VIT - Venice Italian Treebank: Syntactic and Quantitative Features

Rodolfo Delmonte, Sara Tonelli, Antonella Bristot

Department of Language Science
Laboratory Computational Linguistics
Università "Ca Foscari"
30124 - VENEZIA
Tel. 39-041-2345717/52
E-mail: delmont@unive.it
Website: http//project.cgm.unive.it

Outline

- Genesis
- General Linguistic Issues
- Comparison with other Treebanks
- Peculiarities of Italian Language
- Quantitative Data
- Conversion to Depency Structure

Projects of Treebanks of Italian for VIT

- **Project DIGITAL EQ. 1986-88**
- **Manual annotation of a corpus of written Italian at the level**
  - **syntactic**
    - **Constituency structure (~100.000 words)**
- **Internal projects**
- **Automatic annotation of a corpus of written Italian at level - ISST and others**
  - **syntactic**
    - **Constituency structure (~170.000 words)**

National Projects of Italian Treebanks

- Project SITAL

Annotation of a corpus of written Italian at the following levels

- syntactic
  - Constituency structure(~90.000 words)
  - Functional structure (~300.000 words)
- Lexical-semantic (~80.000 open/content words, distribuited amongst nouns, verbs and adjectives)

- Project AVIP/IPAR

Annotation of a corpus of regional Italian at the following levels

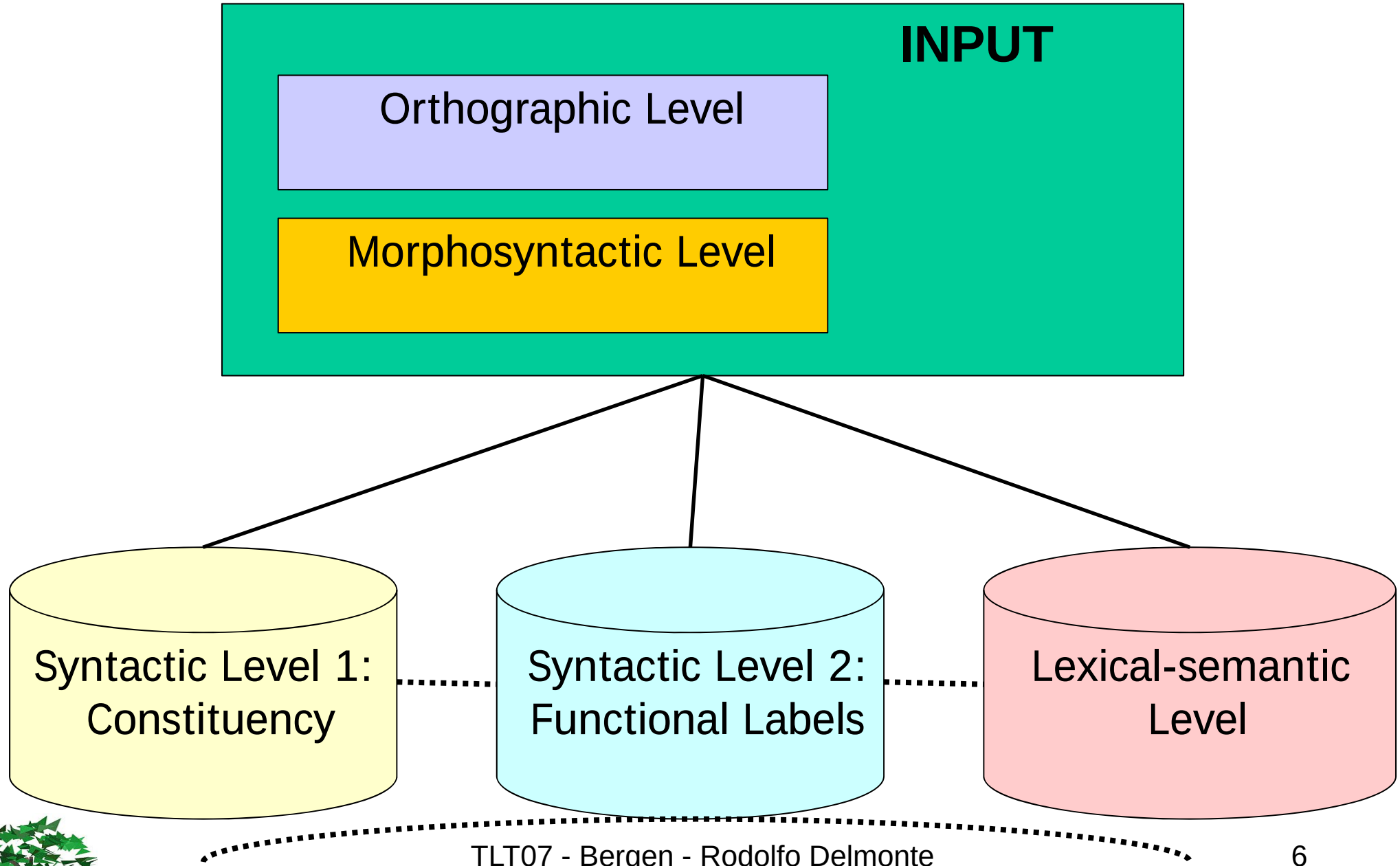
syntactic

- Functional and constituency structure (~60.000 words)

Syntactic-semantic treebank of Italian (ISST)

- annotated at the following levels:
  - orthographic, with indication of macrotextual organization
  - morpho-syntactic, with indication of lemma and of basic multi/poliwords expressions (es. *ad hoc*, *allo scoperto*, *al di là*)
    - Annotation has been validated manually
    - Annotation scheme is compliant with existing standards (EAGLES) and common/shared with the other national project “Annotated Dialogues”
- format of representation: XML accompanied with appropriate DTD

Annotation guidelines: architecture of ISST Treebank



Corpus API/AVIP and DIFFERENCES

- Spoken transcribed Regional Varieties of Italian
- DIFFERENCES
  - Total tokens = 4282 subdivided in:
  - punctuation and turn markers = 1637 tokens
  - words, interjections, quasi words etc. = 2645 tokens
- API/AVIP
  - Total tokens = 56337 subdivided in:
  - punctuation and turn markers = 18710 tokens
  - words, interjections, quasi words etc. = 37627 tokens

