Teaching language variation

The exploitation of Italian dictionaries and corpora

ISABELLA CHIARI
Università La Sapienza di Roma
isabella.chiari@uniroma1.it

Language variation

- topical areas in teaching and developing language awareness
- *genre, register, text type, domain, sublanguage,* and *styles*
- Ambiguous categories and distinctions (e.g. Biber, 1989, 1994; Lee, 2001).

keywords

- **genre** «in this view, is defined as a category assigned on the basis of **external** criteria such as intended audience, purpose, and activity type, that is, it refers to a conventional, culturally recognised grouping of texts» (Lee, 2001)
- a **text type**, as «based on the **internal** (linguistic) properties […*such as*] than lexical or grammatical (co-)occurrence features».
- **register** is commonly used to refer to «the general cover term associated with all aspects of variation in use» (Biber, 1995: 9)
- **Activity types**, **domains**, **topics**

Language variation and corpora

- « …language teachers and researchers need to know exactly what kind of language they are examining or describing» (Lee, 2001).
- To observe language variation two features should be pointed out:
- «individual linguistic features are distributed differently across registers»
- and « the same (or similar) linguistic features can have different functions in different registers» Biber (2001: 221)

Relevance in language learning

- Is it possible to acquire knowledge and competence about variation by exploiting material extracted from electronically available resources?
- What kind of information about linguistic varieties is coded in dictionaries and corpora of the Italian language?
- Is it possible to exploit it in a learning environment in autonomous or in classroom activities?
- What are limitations in the resources for research and practice of language variation aspects?

Reference corpora of Italian

•Maximum language variation coverage
•representativeness, register-diversity and balance among typologies vs. specialized corpora
•linguistic variation in all its aspects is the key feature in reference corpus design

Written Italian	*Spoken Italian*
CORIS/CODIS COrpus di Riferimento dell'Italiano Scritto, 1998-2001 100 million words Free on the web for research purpose	**LIP** Corpus del *Lessico di frequenza dell'italiano parlato*, 1993 500,000 words Free on the web
COLFIS Corpus e Lessico di Frequenza dell'Italiano Scritto, 1995 3 million words Free on the web	**C-ORAL-ROM** Italian section, 2005 300,000 words not on the web, offline purchase

Where and how to look at linguistic variation?

- Corpus design
  - Balance of subcorpora, sections and subsections
- Retrieval tools
  - Capabilities in extracting lexical, morphosyntactic information about variation

CORIS/CODIS

COrpus Dinamico dell'Italiano Scritto

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

CILTA, Bologna (R. Rossini Favretti)

- **subcorpora** (**macro-varieties)** of traditional **text typologies**: press, fiction, academic prose, admin. and legal prose, miscell. ephemera
- **sections**
- **subsections**

FICTION

25 mw

novels, short stories

Italian, foreign,
for adults, for children
crime, adventure, science-fiction, women literature

ACADEMIC PROSE

12 mw

human sciences, natural sciences, physics, experimental sciences

books, reviews
scientific, popular history, philosophy, arts, literary criticism, law, economy, biology, etc.

LEGAL AND ADMIN. PROSE

10 mw

legal, bureaucratic, administrative

books, reviews

MISCELLANEA

10 mw

books on religion, travel, cookery, hobbies, etc.

books, reviews

EPHEMERA

5 mw

letters, leaflets, instructions

private, public printed form, electronic form

CORIS/CODIS

COrpus Dinamico dell'Italiano Scritto

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

CILTA, Bologna (R. Rossini Favretti)

- **subcorpora (macro-varieties)** of traditional **text typologies**: press, fiction, academic prose, admin. and legal prose, miscell. ephemera
- **sections**
- **subsections**

PRESS
(38 mw)

newspapers, periodic, supplement

national, local
specialist, non specialist
connotated, non connotated

CODIS - Corpus query form

Query

dare + un + esame Query Language Help.

