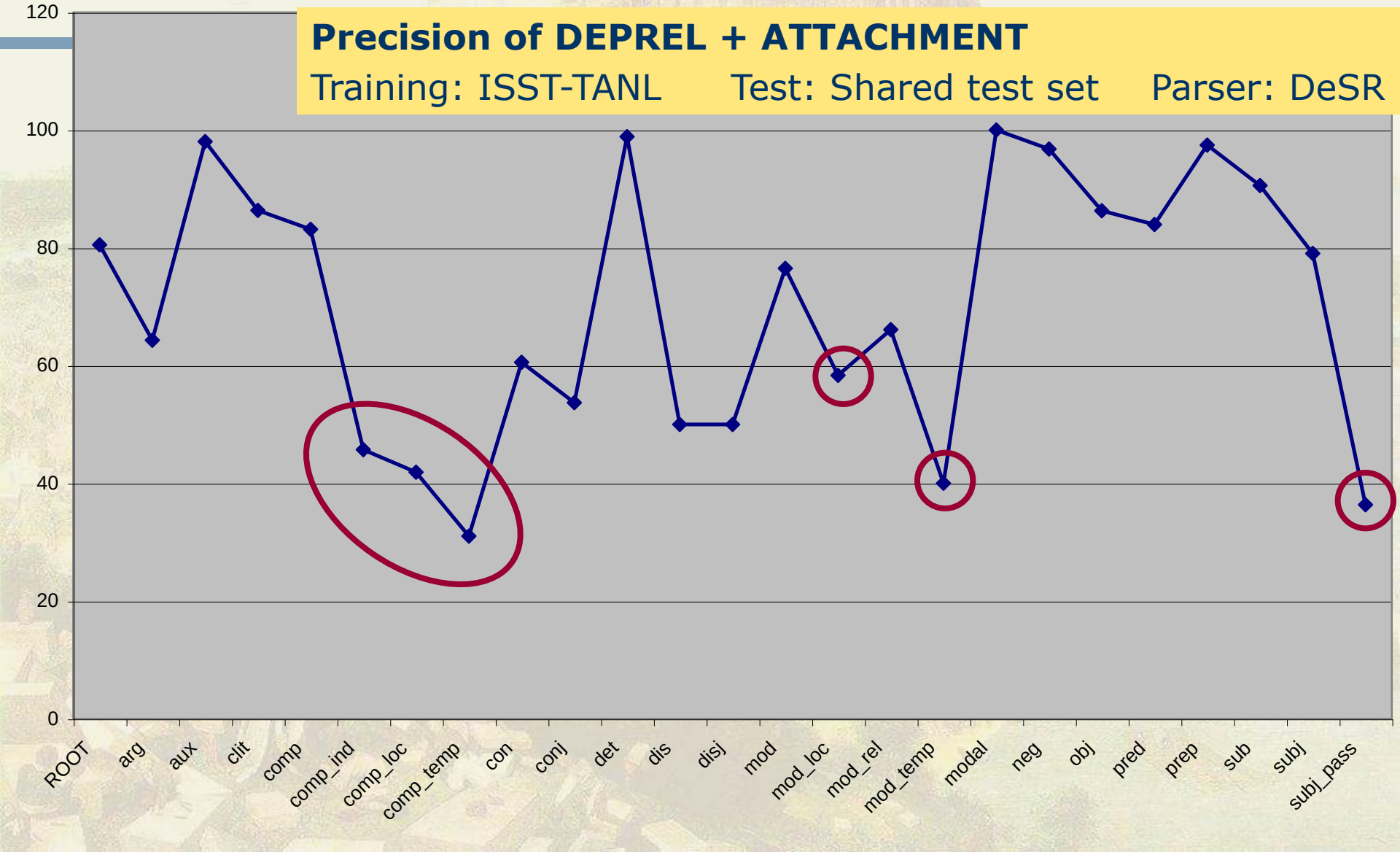


Case study 3
Granularity of the dependency tagset

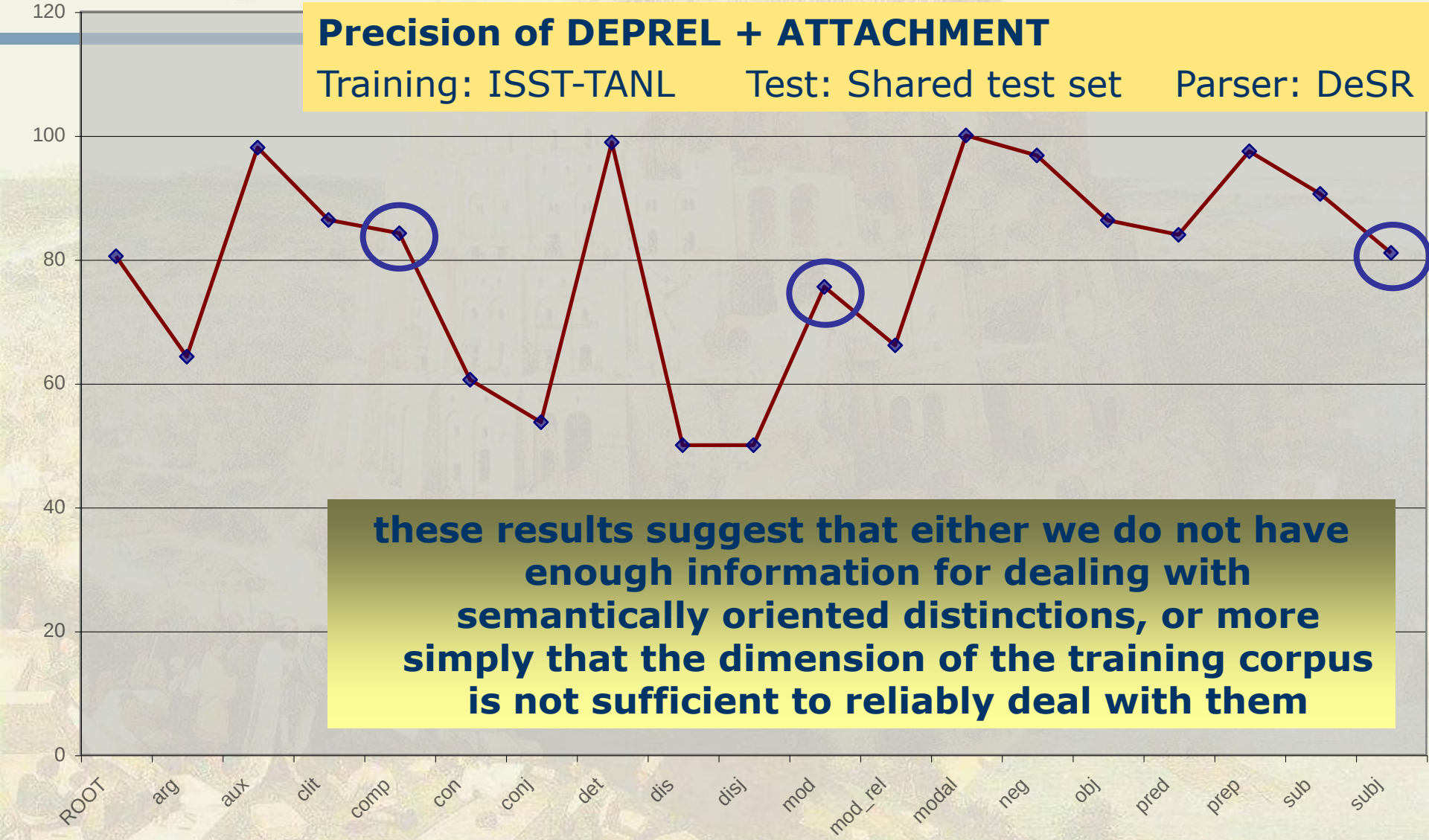
Is it always the case that more is better?

- General desiderata for a tagset
  - Capture interesting linguistic categories
  - Be predictable/learnable for automatic parsers
- High variability in the size of the dependency tagset
  - From 323 relations in TUT to 10 relations in ISST-2001
- What is the appropriate level of tagset granularity to meet the abovementioned desiderata?
  - Interleaving tagset design and parsing experiments
- The ISST-TANL tagset makes available two options
  - Introduce Semantically loaded distinctions
    - Comp_temp/loc/ind, mod_loc/temp, subj_pass
  - Neutralise them

Parser performance with **fine-grained distinctions**



Parser performance with **neutralised** distinctions



these results suggest that either we do not have enough information for dealing with semantically oriented distinctions, or more simply that the dimension of the training corpus is not sufficient to reliably deal with them

