

## RISORSE ITALIANE

Metodi, problemi, comparazioni

## Lo stato dell'arte

2

### Bilanciamento

- Scritto / parlato

### Dimensioni

- 100 milioni di occorrenze o più

### Annotazione / lemmatizzazione

- POS tagging
- lemmatizzazione

### Accesso al corpus

- Internet-based
- Gratuito
- Accesso al testo integrale E mediante interrogazione

## Principali corpora italiani

3

ITALIANO SCRITTO	ITALIANO PARLATO
LIF - Lessico di frequenza della lingua italiana contemporanea	LIP - Lessico di frequenza dell'italiano parlato
CORIS / CODIS Corpus Dinamico dell'Italiano scritto	CLIPS - Corpora Linguistici per l'Italiano Parlato e Scritto
COLFIS - Corpus e Lessico di Frequenza dell'Italiano Scritto	LABLITA - Corpus di italiano parlato
LA REPUBBLICA CORPUS (giornalistico)	<i>Integrated reference corpora for spoken romance languages (C-ORAL-ROM)</i>

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Altri corpora di italiano

4

ITALIANO SCRITTO	ITALIANO PARLATO
TLIO - Tesoro della lingua italiana delle origini (lettarario)	CIT - Corpus di italiano televisivo
LIZ - Letteratura Italiana Zanichelli (lettarario)	LIR - Lessico di frequenza dell'italiano radiofonico
BOnonia Legal Corpus (BoLC)	API/AVIP/IPar
EUROTRA, EuroWordNet, PAROLE, SIMPLE l'Italian Reference Corpus	Child Language Data Exchange System (CHILDES) - italiano
Banca dati di Italiano L2 Osservatorio Linguistico permanente dell'Italiano Diffuso fra Stranieri -Siena	

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

5

## Corpora della lingua scritta

LIF, COLFIS, CORIS/CODIS, LA REPUBBLICA

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Il LIF - *Lessico di frequenza della lingua italiana contemporanea*

6

### Realizzazione

- CNUCE (Centro Nazionale Universitario di Calcolo elettronico) di Pisa
- (1971)
- U. Bortolini, C. Tagliavini, A. Zampolli
- primo grande progetto di costruzione di un lessico di frequenza per la lingua italiana (non tagliato su un singolo autore o su testi specificatamente letterari).

### Reference corpus

- Corpus di 500.000 parole
- **Testi scritti**
  - **1947-1968**
- 15.750 lemmi ordinati per frequenza e secondo l'ordine alfabetico

Oggi il corpus **non** è disponibile.

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Tipologie testuali: rappresentatività dello scritto (e indirettamente del parlato)

7



Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## COLFIS - *Corpus e Lessico di Frequenza dell'Italiano Scritto Contemporaneo*

8

Pier Marco Bertinetto, Cristina Burani, Alessandro Laudanna, Lucia Marconi,

Daniela Ratti, Claudia Rolando, e Anna Maria Thornton

Pubblicato: 2006 (testi anni Novanta)

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

# Progetto COLFIS

9

## Rappresentatività italiano "medio"

- Letture preferite dagli italiani (indagini ISTAT)
- 1992-1994

## Bilanciamento

- differenziati per tipologia (quotidiani, periodici, libri) e per argomento (politica, letteratura, sport, ecc.).
- il bilanciamento delle fonti, che conferisce un carattere di non casualità alle rilevazioni numeriche estraibili dall'archivio lessicale

## Estensione

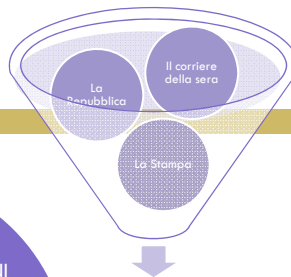
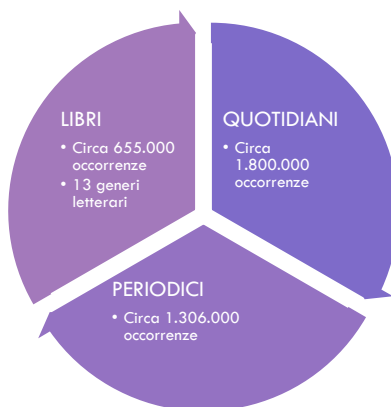
- 3.798.275 parole