☑ Case insensitive search

Subcorpus	Size (in Mw)			
STAMPA	☐ 20	☐ 10	☐ 5	☑ 3
NARRATIVA	☐ 13	☐ 7	☐ 3	☐ 2
PROSA ACCADEMICA	☐ 5	☐ 4	☐ 2	☐ 1
PROSA GIURIDICO-AMM.	☐ 4	☐ 3	☐ 2	☐ 1
MISCELLANEA	☐ 4	☐ 3	☐ 2	☐ 1
EPHEMERA	☐ 2	☐ 1	☐ 1	☐ 1

Query of *praticamente* in CODIS

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 ' osses
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
```


Query syntax and retrieval tools

- *only* **concordancing**
- impossibility of exploring entire texts
- maximum of 300 hits per query, while it still shows actual hit number satisfying query syntax
- **sections and subsections** cannot be queried separately
- le word types entries, specific sequences of words (and OR logical operator), sequences at a distance (which retrieves two words and any number of words in between).
- **Collocational values** (mutual information, t-score or raw frequency) is optionally showed.
- **No grammatical annotation** is given, **nor** any **frequency** count over the corpus.

CoLFIS (Corpus e Lessico di Frequenza dell'Italiano Scritto)

By P. M. Bertinetto, C. Burani, A. Laudanna, L. Marconi, D. Ratti, C. Rolando, A.M. Thornton

3.150.075 tokens

http://alphalinguistica.sns.it/BancheDati.htm

Corpus available (*only authorized portion*) at

www.ge.ilc.cnr.it/strumenti.php

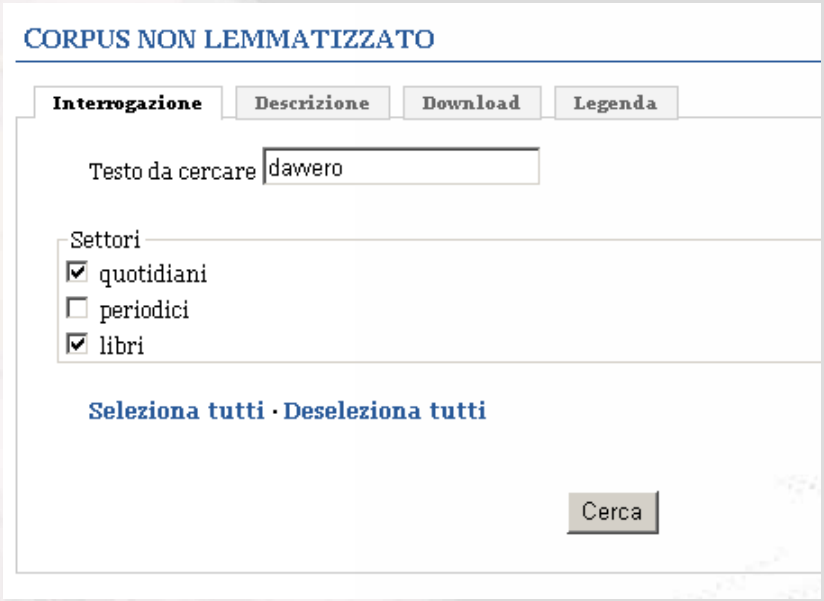
COLFIS balance

Design based on reading typology statistics (National Institute for Statistics) and on specific texts read

Subcorpora	NEWSPAPERS 1.523.167	MAGAZINES 1.084.574	BOOKS 542.334
Sections	*Il Corriere, La Repubblica, La Stampa*		
Subsections	economia, cronaca locale, cronaca mondana, cronaca nera, politica estera, politica interna, scienza, spettacolo e sport	altro, arte scienza e tecnica, auto e nautica, bambini e ragazzi, casa e hobby, femminili, fotoromanzi, informazione generale, cronaca mondana, radio e televisione, sport, viaggi e ecologia	altro, arte, bambini, fantascienza, gialli e spionaggio, hobby e viaggi, narrativa classica, narrativa moderna, rosa, saggistica, scienze naturali e esatte, scienze sociali e umane, teatro e poesia

Retrieval tools

- Only through *concordancing* (no full text access)
- Only a fraction of the entire corpus is available (the authorized portion).
- Queries can be made on the **raw corpus** or on the **lemmatized** version
