

VIT

Venice Italian

Treebank

Rodolfo Delmonte
Dipartimento Scienze Linguaggio
Laboratory Computational Linguistics
Universit "Ca Foscari"
30124 - VENEZIA
Tel. 39-041-2345717/52
E-mail: delmont@unive.it
Website: <http://project.cgm.unive.it>

Progetti Nazionali di Treebanks dell'italiano

• Progetto SITAL

annotazione di un corpus di italiano scritto ai seguenti livelli

- ❖ sintattico
 - ❖ struttura a costituenti (~90.000 parole)
 - ❖ struttura funzionale (~300.000 parole)
- ❖ semantico-lessicale (~80.000 parole piene, distribuite tra nomi, verbi e aggettivi)

• Progetto AVIP/IPAR

annotazione di un corpus di italiano regionale ai seguenti livelli

- ❖ sintattico
 - ❖ struttura a costituenti e *funzionale* (~60.000 parole)

Progetti Locali di Treebanks dell'italiano

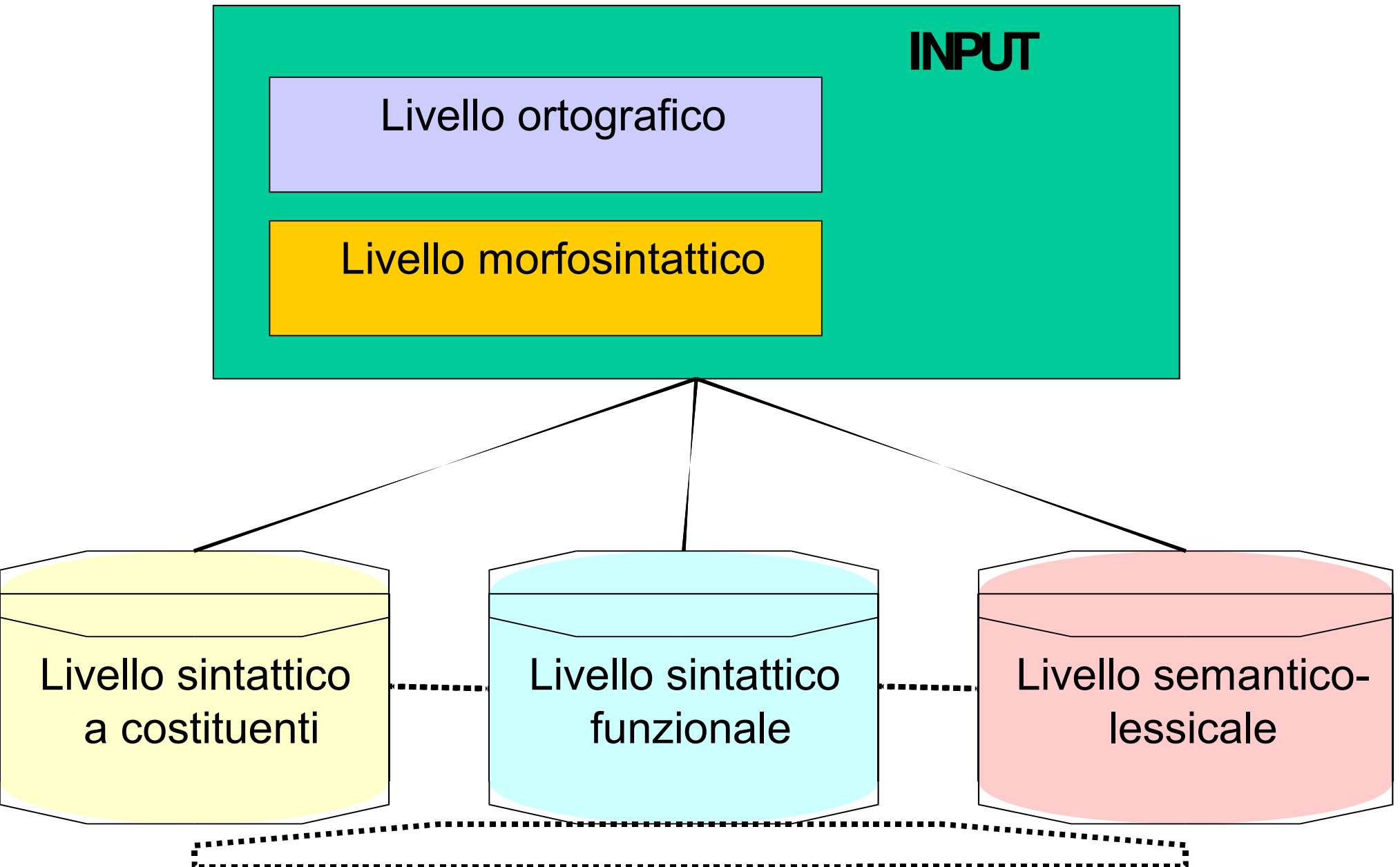
- **Progetto DIGITAL EQ. 1986-88**
- **annotazione manuale di un corpus di italiano scritto a livello**
 - ❖ **sintattico**
 - ❖ **struttura a costituenti (~100.000 parole)**
- **Progetti interni**
- **annotazione automatica di un corpus di italiano scritto a livello**
 - ❖ **sintattico**
 - ❖ **struttura a costituenti (~170.000 parole)**