Corpus API/AVIP and DIFFERENCES

- The most interesting feature to study was Overlaps
- API/AVIP
  - 1100 OVERLAPS
  - 6849 UTTERANCES
  - 4747 TURNS
- DIFFERENCES
  - 147 OVERLAPS
  - 371 UTTERANCES
  - 336 TURNS

Features of Treebanks Relevant for Machine Learning

- Representativeness in terms of text genres
- Representativeness in terms of linguistic theory adherence
- Coherence in allowing Syntactic-Semantic Mapping
- Eventually the distinctive linguistic features of the chosen language

Features of Treebanks Relevant for Machine Learning

- Balanced Corpus Representative of 6/7 different text genres vs. Unbalanced
- Strictly adherent to linguistic principles vs, loosely adherent (e.g. more hierarchical vs. less hierarchical)
- Constituency/Dependency/Functional structures are semantically coherent vs. incoherent
- Language chosen is highly canonical and regular vs. almost free word order language

Criteria inspiring constituency structure in : X-bar

- **Theoretic Schema for X-bar rules**
- CP --> Spec, Cbar

  Spec --> C0
  C0 --> Complementizer
  Cbar --> Adjuncts, XP
  XP --> Spec, Xbar

  Spec--> Subject
  Xbar --> X, Complements
  X --> Verb, Adjective, Noun, Adverb

Criteria inspiring constituency structure in : X-bar

- **NP Specifier: Atomic vs Structured**

  Spec--> Determiners, Quantifiers, Intensifiers

- ***Structure of Verbal Compound***

  Xbar --> Verb - auxiliaries, modals, clitics, negation, adverbials, prepositional phrases, conjunctions

Less generic Schema for X-bar rules

CP --> SpecCP, Cbar

SpecCP -> Adjuncts, Fronted Complements, Focussed Arguments, Dislocated Constituents

Cbar --> C1, IP

Cbar --> C0, CP

C0 --> Complementizer

C1 --> Wh+ word

Less generic Schema for X-bar rules

IP --> SpecIP, Xbar, Complements, Adjuncts , Dislocated Constituents

SpecIP --> Subject

Complements --> COMPT/COMPIN/COMPC/COMPPAS

Xbar --> VerbalCompound

Spec --> Adverbials, Quantified Structures, Preposed Constituents

- F3 --> Fragments
- Distinction between Tensed and Untensed Clauses
- CLAUSE = Semantically transparent syntactic nucleus corresponding to a Semantic Proposition with PAS

TYPOLOGY OF SYNTACTIC CONSTITUENTS

STRUCTURAL CONSTITUENTS

F	sentence
F3	sentence fragment
CP	Dislocated/preposed constituents, adjunct constituents
CP_INT	Dislocated/preposed constituents, adjunct constituents
TOPF	Auxiliary complex constituents
COMPT	Complements governed by Transitive Verbs
COMPIN	Complements governed by Intrasitive Verbs
COMPC	Complements governed by Copulative Verbs
COMPPAS	Complements governed by Passive Verbs
FP	Parenthetical, Apposition with punctuation Šadjunct constituents
DIRSP	Direct speech with punctuation Šany constituent

TYPOLOGY OF SYNTACTIC CONSTITUENTS

LEXICAL FUNCTIONAL CONSTITUENTS

FAC	Complement sentence with/without complementizer (SÕ)
FC	Coordinated sentences with conjunction (SÕ)
FS	Subordinated sentence with subordinator (SÕ)
FINT	Interrogative sentence with/without interrogative pronoun (SÕ)
F2	Relative Clause with relative pronoun (SÕ)
COORD	Coordinate structure for constituents heads with conjunction or punctuation (COORD)
SC	Comparative/Quantified Phrase with conjunction (QP)
SP	Prepositional Phrase with preposition (PP)
SQ	Quantified Phrase with quantifier (QP)
SPD	Prepositional Phrase with preposition (PP) DI (of
SPDA	Prepositional Phrase with preposition (PP) Adjunct

TYPOLOGY OF SYNTACTIC CONSTITUENTS

SUBSTANTIAL CONSTITUENTS

SN	Nominal Phrase (NP) with F2 headed by Indefinite Relative Pronouns
SA	Adjectival Phrase (AP)
SAVV	Adverbial Phrase (ADVP)
AUXCC	Verbal Group (VP) with tensed auxiliary
IBAR	Verbal Group (VP) with tensed verb
IR_INFL	Verbal Group (VP) with tensed verb
SV2	Infinitival Clause (VP)
SV3	Participial Clause (VP)
SV5	Gerundive Clause (VP)

Guidelines for syntactic constituent annotation

- identifying phrasal constituents and their relations of hierarchical embedding
- assignment of specific syntactic category to individual constituents
- Annotation criteria of strict adherence to semantic transparency, mainly as regards the annotation of complex syntactic constructions

UPenn Treebank criteria

- Our approach to developing the syntactic tagset was highly pragmatic and strongly influenced by the need to create a large body of annotated material given limited human resources. The original design of the Treebank called for a level of syntactic analysis comparable to the skeletal analysis used by the Lancaster Treebank... no forced distinction between arguments and adjuncts. A skeletal syntactic context-free representation (parsing).

Example from Upenn Treebank

- In exchange offers that expired Friday, holders of each $1,000 of notes will receive $250 face amount of Series A 7.5% senior secured convertible notes due Jan. 15, 1955, and 200 common shares.

```
( (S (PP-LOC In
             (NP (NP exchange offers)
                    (SBAR (WHNP-1 that)
                          (S (NP-SBJ *T*-1)
                                (VP expired
                                   (NP-TMP Friday))))))
    ,
    (NP-SBJ (NP holders)
           (PP of
              (NP (NP each $ 1,000 *U*)
                 (PP of
                    (NP notes)))))
    (VP will
       (VP receive
          (NP (NP (NP (ADJP $ 250 *U*) face amount)
                 (PP of
                    (NP (NP Series A
                       (ADJP 7.5 %)  senior secured convertible notes)
                       (ADJP due
                            (NP-TMP (NP Jan. 15)    ,
                                                    (NP 1995))))))
                        and
              (NP 200 common shares))))
.))
```

```
( (CP (PP-LOC In
            (NP (NP exchange) offers
                    (CP (WHNP-1 that)
                      (S (IBAR expired)
                              (COMPIN (NP-TMP Friday))))))
    ,
     (S (NP-SBJ (NP holders
           (PP of
              (NP (QP each) $ 1,000 *U*
                   (PP of
                    (NP notes))))))
    (IBAR will receive)
        (COMPT (COORD (NP (NP (ADJP $ 250 *U*)  face amount)
                 (PP of
                    (NP (NP Series A
                       (ADJP 7.5 %)
                          (ADJP senior secured convertible)
                                notes)
                       (ADJP due
                            (NP-TMP (NP Jan. 15)
                                    ,
                                    (NP 1995))))))
                        and
              (NP 200 common shares)))))
.)
```