- subcorpora can be searched, no subsection is searchable
- no query syntax defined at the moment, only word types and sequences can be searched
- Word frequency information is widely available,
  - frequencies, relative frequencies and dispersion values for the words (lemmas and inflected forms in the corpus).
- No hit count is given for a given search



"*sai*" in COLFIS

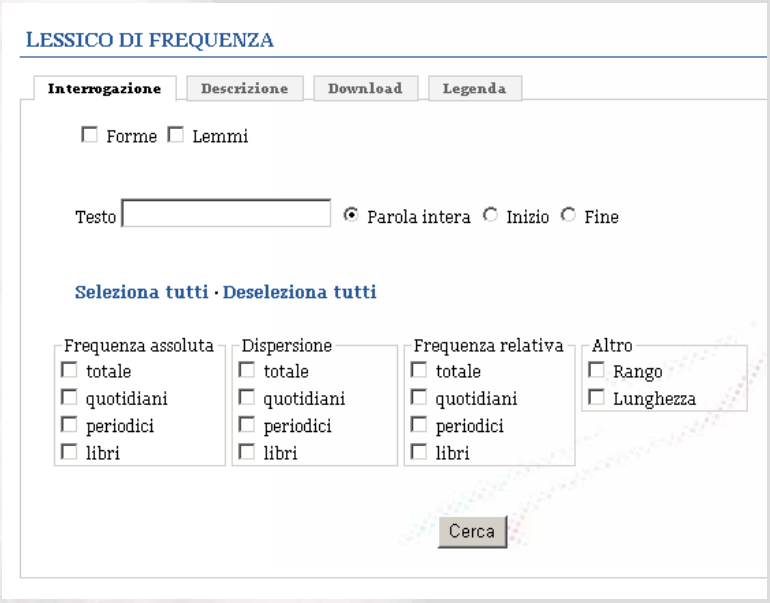
periodico	fotoromanzi	lancio-Lucky	Rienzi Alice	Immagini di una ragazza scomparsa			94-09-13	Janet è sempre piena di risorse, lo *sai*.
periodico	fotoromanzi	sogno	Mancuso Antonino	Sta suonando per me			92-12-01	Si sta divertendo, e poi lo *sai* che ci tiene alla tua presenza. Eppure, *sai*.
periodico	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 *sai*.
periodico	informazione general	espresso	Siciliano Enzo	Il film: che profumo di Cechov!			92-09-13	Non *sai* se abbia letto più Cechov o Trifonov.

"*defungere*" (lemma query)

quotidiano	cronaca mondana	repubblica	LAURA LAURENZI	IN COMUNE O IN CHIESA SPOSARSI È MEGLIO			92-01-04	

POSTUMO	postuma	Agg.
)	)	Punt.
DI	del	Prep.
IL	-l	Art.
MARITO	marito	Sost.
DEFUNGERE	defunto	Verbo
:	:	Punt.
ESSERE	è	Verbo
IL	il	Art.
CASO	caso	Sost.
DI	della	Prep.

- We will find out that the verb is mainly used in participio passato and passato prossimo
- In formal context, in the press



LIP- Lessico di frequenza dell'italiano parlato

De Mauro, Mancini, Vedovelli e Voghera (1993)

500.000 tokens

57 hours of speech

Free on the web

http://languageserver.uni-graz.at/badip/badip/20_corpusLip.php

Text typology	Included texts
A: bidirectional, exchange, face to face, with free turn-taking	-conversations at home; - conversations at work; - conversations at school or at the university; - conversations during recreation or on means of transport.
B: bi-directional exchange, not face to face, with free turn-taking	- normal telephone conversations; - telephone conversations broadcasted on radio; - messages recorded by telephone answering machines.
C: bi-directional exchange, face to face, with regulated turn-taking	- legislative assemblies; - cultural discussions; - assemblies at school; - labor union assemblies; - meetings of workers; - oral exams in the elementary school; - oral exams in the secondary school; - oral exams at the university; - interrogations in the courtroom; - interviews on radio or television