USI POSSIBILI DI UN TREEBANK

- nell'ambito di applicazioni di elaborazione del linguaggio naturale (IR, WSD, acquisizione di conoscenza linguistica)
- per l'addestramento e la specializzazione (training/tuning) di grammatiche e sistemi di disambiguazione di senso
- per la valutazione di sistemi di analisi sintattica e disambiguazione semantica
- per lo studio delle caratteristiche sintattiche e semantiche della lingua

Specifiche di annotazione: architettura della Treebank

Moduli di annotazione



Treebank sintattico-semantica dell'italiano

- ❖ annotato ai seguenti livelli:
 - ❖ ortografico, con indicazione dell'organizzazione macrotestuale
 - ❖ morfo-sintattico, con indicazione del lemma e di espressioni polilessicali di base (es. *ad hoc*, *allo scoperto*, *al di là*)
 - ❖ annotazione verificata manualmente
 - ❖ schema di annotazione conforme agli standards esistenti (EAGLES) e comune al tema "Dialoghi Annotati"
- ❖ formato di rappresentazione: XML corredato di DTD appropriate

Criteri ispiratori struttura in costituenti: X-barra

- **Schema teorico regole X-barra**
 - CP --> Spec, Cbarra
 - Cbarra --> C0, XP
 - XP --> Spec, Xbarra
 - Xbarra --> X, Complementi
 - C0 --> Complementatore
 - X --> Verbo, Aggettivo, Nome, Avverbio
- **Specificatore Atomico vs Strutturato**
 - Spec--> Determinanti, Quantificatori, Intensifiers
- **Schema adattato regole X-barra**
 - XP --> spec(atomico), Xbarra
 - Xbarra --> X, Complementi
 - X --> Nome, Aggettivo, Avverbio

Criteria ispiratori struttura in costituenti: X-barra

- *Schema adattato regole X-barra*

CP --> Spec, Cbarra

Cbarra --> C0, XP

C0 --> Complementatore

XP --> Spec, Xbarra, Complementi, Aggiunti

Spec --> Soggetto

Complementi --> COMPT/COMPIN/COMPC/COMPPAS

Xbarra --> Gruppo Verbale

- *Struttura Gruppo Verbale*

Xbarra --> Verbo - ausiliari, modali, clitici, negazione, avverbiali, sintagmi preposizionali, congiunzioni

TYOLOGY OF SYNTACTIC CONSTITUENTS

Costituenti Sostanziali

SN	sintagma nominale
SA	sintagma aggettivale
SQ	sintagma quantificato
SAVV	sintagma avverbiale
IBAR	nucleo verbale a tempo finito
IR_INFL	nucleo verbale a tempo finito irrealis
SV2	frase infinitiva
SV3	frase participiale
SV5	frase gerundiva

TYOLOGY OF SYNTACTIC CONSTITUENTS

Costituenti Funzionali Lessicali

Simbolo	Tipo di costituente
FAC	frase complemento con o senza complementatore
FC	frase coordinata/frase comparativa
FS	Subordinatore frase subordinata
FINT	Elementi +wh frase interrogativa, anche se il pronome interrogativo è preceduto da preposizione
DIRSP	Frase con discorso diretto/diretto riportato
FP	Introduttore punteggiatura frase parentetica o apposizione
F2	Frase relativa, anche se il pronome relativo è preceduto da preposizione o da articolo
COORD/costituente	Elemento coordinante e costituente coordinato
SP	sintagma preposizionale
SPD	sintagma preposizionale DI
SPDA	sintagma preposizionale DA