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

# Bilanciamento

10

- altro
- arte scienza e tecnica
- auto e nautica
- bambini e ragazzi
- casa e hobby
- femminili
- fotoromanzi
- informazione generale
- cronaca mondiale
- radio e televisione
- sport
- viaggi e ecologia



## QUOTIDIANI

- economia
- cronaca locale
- cronaca mondiale
- cronaca nera
- politica estera
- politica interna
- scienza
- spettacolo
- sport

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Ricerca nel corpus

11

### CORPUS NON LEMMATIZZATO

Interrogazione   Descrizione   Download   Legenda

Testo da cercare

Settori

- quotidiani  
 periodici  
 libri

[Seleziona tutti](#) · [Deseleziona tutti](#)

Cerca

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
 Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## “sai” nel COLFIS

periodico	fotoromanzi	lancio-Lucky	Rienzi Alice	Immagini di una ragazza scomparsa		94-09-13	Janet è sempre piena di risorse, lo <a href="#">sai</a> .
periodico	fotoromanzi	sogno	Mancuso Antonino	Sta suonando per me		92-12-01	Si sta divertendo, e poi lo <a href="#">sai</a> che ci tiene alla tua presenza.  Eppure, <a href="#">sai</a> .
periodico	informazione general	epoca	Gnocchi Laura	Vi ricordate...Serena Cruz? ...Adesso è una bambina felice		92-09-16	Del resto, della sua vita passata ricorda poco, e soltanto ogni tanto dice: "Mamma <a href="#">sai</a> ."
periodico	informazione general	espresso	Siciliano Enzo	Il film: che profumo di Cechov!		92-09-13	Non <a href="#">sai</a> se abbia letto più Cechov o Trifonov.

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
 Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## COLFIS in sintesi

13

### Pregi

#### Rappresentatività

- Criterio esterno fondato
- Comparabilità cronologica con LIP

#### Estensione

- Media per gli standard attuali

#### Distribuzione

- Online
- Liste di frequenza scaricabili in molti formati

### Difetti

#### Interrogazione

- Maschera molto povera
- Non supporta ricerche per categoria

#### Accesso corpus

- Solo alla porzione autorizzata
- Senza esportazione delle concordanze
- Nessun accesso al testo integrale

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

14

## CORIS / CODIS

### *CO*rpus *D*inamico dell'*I*taliano *S*critto

<http://corpus.cilta.unibo.it:8080/>

CILTA, Bologna

Diretto da R. Rossini Favretti

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## II CORIS/CODIS

15

### Due forme

- *Corpus di riferimento dell'italiano scritto (CORIS)*
- *COorpus dinamico dell'italiano scritto (CODIS)*

### 100 milioni di parole

- aggiornato tramite un corpus di monitoraggio con cadenza biennale
- testi: prevalentemente di narrativa prodotta negli anni Ottanta e Novanta

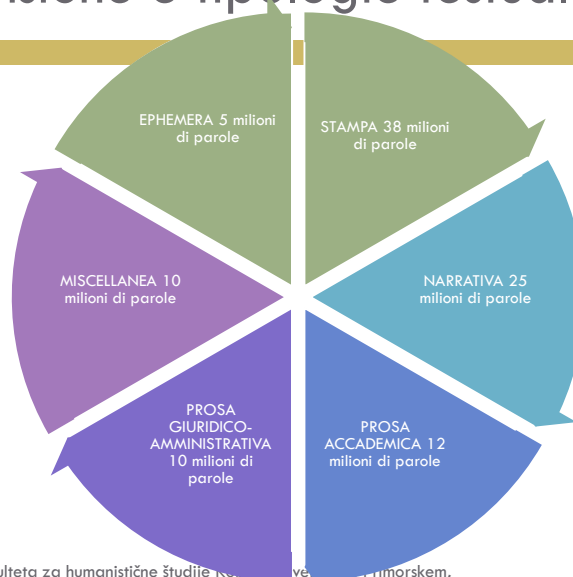
### Accesso online (mediante registrazione)

- [http://corpus.cilta.unibo.it:8080/coris\\_ita.html](http://corpus.cilta.unibo.it:8080/coris_ita.html)

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Estensione e tipologie testuali