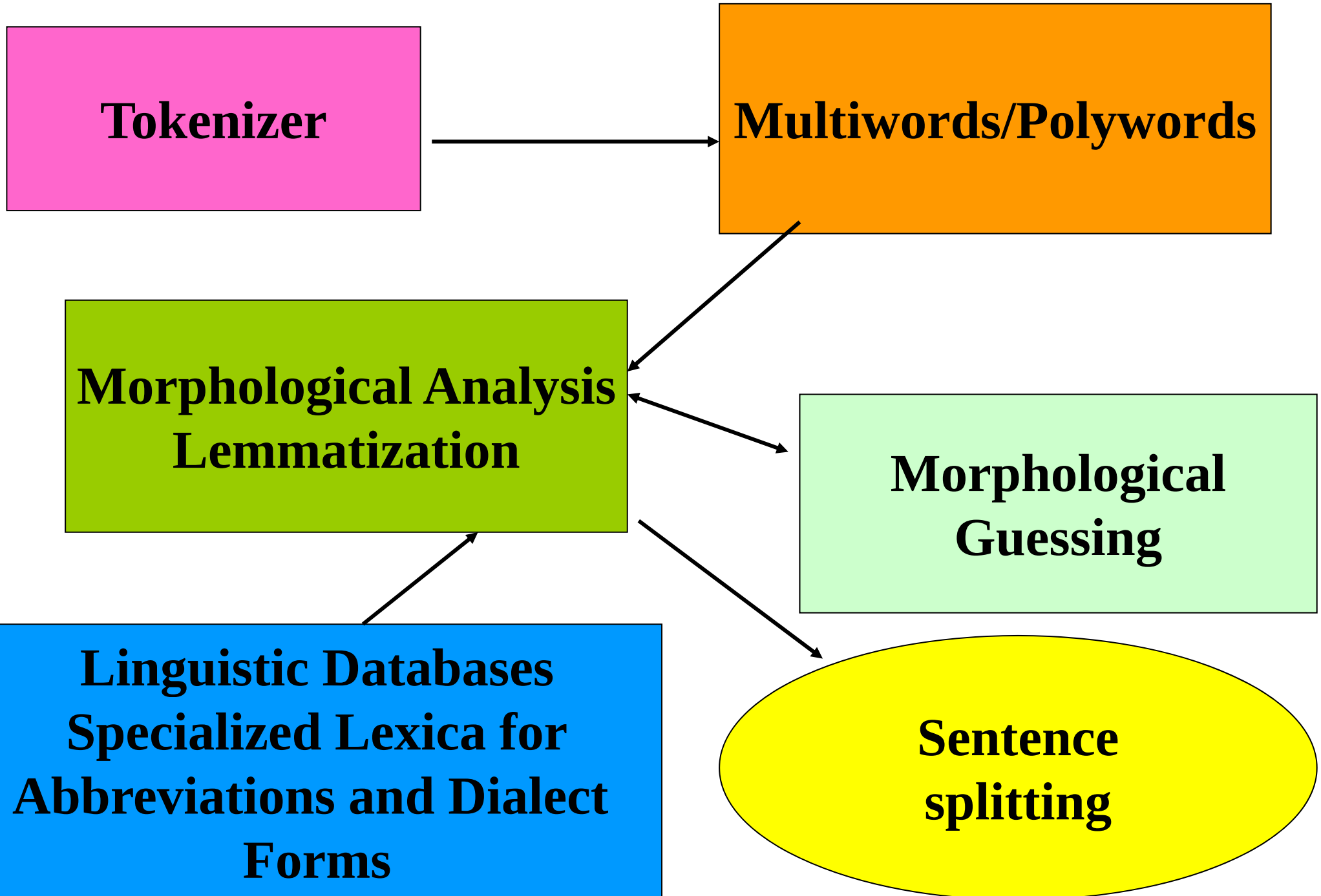
NEGRA Treebank

- Separate constituent for Inflected Verb
- No use of S-BAR
- Only Chomsky-adjunction
- No provision for Verb-Second structures and Inversion
- Fronted auxiliaries and modals are split from their verbal heads