TYPOLOGY OF SYNTACTIC CONSTITUENTS

Costituenti Funzionali Strutturali

Simbolo	Tipo di costituente
F	Frase + soggetto o IBAR
F3	Frase frammento
TOPF	Strutture ad Aux-to-comp
CP	Elementi dislocati o anteposti, aggiunti frasali e non
CP_INT	Elementi dislocati o anteposti, aggiunti frasali e non
COMPT	Complementi retti da Verbi Transitivi
COMPIN	Complementi retti da Verbi Intransitivi
COMPPAS	Complementi retti da Verbi Passivi
COMPC	Complementi retti da Verbi Copulativi

Specifiche per l'annotazione sintattica a costituenti

- **identificazione dei costituenti sintagmatici e loro relazioni di incassamento gerarchico**
- **assegnazione della categoria sintattica ai costituenti individuati**
- **criteri di annotazione di ampia copertura, soprattutto per quanto riguarda l'annotazione di costruzioni sintattiche complesse**

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 f r)

(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 M gen)

(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))

(VFIN-HD r mpfen))) (\$, ,)

(NP-SB

(ART-NK die)

(NN-NK Zukunft)

(NP-GR (ART-NK der)

(NN-NK Musik)))

(VFIN-HD liegt)

(PP-MO

(APPR-AC f r)

(PIDAT-NK viele)

(ADJA-NK junge)

(NN-NK Komponisten))

(PP-MO

(APPRART-AC im)

(NN-NK Crossover-Stil)

)) (\$, .))

Functional Annotation: Main Relations

- dip(head,dependent)
- sogg(head,dependent)
- comp(head,dependent)
- mod(head,dependent)
- arg(head,dependent)
- pred(head,dependent)
- non-pred(head,dependent)
- ogg_d(head,dependent)
- ogg_i(head,dependent)
- obl(head,dependent)

Functional Annotation: Features of Dependent

- intro = introdep/introsim
- caso
- status = open/closed
- mood
- (semantic) role
(agent/temporal/locative/compar)

Functional Annotation: Features of Head

- diat = middle/active/passive/reflexive
- syn_form = pers/impers/si_impers
- reflex = passive/ipron/rifl/rifl_app
- pers
- num
- gen

Functional Annotation: Features of Head/Dependent

- **quant**
- **card**
- **def**

- **aux**
- **perifra**

- **cong**

Caratteri precipui parlato

Parlato e scritto

Trascrizione ortografica e ortografica

Architettura del sistema di annotazione

Le sovrapposizioni

Parlato e Scritto

la trascrizione

orto(ideo)grafica

forma linguistica - parole della lingua e dialettali;

quasi linguistica - quasi parole e interiezioni di vario tipo;

non linguistica - non parole, pause, e altri fenomeni di disfluenza.

Parlato e Scritto

p1#94: no <sp> cioè sì c'ha<aa> <mh> <sp> una specie di tappo

p1#96 <lp> c'ha prima una base un po' altina

p1_94: no, cioè sì c'ha mh, una specie di tappo.

p1_96: - c'ha prima una base un po' altina.

Corpus AVIP

diamo i numeri...

- tokens totali = 56337 di cui:
- punteggiatura e marcatori di turno = 18710 tokens
- parole, interiezioni, quasi parole ecc. = 37627 tokens