16



Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari



## Interrogazione CODIS

17

### CODIS - Corpus query form

		<b>Query</b>			
		<input type="text" value="dare + un + esame"/>	<a href="#">Query Language Help.</a>		
		<input checked="" type="checkbox"/> Case insensitive search			
Subcorpus	Size (in Mw)				
STAMPA	<input type="checkbox"/> 20	<input type="checkbox"/> 10	<input type="checkbox"/> 5	<input checked="" type="checkbox"/> 3	
NARRATIVA	<input type="checkbox"/> 13	<input type="checkbox"/> 7	<input type="checkbox"/> 3	<input type="checkbox"/> 2	
PROSA ACCADEMICA	<input type="checkbox"/> 5	<input type="checkbox"/> 4	<input type="checkbox"/> 2	<input type="checkbox"/> 1	
PROSA GIURIDICO-AMM.	<input type="checkbox"/> 4	<input type="checkbox"/> 3	<input type="checkbox"/> 2	<input type="checkbox"/> 1	
MISCELLANEA	<input type="checkbox"/> 4	<input type="checkbox"/> 3	<input type="checkbox"/> 2	<input type="checkbox"/> 1	
EPHEMERA	<input type="checkbox"/> 2	<input type="checkbox"/> 1	<input type="checkbox"/> 1	<input type="checkbox"/> 1	

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Risultati in concordanza

18

Number of concordances: 300/309

STAMPA\_3 : e giorni dal fattaccio , il caso è praticamente chiuso . Un particolare : prima  
 STAMPA\_3 : tando alla versione ufficiale , ha praticamente consegnato ai carabinieri il Cap  
 STAMPA\_3 : ste dalla riforma Amato del ' 93 ( praticamente con i vecchi requisiti ) , ma ac  
 STAMPA\_3 : tutte quelle riduzioni di pena che praticamente vanificano l ' effetto dissuasiv  
 STAMPA\_3 : stato il giorno buono . La Roma ha praticamente chiuso l ' acquisto di Helguera  
 STAMPA\_3 : icenne nel giardino condominiale , praticamente sotto gli occhi di alcuni compag  
 STAMPA\_3 : " George " ? O quello che Teddy fa praticamente da sempre ? " Anche se l ' osse  
 STAMPA\_3 : " : in questa categoria includeva praticamente tutte le categorie sociali , dai  
 STAMPA\_3 : della fabbrica " . Enzo Ferrari ha praticamente creato la pista di Imola : un am  
 STAMPA\_3 : ace , amici che mi vogliono bene e praticamente tutto ciò che sognavo . Ma quest  
 STAMPA\_3 : i sale cinematografiche a ingresso praticamente gratuito sono comunque ancora mo  
 STAMPA\_3 : connessioni ) facciano il mercato praticamente senza consultarli ( il caso - Ta  
 STAMPA\_3 : nis , costretto dalla crisi a fare praticamente il custode , fa fatica a tirare  
 STAMPA\_3 : tornare indietro " dalla scelta , praticamente esclusiva , dell ' Assemblea Cos  
 STAMPA\_3 : a laboratorio che girerà a Fiorano praticamente per tutto il mese a partire da m  
 STAMPA\_3 : omenica l ' abbiamo a disposizione praticamente tutti . Ma possiede potenzialità  
 STAMPA\_3 : % rispetto a settembre ) . Prezzi praticamente fermi invece a Bari e Palermo (   
 STAMPA\_3 : rtamente . Del resto l ' Africa ha praticamente sovvenzionato lo sviluppo dell '   
 STAMPA\_3 : ungo telex in cui una nobildonna , praticamente , intima allo Stato di darle sub  
 STAMPA\_3 : isticato sistema di trasmissione , praticamente esclusivo , che consente non sol  
 STAMPA\_3 : o la squadra più forte del mondo , praticamente la stessa della passata stagione  
 STAMPA\_3 : esta bambina , una famiglia l ' ha praticamente abbandonata e l ' altra se l ' è  
 STAMPA\_3 : ntitativi venduti nel ' 96 si sono praticamente dimezzati . In realtà - spiega G

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## CORIS/CODIS in sintesi

19

### Pregi

#### Accesso dinamico

- Dimensioni personalizzabili dall'utente
- Sintassi di interrogazione abbastanza elementare