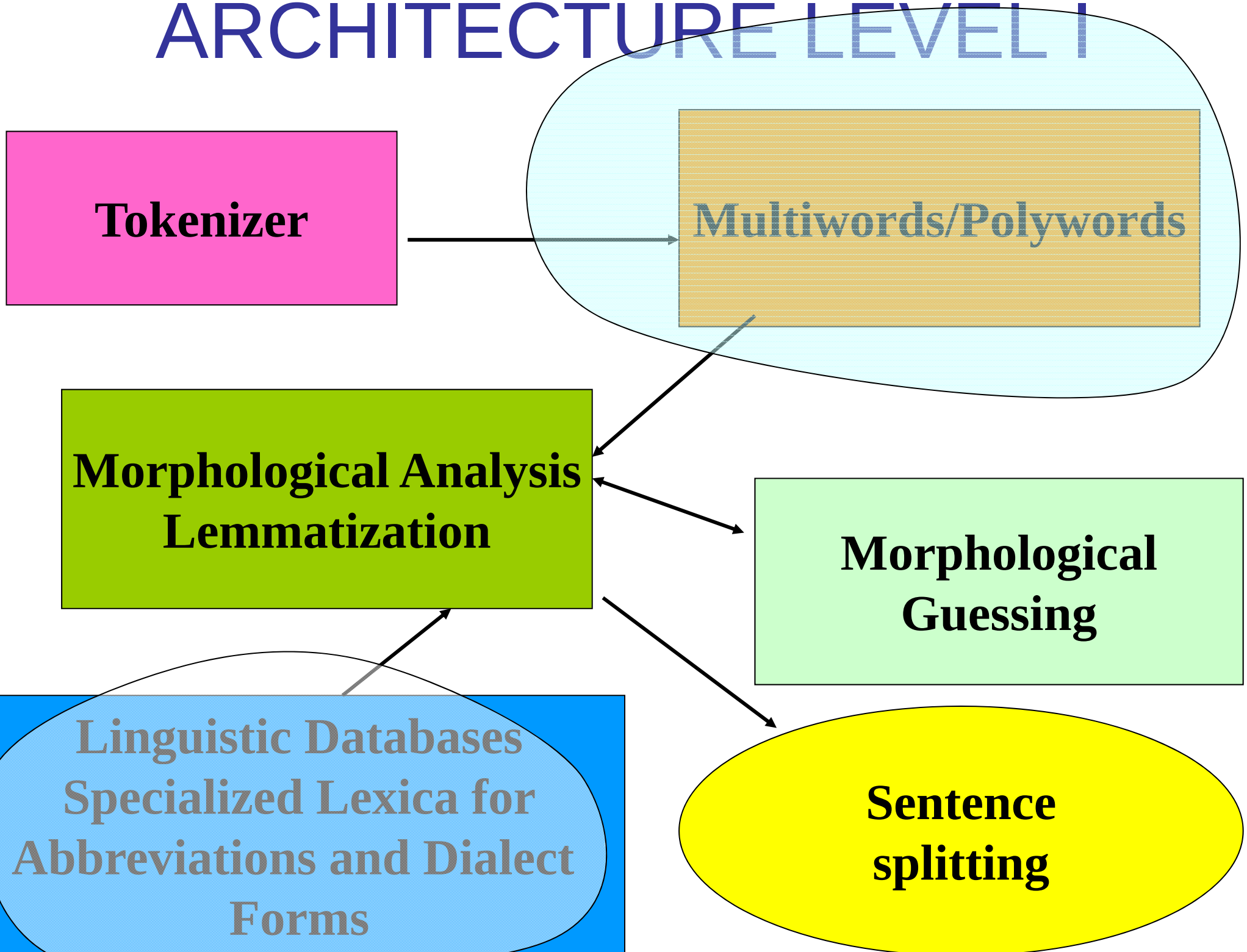
```
( (S
 (NP-PD
        (ART-NK Das)
        (ADJA-NK einzige)
        (NN-NK Forum)
        (PP-MNR
                (APPR-AC für)
                (PDAT-NK diese)
                (NN-NK Musik)
        ))
 (VAFIN-HD ist)
 (NP-SB
        (ART-NK das)
        (ADJA-NK interessierte)
        (NN-NK Publikum)
        (PP-MNR
                (APPR-AC bei)
                (CNP-NK
                        (NN-CJ Konzerten)
                        (KON-CD und)
                        (NN-CJ Festivals)
                )))) ($. .) )
```

```
( (S                (S-MO
                        (VMFIN-HD Mögen)
                        (NP-SB
                                (NN-NK Puristen)
                                (NP-GR
                                        (PIDAT-NK aller)
                                        (NN-NK Musikbereiche) ))
                        (ADV-MO auch)
                        (VP-OC
                                (NP-OA (ART-NK die)
                                        (NN-NK Nase) )
                                (VVINF-HD rümpfen) )) ($, ,)
                (NP-SB (ART-NK die)
                        (NN-NK Zukunft)
                        (NP-GR (ART-NK der)
                                (NN-NK Musik) ))
                (VVFIN-HD liegt)
                (PP-MO (APPR-AC für)
                        (PIDAT-NK viele)
                        (ADJA-NK junge)
                        (NN-NK Komponisten) )
                (PP-MO
                        (APPRART-AC im)
                        (NN-NK Crossover-Stil)
                )) ($. .) )
```