ARCHITETTURA LIVELLO I

Tokenizzatore

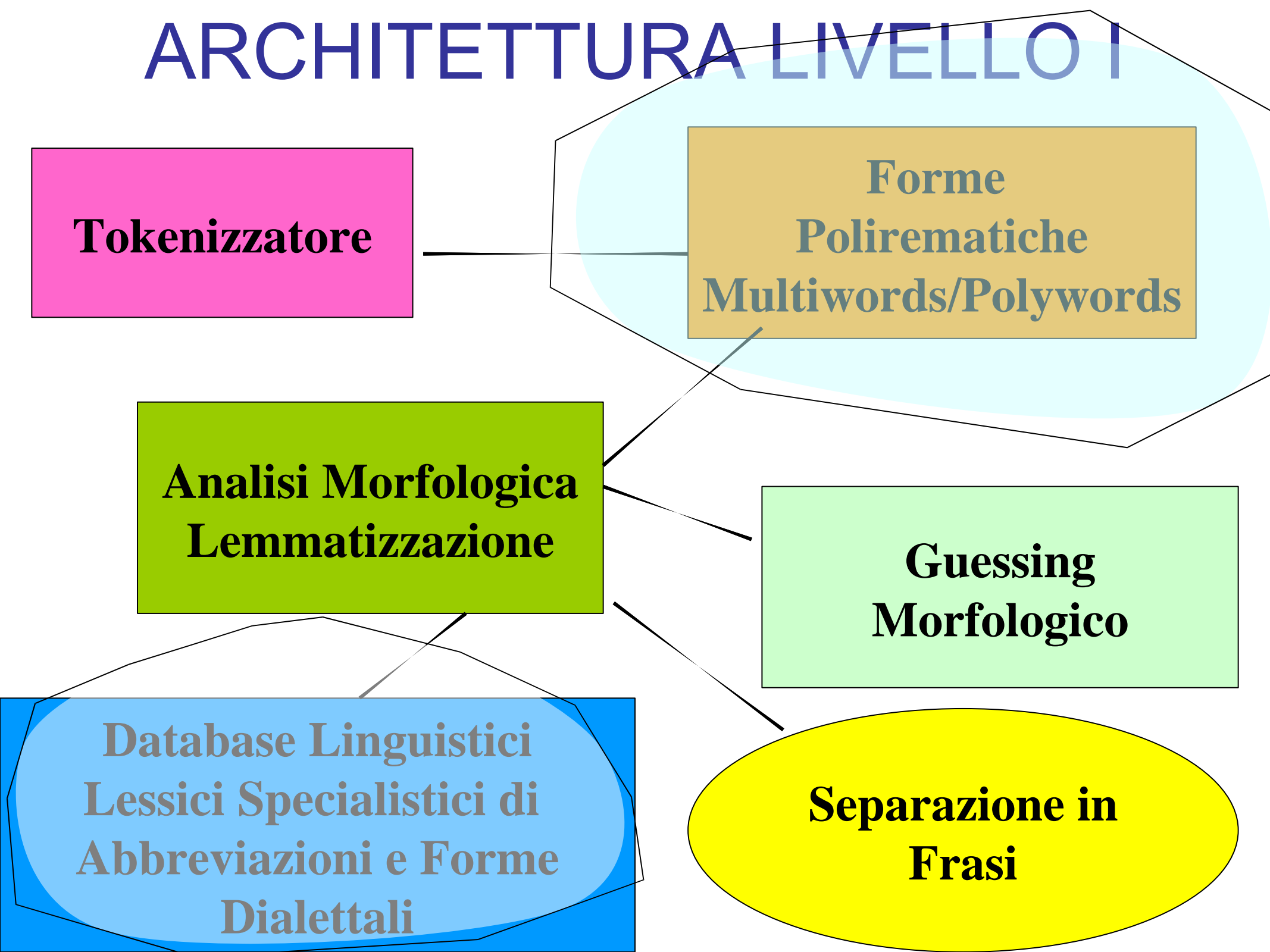
**Forme
Polirematiche
Multiwords/Polywords**