#### Estensione

- Standard

#### Distribuzione

- Online (ma parziale)
- Nessuna risorsa associata

### Difetti

#### Interrogazione

- Maschera poco flessibile
- Non è lemmatizzato
- Limite di concordanza (300)

#### Accesso corpus

- Senza esportazione delle concordanze
- Nessun accesso al testo integrale

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Corpus La Repubblica

20

### SSLMIT (University of Bologna)

- Corpus annotato, lemmatizzato

### Estensione

- Circa 380 milioni di parole

### Accesso

- Gratuito online (mediante registrazione)
- <http://dev.sslmit.unibo.it/corpora/corpus.php>



Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Tipologie testuali

21

### La Repubblica

news

commento

church, culture,  
economics,education,  
news, politics,science, society,  
sport, weather

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Risultati in concordanza

22

su . E sempre innocuo nella possibilità di raccontare la **storia** **attuale** , soprattutto per quanto attiene ai rapporti se  
 edia dei sentimenti per scavare in chiave ironica su **una** **storia** **autobiografica** , lui aveva 34 anni . Io ne ho 35 ades  
 ù possibile all' autore del romanzo , Hunter Thompson , **storia** **autobiografica** , lui aveva 34 anni . Io ne ho 35 ades  
 ui stesso si racconta . Nel 1971 , quando scrisse **questa** **storia** **avventurosa** di Grazia diventa quasi un pretesto per  
 bile Bernardino la famiglia de ' Rossi come gli altri ebrei **storia** **avventurosa** di Grazia diventa quasi un pretesto per  
 tova è stata costretta a fuggire abbandonando tutto , **la** **storia** **avvolta** nel silenzio di un personaggio che soltanto in  
 alla voglia di silenzio . Il percorso di Filippini narratore è **storia** **avvolta** nel silenzio di un personaggio che soltanto in  
 ria esemplare di autonegazione ironica e passionale , **la** **storia** **avvolta** nel silenzio di un personaggio che soltanto in  
 familista " I gay italiani stroncano Kubrick I GAY italiani **storia** **banalotta** e " familista " " dice Franco Grillini , dell' Ai  
 bocciano " Eyes Wide Shut " . " **Una** **storia** **avvolta** nel silenzio di un personaggio che soltanto in  
 skowitz , che aveva trovato una buona distribuzione , **la** **storia** **bella** e inconsueta , che raccontava il lungo percorso  
 Universal , e , finalmente , un mercato . Con **questa** **storia** **bella** e inconsueta , che raccontava il lungo percorso  
 libri , ma non fu un intellettuale , li scrisse per la gente **storia** **biblica** , la storia d' Italia , sempre per il popolo . Il b  
 semplice . Narrò la storia ecclesiastica , narrò **la** **storia** **biblica** , la storia d' Italia , sempre per il popolo . Il b  
 Don Bosco aveva il buon senso e la  
 le proporzioni di un Watergate della City . Nomi famosi **storia** **britannica** mentre sulle poltrone dorate delle banche  
 :ravolti nella polvere di uno dei più grandi imbrogli **della** **storia** **britannica** mentre sulle poltrone dorate delle banche  
 della City vi sono uomini di gran potere che tremano

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## LA REPUBBLICA in sintesi

23

### Pregi

#### Trattamento

- Lemmatizzato e analizzato morfologicamente (in modo automatico, Treetagger)

#### Interrogazione

- Sintassi di interrogazione molto ricca (un po' complessa)

#### Estensione

- 380 milioni (grande per gli standard attuali)

#### Distribuzione

- Online gratuita

### Difetti

#### Disegno

- Non è un corpus di riferimento

#### Accesso corpus

- Senza esportazione delle concordanze
- Nessun accesso al testo integrale
- Liste di frequenza non esportabili nella totalità

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

24

## Corpora della lingua parlata

LIP, CLIPS, C-ORAL-ROM

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## II LIP

25

### *Lessico di frequenza dell'italiano parlato*

A cura di Tullio De Mauro, Federico Mancini, Massimo Vedovelli e Miriam Voghera  
(1993)

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Il lessico di frequenza del LIP