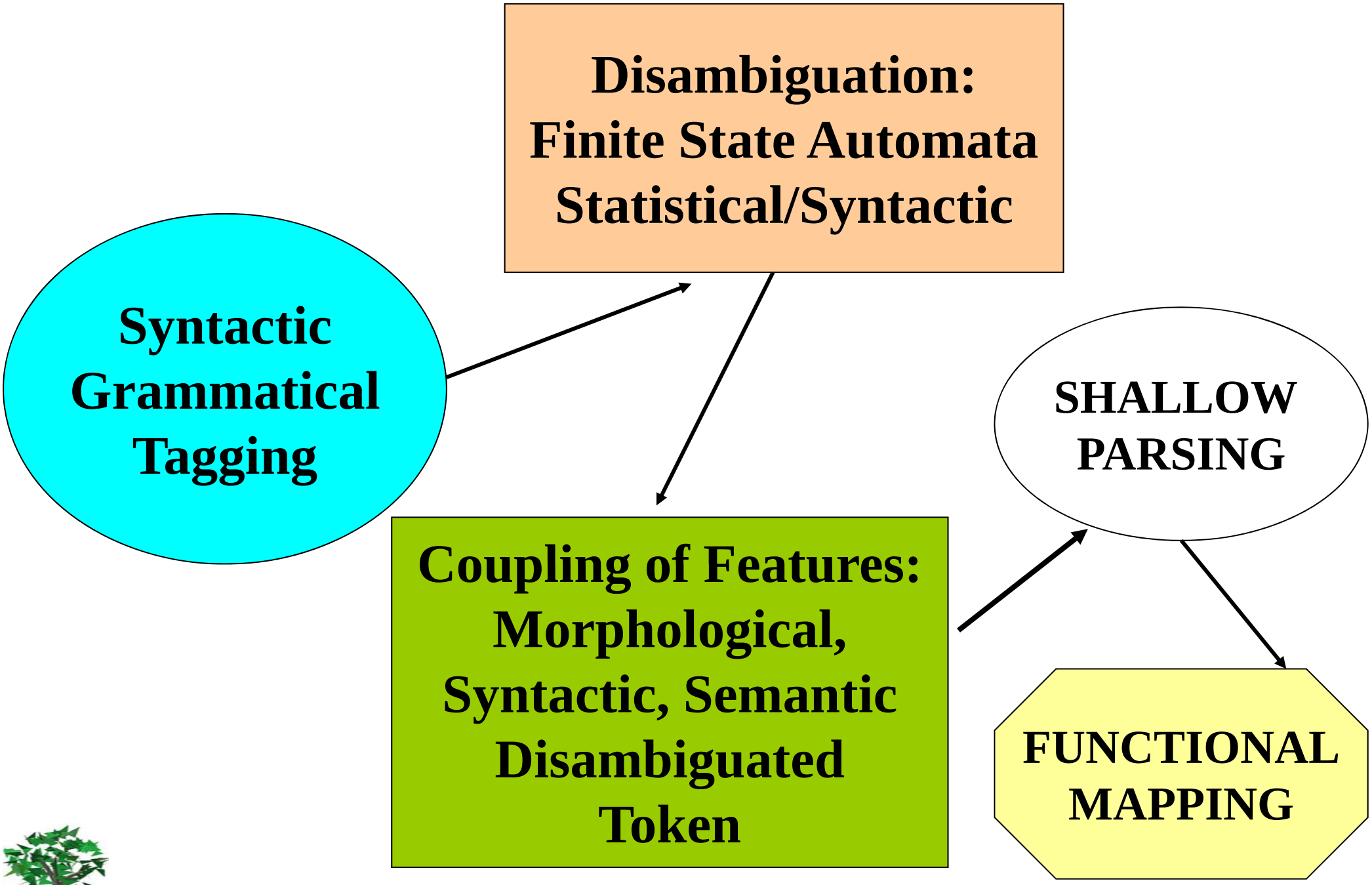
ARCHITECTURE LEVEL I



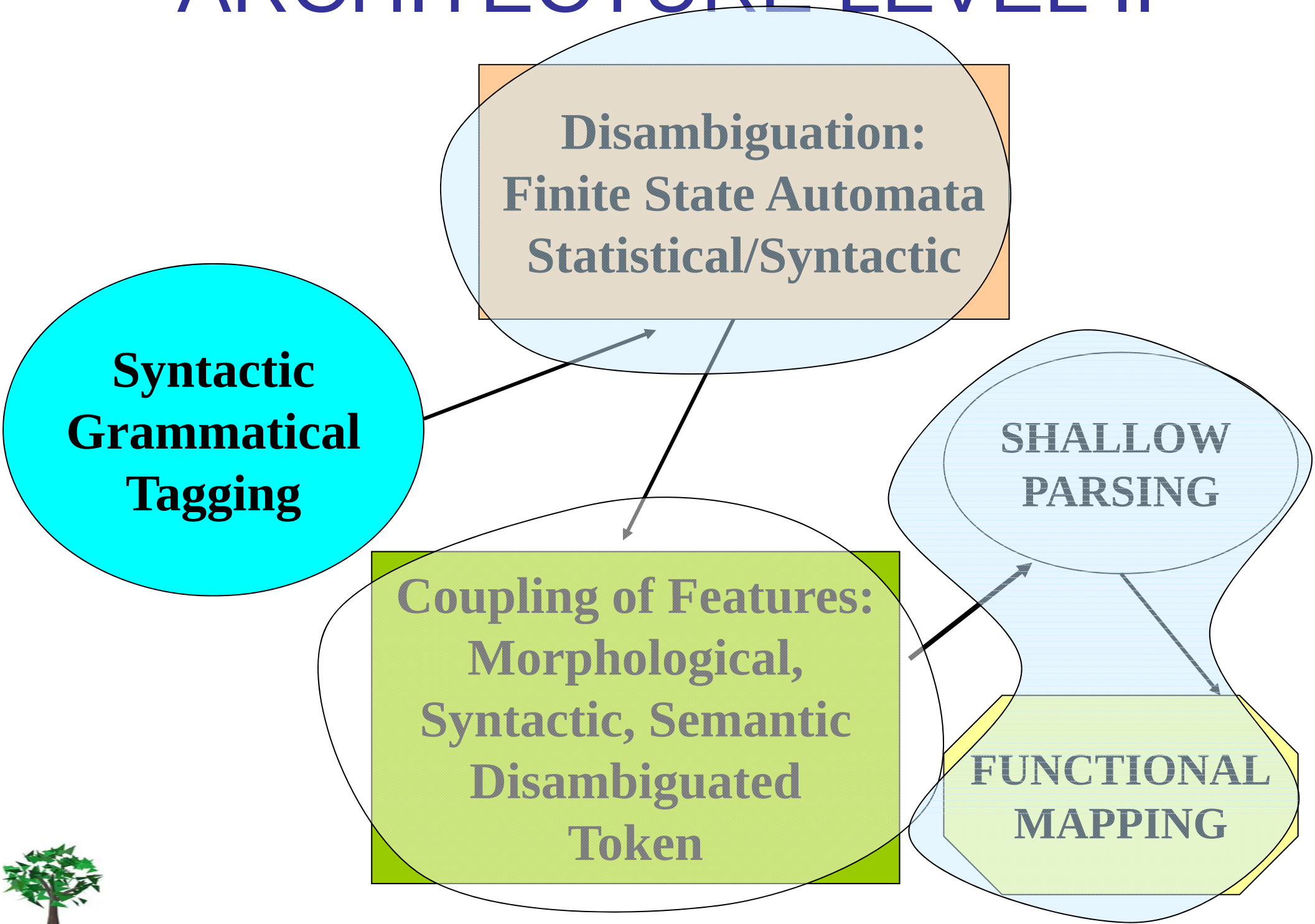
ARCHITECTURE LEVEL I



ARCHITECTURE LEVEL II



ARCHITECTURE LEVEL II



Quantitative Data for Written Text Treebanks

- 10,200 Utterance
- 257,797 Tokens
- 274.000 Tokens +Arts
- 229,067 Constituents
- 69580 NPs
- 41985 PPs
- 21205 APs
- 15930 IBARs

- **20592 PPs DI/DA**
- **7565 Untensed Clauses**
- **1724 QPs**
- **4575 ADVPs**
- **3425 RELCls(F2)**
- **3552 Fragments(F3)**
- **959 COMPLCls(FAC)**

Quantitative Data for Aps and F/IBAR

- **21205 APs**
- **1227 possessives**
- **6880 PostNominal APs**
- **2321 PreNominal APs**
- **940 internal PreNominal APs**
- **577 Complement APs**

- **2526 IR_INFL Irrealis**
- **8603 F/IBAR little_pros - 1156 IR_INFL**
- **23621 Tensed/Untensed Cl**
- **4906 CPs**
- **1063 SUBORDCls(FS)**
- **585 InterCls(FINT)**