**Analisi Morfologica
Lemmatizzazione**

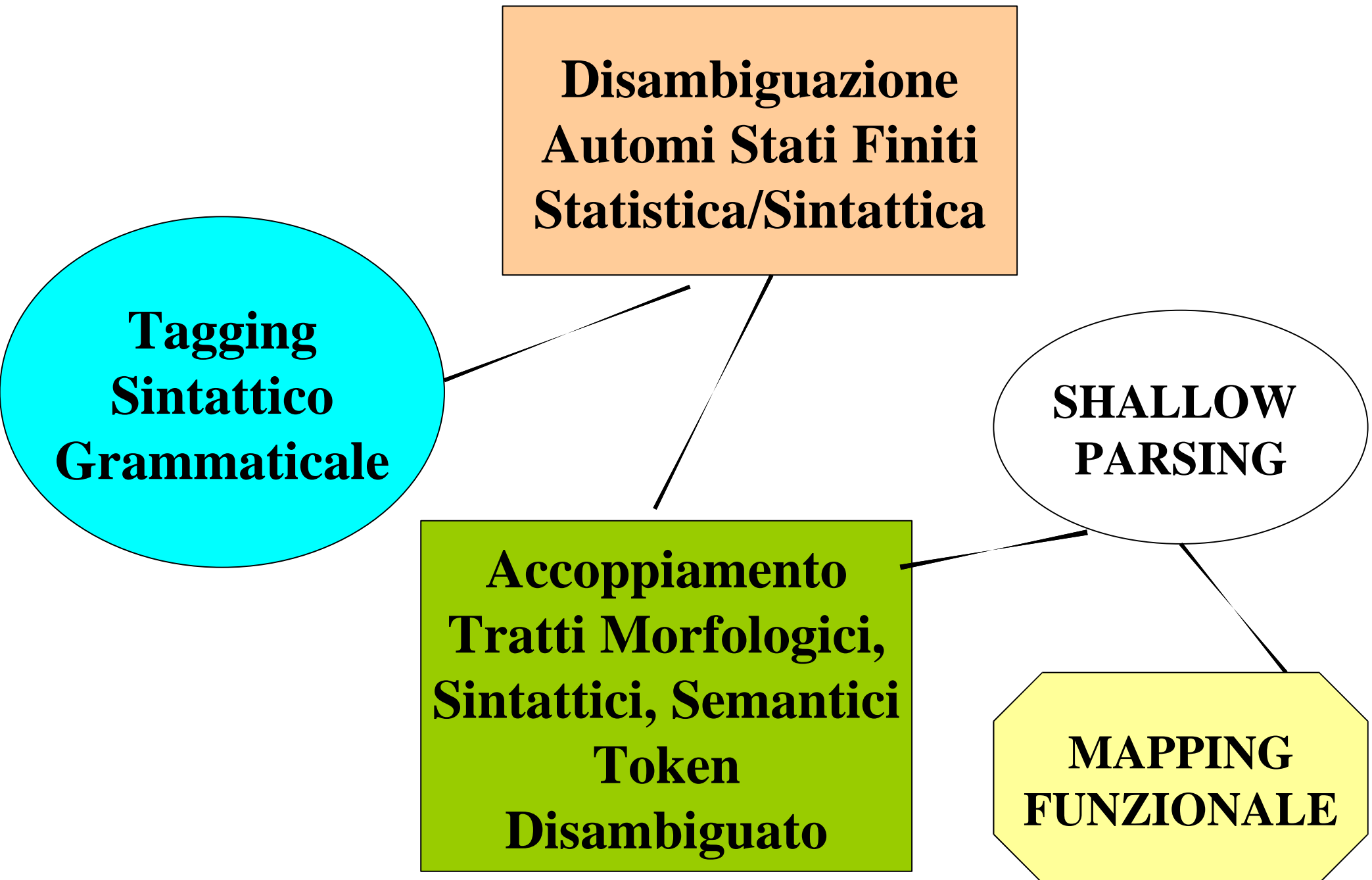
**Guessing
Morfologico**

**Database Linguistici
Lessici Specialistici di
Abbreviazioni e Forme
Dialettali**

**Separazione in
Frase**



ARCHITETTURA LIVELLO II



ARCHITETTURA LIVELLO II

**Tagging
Sintattico
Grammaticale**

**Disambiguazione
Automati Stati Finiti
Statistica/Sintattica**

**Accoppiamento
Tratti Morfologici,
Sintattici, Semantici
Token
Disambiguato**

**SHALLOW
PARSING**

**MAPPING
FUNZIONALE**



Quantitative Data for Written Text Treebanks

- 10,200 Utterance
 - 257,797 Tokens
 - 229,067 Constituents
 - 69863 NPs
 - 42541 PPs
 - 21706 APs
 - 15958 IBARs
- 21014 PPs DI/DA
 - 7663 Untensed Clauses
 - 1724 QPs
 - 4234 ADVPs
 - 3439 RELClS(F2)
 - 3206 Fragments(F3)
 - 959 COMPLClS(FAC)