26

A		0	0	3	1	3	7	4	abbraccio	0	10	0	0	0	10	0
a	E	3093							ABBACCIO	0	2	0	0	13	15	2
									S	3678						
									abbraccio	0	2	0	0	11	13	2
									abbraccione	0	0	0	0	2	2	0
									ABETE	0	0	5	0	0	5	0
									Cg	6505						
									abete	0	0	5	0	0	5	0
									ABILE	4493						
									Ag	0	2	2	0	0	4	2
									abile	0	0	1	0	0	1	0
									abili	0	0	1	0	0	1	0
									abilissimo	0	2	0	0	0	2	0
									ABILITA'	2516						
									S	5	3	1	0	1	10	6
									abilita'	5	3	1	0	1	10	6
									ABILITARE	6505						
									V	0	4	0	0	0	4	0
									abilitarmi	0	1	0	0	0	1	0
									abilitata	0	1	0	0	0	1	0
									abiliti	0	2	0	0	0	2	0
									ABILITATO	6505						
									S	0	4	0	0	0	4	0
									abilitati	0	4	0	0	0	4	0
									ABILITAZIONE	0	7	0	1	0	8	1

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## LIP – «Corpus del Lessico di frequenza dell'italiano parlato»

27

A cura di Tullio De Mauro, Federico Mancini, Massimo Vedovelli e Miriam Voghera (1993)

57h di registrazione di parlato (1990-1992)

- 475.883 parole grafiche
- 496.335 occorrenze di lemmi

**Rappresentatività geografica:** Milano, Firenze, Roma e Napoli: ogni città 125.000 occorrenze

Interrogazione completa e gratuita

- sito BADIP (banca dati dell'italiano parlato)
- [http://languageserver.uni-graz.at/badip/badip/20\\_corpusLip.php](http://languageserver.uni-graz.at/badip/badip/20_corpusLip.php)

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Rappresentatività

28

Tipologie  
testuali

- 1) scambio *bidirezionale faccia a faccia* con presa di parola **libera**
- 2) scambio *bidirezionale non faccia a faccia* con presa di parola **libera** (conversazioni telefoniche)
- 3) scambio *bidirezionale faccia a faccia* con presa di parola **non libera** (dibattiti, interviste, interrogazioni)
- 4) scambio *unidirezionale in presenza di destinatario/i* (lezioni, conferenze, omelie, comizi, ecc.)
- 5) scambio *unidirezionale o bidirezionale a distanza* (trasmissioni radiofoniche e televisive)

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Esempio: il testo RA1 (formato grezzo)

29

```
A: chi e'? chi e'? fatti vedere?
B: come stai?
A: bene_ [BACI]
A: ciao
B: chi e'? chi e'? e' zia Vania <??> ahah che magro
A: come che magro? mangia come una bestia rara
B:          e' magrissimo          <?> che bella che sei
ammazza sei
A:          ma <?> #
A:          una ciofecca
B:          una show girl
A:          una show gir<l> RIDE
B: <?> io sono un mostro allora?
A: e allora?
B: il brodo stai facendo?
A: ti sto facendo il risotto ai funghi e gli asparagi eh eh poi_ #
non so? vuoi qualcos'altro?
B: no basta grazie sai ho mangiato tutto il giorno a voglia
A:          sul serio un suppli'_?          hai mang<iato>
vomitato tutto il giorno?
```

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Interrogazione BADIP

30

**badip**

banca dati dell'italiano parlato

[home](#) [corpus](#) [lip](#) [collaboratori](#) [consulenti](#) [contatto](#) [sponsor](#) [lista di corpora](#) [english](#)

ricerca

testi

tipologia dei testi

classi di parola

simboli e notazioni

durata delle registrazioni

parlanti

lista dei lemmi



Cerca tutte le sequenze che contengano: ( aiuto: )

.V.comprare	digita la prima parola oppure <a href="#">clicca</a>
%	digita la seconda parola oppure <a href="#">clicca</a>
.S	digita la terza parola oppure <a href="#">clicca</a>

e che non contengano:

	digita la prima parola
	digita la seconda parola
	digita la terza parola

nelle città:  Firenze  Milano  Napoli  Romanei tipi di testo:  A  B  C  D  E

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Risultati

31

cerca

occorrenze del lemma/forma trovato: **24**  
 sul totale di parole grafiche: **146462**

statistiche:

- 1) .V.comprare
- 2) %
- 3) .S

salva i dati visualizzati:

cit	tip	co	en	pa	enunciato
ta'	o	n	un	ria	
		ver	cia	e	
		sa	nt		
		zio	to		
		ne			
F	A	2	346	C	senti scusa se t' interrompo gli ho <b>comprato una crema</b> per il corpo secondo me domani avra' l' avra' finita
F	D	14	1	A	lecco in questa terza domenica di quaresima e anche nelle due

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
 Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Osservare le concordanze

32

F	E	1	1	A	forza i carciofi duemila # quattro duemila lire il carciofino # * # <b>comprate il arancino</b> bono dumila lire guarda # un chilo e mezzo dumila spinaci e finocchio un chilo e mezzo duemila lire
R	A	2	38	B	lo devo comprare me lo devo <b>comprare grande idea</b>
R	A	2	158	B	si trova nelle erboristerie \$ solo nelle erboristerie costa relativamente tanto perche' un \$ da mezzo chilo costa seimila lire invece Anna ne ha <b>comprato in quantita</b> industriale \$ \$ \$ per far \$ i suoi amici e questa persona che da' queste diete e tremila lire al chilo lo ha pagato
R	A	5	1	A	della serie ve lo cuccate cosi' come dio ve l' ha mandato * e c' e' Letizia che che dovrebbe andar via per adesso sono dieci giorni e' a Parigi perche' s' e' <b>comprata una casa</b> # e allora ci dovrebbe lasciare Brigitte come insegnante e poi rimaniamo io e Monica o spesso io solamente con lei e questo Claudio
R	E	4	185	H	buongiorno senta eh io so' una signora nuovo che mio zio ha <b>comprato due apparecchi</b> due apparecchi
R	E	5	45	B	dice senti io sono pisano ah qui siamo quasi tutti pisani guarda non mangiare non <b>comprare di cavallo</b> ah *

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
 Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari



## Usare i testi integrali

33

**badip**  
banca dati dell'italiano parlato

home corpus lip collaboratori consulenti contatto sponsor lista di corpora english

ricerca  
testi  
tipologia dei testi  
classi di parole  
simboli e notazioni  
durata delle registrazioni  
parlanti  
lista dei lemmi

Testo:

http://language-server.uni-graz.at/~badip - file FA1 - versione badip - Mozilla Firefox

File Modifica Visualizza Vg Segnalibri Strumenti 2

```
A: come sei fine
B: eh e tutto quanto il capitolo del $ # si' pero' e' un capitolo che non mette
la fine
C: perche' qui si sposta * *
D: che c' e' *
C: c' e' la carne la straniera alla pizzaiola
B: cioe' a te Giovanna non ti piace l' aglio vero *
C: ue ne sono novantasette grammi a testa
B: aspetta che mi se mi dai $ li' raccolgo l' aglio
C: ma l' aglio lo vogliamo anche noi
B: no la Giovanna ha detto che non gli piace
D: niente lo vogliamo anche noi
B: la Giovanna ha detto che 'un gli piace
D: non ti piace l' aglio *
A: mi piace mi piace * *
B: quanti siamo *
D: dieci
C: piu' quattro *
A: ahah se stata soddisfatta di questo studio
D: di quale * * la storia
A: * storia
D: ahah perso perche' quando uno esercita l' intelligenza eh certo che e'
Completato
```

## LIP in sintesi

34

### Pregi

#### Interrogazione

- Ricca anche per categorie vuote
- Accesso ai sottocorpora
- Esportabilità di tutti i risultati in molteplici formati (XML, Excel, testo, ecc.)

#### Distribuzione

- Online gratuita
- Accesso al corpus integrale (scaricabile)
- Accesso con interrogazione

### Difetti

#### Estensione

- Il corpus è piccolo (secondo gli standard attuali)

#### Accesso corpus

- Nella versione online non si accede alle liste di frequenza
- La versione online non è identica a quella su carta
- Non è distribuito l'audio

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

35 **CLIPS**

**Corpora e Lessici dell'Italiano Parlato e Scritto**  
**Sezione parlato**

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
 Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

**CLIPS**

36

**Progetto**

- diretto da Federico Albano Leoni
- 1999-2004
- voci maschili e voci femminili, in parte trascritto ortograficamente e etichettato foneticamente