DATA for F/FINT Structures

	F-[IBAR ***F-[IR_INFL*** ***Empty Subject***	***F-[SN*** ***Lexic alized Subj.***	*F3*	*FINT*	***CP_INT*** ***FINT - excluded***	***No. Utter.*** ***F + F3***
AVIP/API	***4179***	***622***	*2038*	*620*	*836*	*6849*
DIALOGUE ***DIFFER-ENCES***	*231*	*36*	*104*	*81*	*28*	*371*
CORPUS *WRITTEN*	***9257***	***6636***	*3206*	*557*	*204*	*19099*

TYPOLOGIES of Adjectival STRUCTURES

- **Coordination**
- **Pronominalization**
- **Internal Dependency**
- **Complements**
- **Quantified and Comparative Structures**
- **Spec of SA**

TYPOLOGIES of Non-Canonical STRUCTURES

- **Inverse Focalized**
- **Reported Speech**
- **Fragments governed by SA**
- **Left Dislocazions**
  - **Within CP**
  - **Within Complement**

NON-CANONICAL STRUCTURES IN VIT

Treebanks Vs. Non-canonical Structures	Non-canonical Structures (TU)	Structures with Non-Canonical Subject (TS	Total (T Utteran	Total (TS) Simple Sentenc	Totale Comple Sentenc
VIT	3719	9800	10,200	19,099	6782
Percenta	27.43%	51.31%	63.75%		66.5%
PT	7234	2587	5,600	9,532	3,600
Percenta	13.01%	02.7%	59.44%		69.4%

NON-CANONICAL STRUCTURES IN VIT

Treebank Vs. Non canonical Structure	LDC (left dislocated complex)	S_DISL (dislocated subject)	S_OPV (topicalized subject)	SLOC (Fronted Subject)	Total Non-Canonical	Total Complete Sentences
VIT	25	107	265	26	379	3000

Conversion to Dependency

- Manually relabeling of all non-canonical structures
- Introduction of a subcategorization lexicon of 17,000 Italian verbs
- We used agreement for Subj and position
- For remaining constituents only position to assign Argument/Adjunct grammatical labels
- Adjuncts with spatiotemporal locations detected from head tag

Conversion to Dependency: Preliminary Evaluation

- The treebank has 10,607 constituents with subject role, 3,423 of which have been manually assigned because they are in non-canonical position. Among the 7,184 SUBJ labels which were automatically identified, 46 constituents should have been assigned another function, with a precision of 0.99. On the other hand, 218 constituents should bear a SUBJ label instead of their actual label, with a recall of 0.97

High Structural Ambiguity

- It is easy to guess that the constituents with a higher structural ambiguity in Italian are those whose position in respect to the head is less predictable: respectively AP>VP>PP>RC>PP-BY>PP-OF. Two criteria apply when looking for ambiguity measures:
- Semantic function of constituent : Argument vs. Adjunct or Modifier
- Attachment Position : Head Adjacent vs. Non-Head Adjacent

High Structural Discontinuity

- Other elements that can lead to discontinuity or non-canonicity problems are:
- the number of F3 or sentence fragments is quite high compared to the number of total utterances, 3552 (35%);
- the number of complex utterances is quite high – 6782 if compared to the total number (10.200) of utterances, therefore much higher than the 41% of PT.

Modifier Discontinuity Table

Constituent/ Distance	SP	SPD	SPDA	SV	F2	SA	TOTAL
Head Adjacent (HA)	4726	13.798	509	3249	1560	13.932	37,774
Distance=1	2677	1827	266	941	460	908	7,079
Distance=2	1718	494	203	485	305	179	3,384
Distance=3	624	81	58	130	82	24	999
Distance=4	600	45	32	175	100	23	975
Total All Mods (AM)	10.345	16.245	1068	4980	2507	15.066	50,211
Ratio AM/AC	***0,483***	***0.912***	***0.384***	***0.658***	***0.73***	***0.71***	***0.652***
Totals Non HA	5619	2447	559	1731	947	1134	12,437
Ratio Non HA/AM	***0,54***	***0,15***	***0,523***	***0,347***	***0,378***	***0,075***	***0.652***
All Constituents	21.393	17.812	2780	7568	3425	21.205	76,971

Adjectival APs & their Functions

- A count of the functional conversion of adjectival structures is presented here below:

  - **1296 Complement APs (ACOMP)**
  - **18748 Modifiers (MOD)**
  - **324 Adjuncts (ADJ)**
  - **2001 COORDinate APs**

AMBIGUOUS PREDICATIVE SA

Adjectives may be positioned in front or after the noun they modify almost freely for most classes

sn-[art-i, n-posti, spd-[partd-della, sn-[n-dotazione, sa-[ag-organica_aggiuntiva]]], sa-[ag-disponibili, sp-[p-a,

the posts of the pool organic additive available to

AMBIGUOUS PREDICATIVE SA

Syntactic ambiguity arise and needs agreement to be checked

sn-[sa-[ag-significativi], n-ritardi]], sn-[sa-[ag-profonde], n-trasformazioni], ibar-[vt-investono],

significative delays profound transformations affect

AMBIGUOUS PREDICATIVE SA

Syntactic ambiguity arise and needs agreement to be checked also in a row

sn-[art-il, n-totale, spd-[partd-dei, sn-[n-posti, spd-[partd-della, sn-[n-dotazione, sa-[ag-organica]]], ag-vacanti], sa-[ag-disponibili

the total of the posts of the pool
organic additive vacant available

AMBIGUOUS PREDICATIVE SA

Syntactic ambiguity arise and needs agreement to be checked also in a row, however the adjective may belong to a following noun phrase

ibar-[vin-darebbe], compin-[sp-[in-anche, part-agli, sn-[n-orientamenti, spd-[pd-di, sn-[n-democrazia, sa-[ag-laica]]]]], sn-[sa-[ag-maggiori

would give also to the viewpoints of democracy laic main

MINOR PHENOMENA

- COORDINATION
- DEPENDENCY
- PRONOMINALIZATION
- COMPLEMENTATION
- SPEC SA | Neg, Adv, Int, Quant…

SENTENCE COMPLEMENT

f-[sn-[art-il,
sa-[ag-bello]],
ibar-[vc-è],
compc-[fac-[pk-che]