Quantitative Data for Aps and F/IBAR

- 17372 Adjectives
- 21706 APs
- 1227 possessives
- 6880 PostNominal APs
- 2321 PreNominal APs
- 940 internal PreNominal APs
- 577 Complement APs
- 2286 IR_INFL Irrealis
- 8603 F/IBAR little_pros -
1196 IR_INFL
- 23621 Tensed/Untensed Cl
- 4581 CPs
- 1026 SUBORDClS(FS)
- 536 InterClS(FINT)

DATI QUANTITATIVI

	<i>Punteggiatura</i>	<i>Tokens</i>	<i>Tokens Totali</i>	<i>Sovrapposizioni</i>	<i>No. Turni</i>	<i>No. Enunciati F + F3</i>
<i>AVIP/API</i>	<i>18710</i>	<i>37627</i>	<i>56337</i>	<i>1110</i>	<i>4747</i>	<i>6849</i>
<i>DIALOGO DIFFERENZE</i>	<i>1637</i>	<i>2645</i>	<i>4282</i>	<i>147</i>	<i>336</i>	<i>371</i>
<i>CORPUS SCRITTO</i>	<i>32214</i>	<i>225440</i>	<i>257654</i>	<i>--</i>	<i>--</i>	<i>19099</i>

RAPPORTI

	<i>Rapporto Sovr/Turn</i>	<i>Rapporto Sovr/Token</i>	<i>Rapporto Sovr/Enunc</i>	<i>Rapporto Token/Enun.</i>
<i>AVIP/API</i>	<i>1 x 4,27</i>	<i>1 x 31,19</i>	<i>1 x 6,15</i>	<i>5,5 tok x En.</i>
<i>DIALOGO DIFFERENZE</i>	<i>1 x 2,28</i>	<i>1 x 17,99</i>	<i>1 x 2,52</i>	<i>7,14 tok x En.</i>
<i>CORPUS SCRITTO</i>	<i>--</i>	<i>--</i>	<i>--</i>	<i>11,8 tok x En.</i>

DATI STRUTTURE F/FINT

	<i>F-[IBAR</i> <i>F-[IR_INFL</i> <i>Sogg.vuoto</i>	<i>F-[SN</i> <i>Sogg.</i> <i>pieno</i>	<i>F3</i>	<i>FINT</i>	<i>CP_INT</i> <i>Esclusi</i> <i>FINT</i>	<i>No.</i> <i>Enunciati</i> <i>F + F3</i>
<i>AVIP/API</i>	<i>4179</i>	<i>622</i>	<i>2038</i>	<i>620</i>	<i>836</i>	<i>6849</i>
<i>DIALOGO</i> <i>DIFFERENZE</i>	<i>231</i>	<i>36</i>	<i>104</i>	<i>81</i>	<i>28</i>	<i>371</i>
<i>CORPUS</i> <i>SCRITTO</i>	<i>9257</i>	<i>6636</i>	<i>3206</i>	<i>557</i>	<i>204</i>	<i>19099</i>

DATI % STRUTTURE F/FINT

	<i>F-[IBAR F- [IR_INFL Sogg.vuoto</i>	<i>F-[SN Sogg. pieno</i>	<i>F3</i>	<i>FINT CPINT</i>	<i>No. Enuncia ti F + F3</i>
<i>AVIP/API</i>	<i>61,01%</i>	<i>0,91%</i>	<i>29,75 %</i>	<i>21,25 %</i>	<i>6849</i>
<i>DIALOGO DIFFERENZE</i>	<i>62,2%</i>	<i>0,97%</i>	<i>28,03 %</i>	<i>29,38 %</i>	<i>371</i>
<i>CORPUS SCRITTO</i>	<i>48,47%</i>	<i>34,74 %</i>	<i>13,79 %</i>	<i>0,04%</i>	<i>19099</i>

TIPOLOGIE di STRUTTURE Aggettivali

- **Coordinazione**
- **Pronominalizzazione**
- **Dipendenza interna**
- **Complementi**
- **Strutture Quantificate e Comparative**
- **Spec di SA**

TIPOLOGIE di STRUTTURE Aggettivali

- **Focalizzate Inverse**
- **Frammenti retti da SA**
- **Dislocazioni a Sinistra**
 - **Nel CP**
 - **Nel Complemento**