**Struttura**

- **100 ore di parlato**
- Distribuzione sia dell'audio sia delle trascrizioni

**Località**

- 15 località italiane
- Bari, Bergamo, Bologna, Cagliari, Catanzaro, Firenze, Genova, Lecce, Milano, Napoli, Palermo, Parma, Perugia, Roma, Venezia

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
 Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Tipologie testuali

37

### a) parlato radiotelevisivo

- (notiziari, interviste, talk shows);

### b) parlato dialogico

- (240 dialoghi raccolti secondo le modalità del map task e del 'gioco delle differenze', dei quali 30 etichettati foneticamente, 90 trascritti ortograficamente, studenti universitari);

### c) parlato letto da parlanti non professionisti

- (20 frasi atte a garantire la copertura delle frequenze medio-alte del lessico italiano);

### d) parlato telefonico

- (conversazioni tra circa 300 parlatori e un portiere d'albergo simulato)

### e) parlato letto da 20 parlanti professionisti

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## www.clips.unina.it

38

**CLIPS** Corpora e Lessici di Italiano Parlato e Scritto

Corpus CLIPS - [www.clips.unina.it](http://www.clips.unina.it) 03/04/2007

CLIPS (/corpus/) 7 subfolder(s) 0 file(s) Total Size: 21,70 GB

Nome	Dimensione	Tipo	Azione
DIALOGICO		Cartella di file	
LETTO		Cartella di file	
ORTOFONICO		Cartella di file	
RTV		Cartella di file	
TELEFONICO		Cartella di file	
trascrizioni ZIP		Cartella di file	
X_Materiali utili		Cartella di file	

WebExplorer Lite v 2.13 - Copyright © 2000-2004 GleanTech

menù

- presentazione
- coordinamento
- partners
- ringraziamenti
- documenti
- programmi
- descrizione del corpus
- accesso al corpus
- database
- links
- Area privata

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

<inspiration> in realtà la cultura svedese , è come se in qualche modo un po' ha sempre un po'<oo>' <inspiration> <eh> subito una<aa> <inspiration> una un po' una mancanza di fiducia in se stessa e quindi si è sempre nutrita di modelli , che in negli anni sessanta principalmente erano fondamentalmente l'America <inspiration>

add  
 wrd | in | realtà | ia | cultura |  
 std | in | realtà | ia | cultura |  
 phn | [h] [r] [e] [a] [l] [t] [a] [l] [a] [k] [u] [l] [t] [v] [u] [z] [a] [z] [v] [e] [d] |  
 acc | [ci] | [k\_c] | [t\_c] |

Spectrogram - 00.993 7273Hz -70.35dB - Formants(Hz):387 1636 2024 2497 3703 0 0.0763899

39

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
 Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## CLIPS in sintesi

40

### Pregi

Trascrizione e Annotazione

- Ortografica, fonetica e fonologica
- Standardizzata Eagles

Distribuzione

- Online gratuita
- Accesso al corpus integrale (scaricabile)
- Accesso anche all'audio
- Software per le analisi gratuiti online
- Documentazione dettagliata

### Difetti

Estensione

- Grande per indagini fonetiche, ma piccolo per gli altri livelli

Interrogazione e annotazione

- Nessuna annotazione grammaticale (attualmente)

Accesso corpus

- Mediante Ftp (un po' lento)

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
 Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

41

## CORPORA DI LABLITA

Laboratorio Linguistico del Dipartimento di  
Italianistica dell'Università di Firenze  
Diretto da Emanuela Cresti

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Corpora Lablita

42

1) un *corpus* di italiano parlato spontaneo adulto che raccoglie circa centoventi testi che riguardano situazioni comunicative diafasiche diverse per un totale di sessanta ore;

2) un *corpus* della lingua dei media (cinema, radio e televisione);

3) un *corpus* di cento ore di italiano registrato nella fase del primo apprendimento (in bambini di diciotto-trentasei mesi).

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Annotazioni e distribuzione

43

### Audio

- In questi *corpora* i testi sono trascritti, ma l'audio è disponibile in formato digitalizzato (.wav).