Text typology	Included texts
D: unidirectional exchange, with the addressee being present	- lessons in the elementary school; - lessons in the secondary school; - university lectures; - speeches held during party conventions or labor union meetings; - presentations at scientific meetings; - speeches held during electoral campaigns; - sermons; - presentations at non-specialist meetings; - court pleadings.
E: distanced unidirectional exchange	- television programs; - radio programs.

LIP (BADIP) search

search for all sequences that contain: (help)

type the first word or click
type the second word or click
type the third word or click

and that do not contain:

type the first word
type the second word
type the third word

in the cities: ☑ Firenze ☑ Milano ☑ Napoli ☑ Roma
in the text types: ☑ A ☑ B ☑ C ☑ D ☑ E

search

praticamente in LIP

occurrences of found lemma/form: **203**
of total of graphical words: **489178**

statistics:
1) praticamente

save displayed data: Microsoft Excel (xls)

parola	TIPO A	TIPO B	TIPO C	TIPO D	TIPO E	TOT
praticamente	**48**	**42**	**56**	**20**	**37**	**203**

concordance of *praticamente*

city	type	conversation	utterance	speaker	utterance
F	A	9	35	A	non deambulante cioe' non deve praticamente non potrebbe eh dovrebbe essere in condizione di non muoversi da solo
F	A	10	76	B	e' solo quello di sotto * ah si' perche' praticamente e' due metri e quarantacinque eh quindi a parte che non si sa se le abitabilita' queste abitabilita' le daranno o no ti ricordi $ $ *
F	A	10	249	A	quindi praticamente tu paghi adesso
F	A	12	162	B	* si lascia cosi' com' e' perche' praticamente qui dice
F	A	12	170	B	* cioe' il sabato e la domenica la dovrebbe fa il portiere praticamente
F	A	13	97	A	discretamente discretamente in quegli altri che erano eh una volta doveva inventare un racconto un breve racconto su una traccia che avevo dato non l' ha fatto bene gliel' ho fatto rifare ma praticamente anche il recupero
F	B	5	71	A	e perche' e' inutile io gli dico anche le cose belle sono la prima a

Strong point in retrieval tasks

- Query syntax accepts:
  - wildcards,
  - some of tags,
  - lemma search.
- **Subcorpora** can be queried separately
- Simple **statistics** (frequency and subcorpora frequency) with single query