the beatiful is that

SENTENCE COMPLEMENT

f-[sn-[art-l_,
sa-[ag-importante]],
savv-[avv-ora],
ibar-[vc-è],
compc-[sv2-[vcl-aprirlo,
compt-[clitac-lo],
savv-[pd-di, avv-più]]]]

the important now is to open it of more

TOUGH PROBLEMS: QUANTIFICATION

sq-[in-molto, q-più, coord-[sa-[ag-efficace, punt-,, ag-controllabile, cong-e, ag-democratico]],
sc-[ccom-di,
f2-[sq-[relq-quanto],
cp-[savv-[avv-oggi],
f-[ibar-[neg-non, vcir-sia]

much more effective , controllable and democratic of how much today not be

TOUGH PROBLEMS: QUANTIFICATION

cp-[-sq-[in-Più, sa-[ag-buono],
sc-[ccom-di, savv-[avv-così]]],
f-[ibar-[neg-non, vsupp-poteva, vci-essere

more good than so not could be

TOUGH PROBLEMS: QUANTIFICATION

cp-[sc-[ccom-tanto, sq-[q-più],
f-[ibar-[vc-sono], compc-[sa-[ag-lunghi]]],
sc-[ccom-tanto, sq-[q-maggiore],
f-[ibar-[vc-è],
compc-[sn-[art-la, n-soddisfazione, sa-[ag-finale]

much more are long much higher is the satisfaction final

TOUGH PROBLEMS: QUANTIFICATION

```
cp-[
    cp-[sa-[ag-generali],
        sp-[p-per, f2-[relq-quanto,
            f-[ir_infl-[vcir-siano]]]]], punt-,,
                f-[sn-[art-le, n-regole], ibar-[vt-investono
```

general for as much as be the rules involve

FRONTED SPs in PARTICIPIALS

sp-[p-in, sn-[n-base,
sp-[part-al, sn-[n-punteggio,
sv3-[sp-[p-ad, sn-[pron-essi]],
ppas-attribuito, compin-[sp-[p-con,

on the basis of the scoring to them attributed with

FRONTED SPs in PARTICIPIALS

sp-[p-a,
coord-[sn-[sa-[ag-singoli], n-plessi],
cong-o,
sn-[n-distretti],
sv3-[sp-[p-in, sn-[pron-essi]],
ppas-compresi, punto-.]]]]]]]]]

to single groups or districts in them comprised

FRONTED SPs in PARTICIPIALS

spd-[partd-degli,
sn-[n-importi,
sv3-[sp-[p-ad, sn-[pron-essi]],
ppre-spettanti]]], cong-e,

of the amounts to them owed and

FRONTED SPs in PARTICIPIALS

spd-[partd-della,
sn-[n-cortesia,
sv3-[sp-[p-in,
sq-[q-più, pd-di, sn-[art-un_, n-occasione]]],
vppt-dimostrata,
compin-[coord-[sp-[p-a, sn-[pron-me]],

of the courtesy in more than one occasion demonstrated to me

SUBJECT INVERSION

f-[ibar-[vc-diventa],
compc-[savv-[avv-così],
sa-[in-più, ag-acuta],
sn-[art-la, n-contraddizione], sp-[p-tra

becomes so more acute the contradiction between

SUBJECT INVERSION

f-ibar-[vc-è],
compc-[sa-[ag-peculiare,
sp-[part-all, sn-[np-Italia**]]],**
sn-[art-l, n-esistenza, spd-[pd-di

is peculiar to Italy the existence of

FRAGMENTS & SENTENCES WITH FOCUS INVERTED SA

cp-[s_foc-[ag-Buono],
f3-[sn-[cong-anche, art-l, n-andamento,
spd-[partd-delle, sn-[n-vendite

good also the behaviour of the sales

FRAGMENTS & SENTENCES WITH FOCUS INVERTED SA

cp-[s_foc-[ag-Calmo],
f3-[sn-[art-il, n-listino,
spd-[partd-del, sn-[n-granoturco]

quite the price list of mais

FRAGMENTS & SENTENCES WITH FOCUS INVERTED SA

cp-[s_foc-[ag-buono], congf-invece,
savv-[p-nel, avvl-complesso],
f3-[sn-[art-il, n-resto

good instead on the whole the rest

FRAGMENTS & SENTENCES WITH FOCUS INVERTED SA

cp-[ldc-[sa-[ag-altra], n-fonte,
spd-[pd-di, sn-[n-finanziamento]]],
f-[ibar-[vc-sarà],
compc-[sn-[art-il, n-trattamento

other source of funding will be the treatment

HANGING TOPICS & LEFT DISLOCATION

cp-[sn-[sa-[ag-brutta], n-faccenda], punt-,,
f-[sn-[art-i, n-sudditi],
ibar-[clit-si, vt-ribellano, punto-.]]

bad story , the populace self rebel

HANGING TOPICS & LEFT DISLOCATION

cp-[ldc-[art-una, n-decisione, sa-[ag-importante]],
f-[sn-[nh-Ghitti],
ibar-[clitac-l, ausa-ha, vppt-riservata],

a decision important Ghitti it has reserved

HANGING TOPICS & LEFT DISLOCATION

cp_int-[ldc-[art-il, n-concorso],
f-[ibar-[clitac-l, ausa-ha, vppt-vinto],
compt-[coord-[sn-[nh-Francesco],
cong-o,
sn-[nh-Giovanni]]]],
puntint-?]

the competition it has won Francesco or Giovanni ?

AUX-TO-COMP STRUCTURES

```
cp-[f-[sn-[art-La, n-perdita],
        sp-[p-per, sn-[art-il, npro-Rolo]],
        ibar-[vcir-sarebbe],
        compc-[congf-però,
                spd-[pd-di, sn-[in-circa, num-'30', num-miliardi]]]],
        topf-[auxtoc-[auag-avendo],
                f-[sn-[art-la, npro-Holding],
                        sv3-[vppt-incassato,
                        compt-[sn-[n-indennizzi,
                                sp-[p-per, sn-[num-'28',
                                        num-miliardi]]]]]], punto-.]
```

the loss for the Rolo would be then of about 30 billion having the Holding cashed payments for 28 billions