### Trascrizioni

- Le trascrizioni sono in formato CHAT (cfr. Childes)

### Distribuzione

- A richiesta
- Non interrogabili online

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

44

## C-ORAL-ROM

### *Integrated reference corpora for spoken romance languages*

E. Cresti - M. Moneglia  
2005

comparable set of corpora of spontaneous spoken language for the  
main romance languages, namely French, **Italian**, Portuguese and  
Spanish

**300,000** words for each language

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Caratteristiche C-ORAL-ROM

Comparabilità tra le quattro lingue romanze

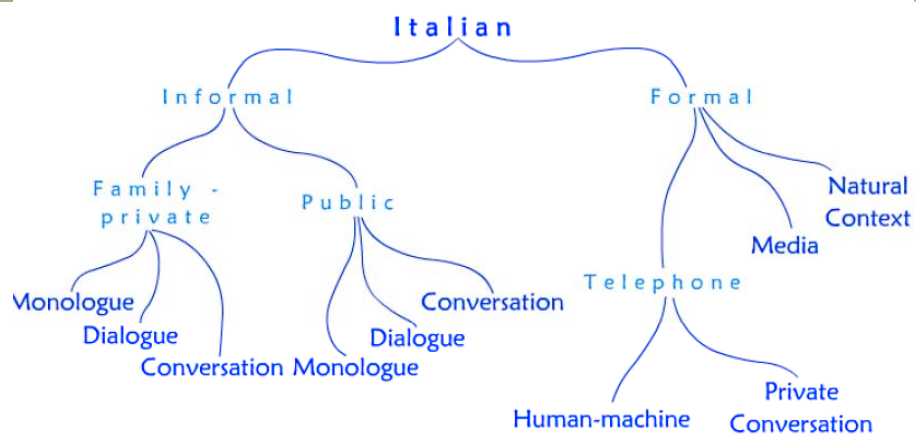
Distribuzione di Audio e trascrizione

Allineamento di audio e trascrizione con software (WinPitch)

tagging prosodico & grammaticale

45  
Fakulteta za humanistične študije  
Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile  
2007 - Isabella Chiari

## C-ORAL-ROM design



Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari





## C-ORAL-ROM in sintesi

49

### Pregi

#### Trascrizione e Annotazione

- Ortografica
- Annotazione prosodica e grammaticale
- Standardizzata CHAT
- Esportazione di concordanze e liste selezionate

#### Distribuzione

- Accesso al corpus integrale
- Accesso anche all'audio
- Software per le analisi gratuiti online
- Documentazione dettagliata

### Difetti

#### Estensione

- Piccola per indagini diverse da fonetica e prosodica

#### Interrogazione

- Non si possono interrogare sequenze

#### Accesso corpus

- A pagamento su cd-rom

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

50

## Piccola "guida" all'uso dei corpora

nella ricerca linguistica

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

## Analisi comparativa

51

	LIP	CLIPS	CORIS/CODIS	COLFIS	LA REPUBBLICA	C-ORAL- ROM
<b>Analisi fonetiche</b>		✓				(✓)
<b>Analisi morfosintattiche</b>	✓			(✓)	✓	✓
<b>Analisi lessicali</b>			(✓)	✓	✓	
<b>Accesso al testo integrale</b>	✓	✓				✓
<b>Accesso ai sottocorpora</b>	✓	✓	✓	✓	✓	✓
<b>Gratuito</b>	✓	✓	✓	✓	✓	
<b>Accesso online</b>	✓	✓	✓	✓	✓	

## Applicazioni

52

Linguistica descrittiva e statistica

Applicazioni lessicografiche

- dizionari cartacei ed elettronici

Grammatiche

- corpus-driven grammar

Trattamento automatico del linguaggio

- costruzione di parsers, taggers e lemmatizzatori che includano moduli di tipo statistico

Traduzione automatica

- corpus-based, example-based e statistica

Speech technologies

Didattica delle lingue

Fakulteta za humanistične študije Koper, Univerza na Primorskem,  
Capodistria, Slovenia, 5 aprile 2007 - Isabella Chiari

Fakulteta za humanistične študije Koper, Univerza na Primorskem, Capodistria,  
Slovenia, 5 aprile 2007 - Isabella Chiari

53

GRAZIE!

Isabella Chiari, le slides al sito: [www.alphabit.net](http://www.alphabit.net)