- **Exportation** in various formats (txt, excel, html; xml, xsl, dtd)
- Very useful and easy in classroom activities.

Disadvantages of LIP online

- Size is small and not reliable for medium frequency words
- No access to audio is available
- No phonetic or prosodic annotation
- No queries possible over grammatical tags (without specification of lemma or form)
- Full access to whole texts is not available online (only on floppy)

C-ORAL-ROM

Integrated reference corpora for spoken romance languages

E. Cresti - M. Moneglia

2005

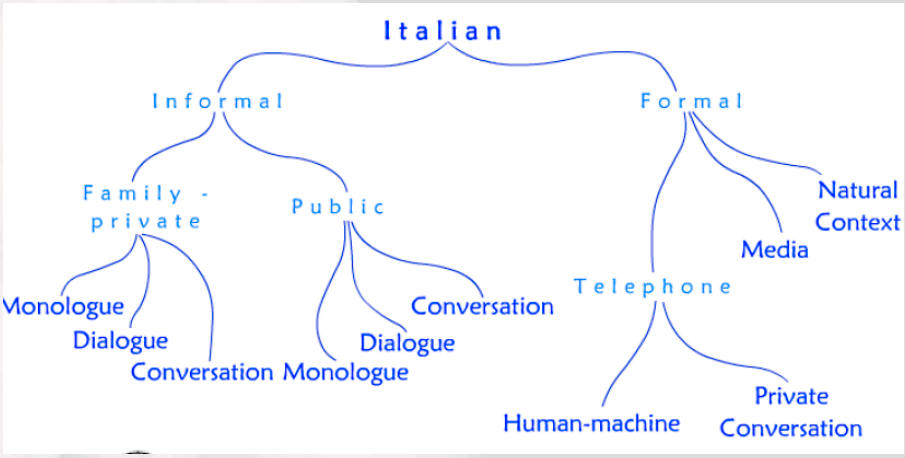
comparable set of corpora of spontaneous spoken language for the main romance languages, namely French, Italian, Portuguese and Spanish

300,000 words for each language

Features of C-ORAL-ROM

- *comparability* throughout the four Romance languages corpora
- both Audio and Text Analysis
- Simultaneous and aligned access to acoustic and textual information
- tagging with respect to prosodic parsing & action values of the all textual information

C-ORAL-ROM design



INFORMAL	155,048
type	**sub-type**
family/private	conversation
family/private	dialogue
family/private	monologue
public	conversation
public	monologue
FORMAL - natural context	68,324
Political speech	monologue
Preaching	monologue
Conferences	monologue
Prof. Explanation	conversation
Business	dialogue
Law	monologue
Teaching	monologue

FORMAL - natural context	68,324
Political speech	monologue
Preaching	monologue
Conferences	monologue
Prof. Explanation	conversation
Business	dialogue
Law	monologue
Teaching	monologue

FORMAL- Telephone	26,582
Genre	
private conversation	
man-machine interaction	

FORMAL - media	61,638
Genre	
interviews	
meteo	
news	
sport	
scientific press	
reportages	
talk show	

TOTAL FORMAL	129,962
TOTAL CORPUS	285,010

subordinative conjunction *che*

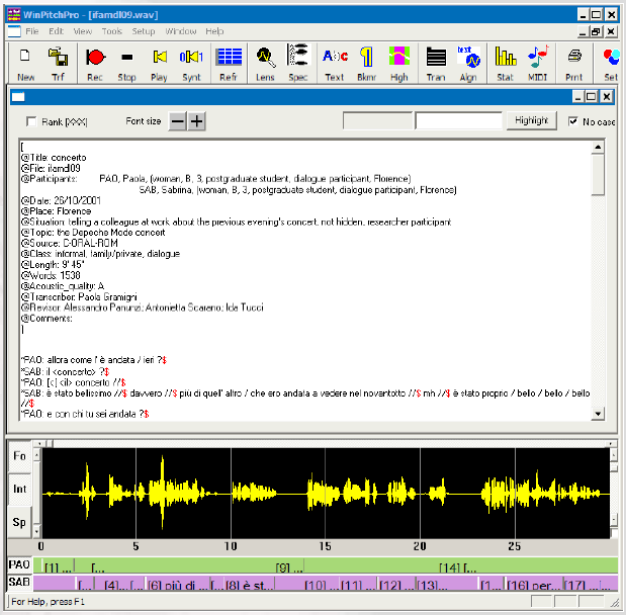
Contextes

File Edit Options Language ?

\CHE\CS | Word | Reg. exp.

#	File	Left context	Match	Right context
1	ifamcv01	? per\PER\E farsi\FARE\Vf perdonare\PERDONARE\Vf / che	\CHE\CS	+ che\CHE\REL si\SI\PER doveva\DOVERE\Vs3ii fa'\FARE\
2	ifamcv01	*MAX: ma\MA\CC che	\CHE\CS	/ Baratti\BARATTI\SP è\ESSERE\Vs3ip / in\IN\E Toscana
3	ifamcv01	ualcosa\QUALCOSA\IND di'\DI\E_R genere\GENERE\S // che	\CHE\CS	l'\LO\PER ho\AVERE\Vslip addosso\ADDOSSO\E / questo\Q