AUX-TO-COMP STRUCTURES

fc-[congf-e, punt-',',
topf-[auxtoc-[clit-si, aueir-fosse],
f-[sn-[pron-egli],
sv3-[vppin-trasferito, cong-pure,
compin-[sp-[part-nel,
sn-[sa-[in-più, ag-remoto], n-continente]]]]]]

and , self would be he moved also in the more remote continent , SENTENCE

AUX-TO-COMP STRUCTURES

cp-[sn-[topf-[auxtoc-[art-l, ausai-avere],
f-[sn-[art-il, n-figlio],
sv3-[vppt-abbandonato,
compt-[sn-[art-il, n-mare],
sp-[p-per, sn-[art-la, n-città]]]]]]],
f-[ibar-[clitdat-le, ause-era, avv-sempre, vppt-sembrato]

the have the son abandoned the sea for the city her was always seemed

(IN)DIRECT REPORTED SPEECH:

- A. parenthetical inserted between SUBJ and IBAR
- B. parenthetical inserted between material in CP and the F
- C. free reported direct speech and then quoted direct speech
- D. Direct speech is ascribed to an anonymous "someone" quoted anyhow

(IN)DIRECT REPORTED SPEECH: A. parenthetical inserted between SUBJ and IBAR

dirsp-[par-",
cp-[sp-[p-a, sn-[sa-[dim-questo], n-punto]],
f-[sn-[art-la, n-data], par-",
fp-[punt-,, f-[ibar-[ausa-ha, vppt-detto],
compt-[sn-[npro-d_, npro-Alema],
savv-[avv-ieri], nt-sera]]], punt-,],
par-", ibar-[vin-dipende],

" at this point the date " , said D'Alema last night , "
depends

(IN)DIRECT REPORTED SPEECH: A. parenthetical inserted between SUBJ and IBAR

dirsp-[par-",
cp-[sp-[p-in, sn-[sa-[dim-questo], n-libro]],
f-[sn-[nh-madre, npro-Teresa],
fp-[par--, f-[ibar-[vt-spiegano],
compt-[sp-[part-alla,
sn-[npro-Mondadori]]]], par--],
ir_infl-[vcir-darà],

in this book Mother Theresa -- explain at the Mondadori - will give

RESIDUAL PROBLEMS: RELATIVES AND COMPLEMENT CLAUSES AS MAIN SENTENCES

cp-[f2-[rel-Che,
cp-[fp-[punt-,, f-[ibar-[vt-sostengono],
compt-[sp-[part-alla, sn-[npro-Farnesina]]]], punt-,],
f-[ibar-[neg-non, ausa-ha,
sp-[p-per, avvl-niente],
vppt-gradito],
compt-[sn-[art-l, n-operazione, n-by_pass]],
punto-.]]]]

That , maintain at the Farnesina , not has in no case liked the operation by_pass .

RESIDUAL PROBLEMS: RELATIVES AND COMPLEMENT CLAUSES AS MAIN SENTENCES

cp-[f2-[rel-che,
f-[ibar-[neg-non, vc-è],
compc-[sn-[n-cifra, spda-[pda-da, sn-[in-poco]]]]]],
fp-[punt-,, fc-[ccong-così, conjl-come,
f-[ibar-[neg-non, vc-è],
compc-[sn-[n-cosa, spda-[pda-da, sn-[qc-tutti, art-i, nt-giorni]]],
sv2-[vci-avere, compc-[sn-[num-un, n-erede,
sp-[part-al, sn-[n-trono,
sc-[ccom-come, n-guida,
sa-[ag-turistica], punto-.]]]]]]]]]

That not is figure by nothing , so as not is thing by every day to have one heir to the throne as guide turistic.

RESIDUAL PROBLEMS: RELATIVES AND COMPLEMENT CLAUSES AS MAIN SENTENCES

cp-[fac-[pk-che, savv-[avv-poi],
f-[sn-[art-la, n-legge],
ibar-[neg-non, virin-riesca],
compin-[sv2-[pt-a, viin-funzionare]]]],
punt-,, f-[ibar-[vc-è],
compc-[sn-[art-un, n-discorso, f2-[rel-che

That then the law not manages to work , is a matter that

BIKEL'S Model Implemented on a subset (homogeneous!!)

Number of sentence	=	3109
Number of Error sentence	=	0
Number of Skip sentence	=	0
Number of Valid sentence	=	3109
Bracketing Recall	=	67.47
Bracketing Precision	=	66.48
Complete match	=	6.66
Average crossing	=	4.17
No crossing	=	30.33
2 or less crossing	=	53.43
Tagging accuracy	=	97.26

This work has been carried out by Alberto Lavelli

BIKEL'S Model Implemented on a subset (homogeneous!!)

-- len<=40 --		
Number of sentence	=	2458
Number of Error sentence	=	0
Number of Skip sentence	=	0
Number of Valid sentence	=	2458
Bracketing Recall	=	71.16
Bracketing Precision	=	70.08
Complete match	=	8.42
Average crossing	=	2.40
No crossing	=	38.00
2 or less crossing	=	65.13
Tagging accuracy	=	97.20

This work has been carried out by Alberto Lavelli

BIKEL'S Model Implemented on the whole of VIT

Number of sentence	=	10189
Number of Error sentence	=	12
Number of Skip sentence	=	0
Number of Valid sentence	=	10177
Bracketing Recall	=	68.61
Bracketing Precision	=	68.29
Complete match	=	8.70
Average crossing	=	3.25
No crossing	=	38.37
2 or less crossing	=	61.73
Tagging accuracy	=	96.65

This work has been carried out by Alberto Lavelli

BIKEL'S Model Implemented on the whole of VIT

-- len<=40 --

Number of sentence	=	8519
Number of Error sentence	=	12
Number of Skip sentence	=	0
Number of Valid sentence	=	8507
Bracketing Recall	=	71.87
Bracketing Precision	=	71.58
Complete match	=	10.40
Average crossing	=	1.94
No crossing	=	45.47
2 or less crossing	=	71.72
Tagging accuracy	=	96.55

This work has been carried out by Alberto Lavelli

CONCLUSIONS

- We have shown with sufficient evidence and clarity that machine learning and statistical methods FAIL whenever sufficient conditions of homogeneousness do not obtain. Related issues concern language typology as well as strict adherence to linguistic theories.

CONCLUSIONS

- Eventually, if the main goal is the attainment of syntactic-semantic transparency and thus facilitate as much as possible the conversion of the treebank into a semantically “complete” representation, syntactic/dependency structure should be at least semantically coherent from the start. That’s what we did in VIT.