Conclusions

Desiderata

usability for research and applicative purposes

robustness and wide-coverage

flexibility and customisability

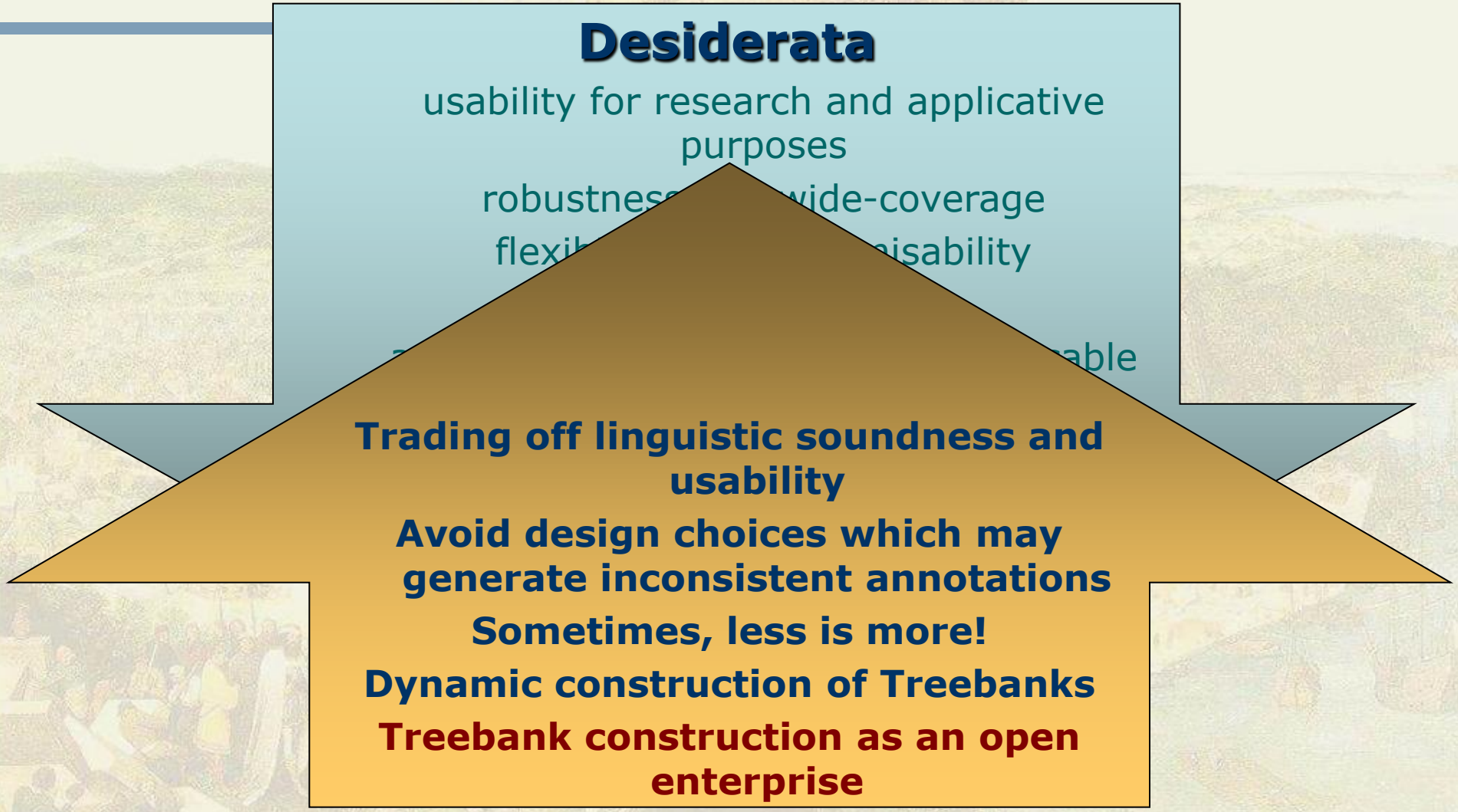
reliability

applicability in a coherent and replicable way

portability to different language varieties

amenability to semi-automatic annotation

Conclusions



ISST-* distribution

- ISST-2001 is distributed through the ***European Language Resources Association*** (**ELRA**)
  - http://www.elra.info or http://www.elda.org
  - Benefits
    - Servicing of bug reporting through ELRA
    - Organisational embedding into other lexical resources
    - Long-term availability
    - Support to European language infrastructures
  - Different licence types for
    - Research use
    - Commercial use
  - Soon also ISST-TANL will be available

- ISST-CoNLL is going to be distributed by the Linguistic Data Consortium (LDC) as part of the 2007 CoNLL Shared Task Data

Credits

Felice Dell'Orletta
Alessandro Lenci
Eva Maria Vecchi

DyLan Lab

Lab for computational models of the dynamics of language and cognition