4	ifamcv01	*ELA: che\CHE\REL [/] che	\CHE\CS	[///] la\IL\R seicento\SEICENTO\S //
5	ifamcv01	ON\B so'\ESSERE\V come\COME\B Luigina\LUIGINA\SP / che	\CHE\CS	lei\LEI\PER / &sco\PLG [/] scorreva\SCORRERE\Vs3ii tu
6	ifamcv01	QUESTO\DIM / son\ESSERE\VW persone\PERDERE\Vmspr_E che	\CHE\CS	non\NON\B conosco\CONOSCERE\Vslip // questa\QUESTO\DI
7	ifamcv01	*LIA: che	\CHE\CS	era\ESSERE\Vs3ii ? se\SE\CS avevo\AVERE\Vslii / &dici
8	ifamcv01	*ELA: che	\CHE\CS	l'\LO\PER è\ESSERE\Vs3ip / mezza\MEZZO\A montagna\MON
9	ifamcv01	*LIA: o\O\CC che	\CHE\CS	lo\IL\R so\SAPERE\Vslip / se\SE\PER saranno\ESSERE\VW
10	ifamcv01	RE\Vs2mp un\UNO\IND po'\PO'\B vedere\VEDERE\Vfp // che	\CHE\CS	è\ESSERE\Vs3ip ?
11	ifamcv01	*MAX: che	\CHE\CS	poteva\POTERE\Vs3ii fare\FARE\Vfp / disgraziata\DISGR
12	ifamcv01	Vs3ii i'\IL\R pesce\PESCE\S // voleva\VOLERE\Vs3ii che	\CHE\CS	lo\LO\PER cucinassi\CUCINARE\V //
13	ifamcv01	\CC pensa\PENSARE\V [/] ma\MA\CC pensa\PENSARE\V / che	\CHE\CS	/ alla\A\E_R sorella\SORELLA\S di\DI\E Virgilio\VIRGI
14	ifamcv01	l\A\E_R cervello\CERVELLO\S // senti\SENTIRE\Vs2ip che	\CHE\CS	lavoro\LAVORO\S // hhh\XLG 'un\NON\B ne\NE\E vòi\VOI\
15	ifamcv01	BBASTANZA\B giovani\GIOVANE\A // questi\QUESTO\DIM che	\CHE\CS	son\ESSERE\VW morti\MORIRE\Vmppr / che\CHE\REL tu\TU\
16	ifamcv01	*LIA: che	\CHE\CS	sarebbe\ESSERE\Vs3dp / la\IL\R sorella\SORELLA\S / di
17	ifamcv01	\CHE\REL ho\AVERE\VWslip scoperto\SCOPRIRE\Vmspr / che	\CHE\CS	ero\ESSERE\VWslii cattivo\CATTIVO\A //
18	ifamcv01	*LIA: che	\CHE\CS	dici\DIRE\Vs2ip te\TE\PER ?
19	ifamcv01	// bella\BELLO\A / 'esta\QUESTO\DIM foto\FOTO\S // che	\CHE\CS	l'\LO\PER hanno\AVERE\VWp3ip preso\PRENDERE\Vmspr / m
20	ifamcv01	STARE\Vs3ii lavorando\LAVORARE\Vg / lui\LUI\PER // che	\CHE\CS	c'\CI\PER aveva\AVERE\VWs3ii + Liana\LIANA\SP / non\N
21	ifamcv01	PERCHE'\CS / hanno\AVERE\VWp3ip detto\DIRE\Vmspr / che	\CHE\CS	il\IL\R prossimo\PROSSIMO\S anno\ANNO\S / ci\CI\PER s
22	ifamcv01	E questa\QUESTO\DIM mostra\MOSTRA\S / si\SI\PER // che	\CHE\CS	siccome\SICCOME\CS +
23	ifamcv01	*ELA: hhh\XLG che	\CHE\CS	[/] che_cosa\CHE_COSA\REL //
24	ifamcv01	NO\S in\IN\E questa\QUESTO\DIM maniera\MANIERA\S / che	\CHE\CS	ce\CI\B ne\NE\PER sono\ESSERE\V trenta\TRENTA\N //
25	ifamcv01	\DIPENDERE\Vs3ip anche\ANCHE\CC / Liana\LIANA\SP / che	\CHE\CS	lei\LEI\PER è\ESSERE\VWs3ip stata\ESSERE\Vfspr tanto\
26	ifamcv01	SERE\VWs3ip una\UNO\R cosa\COSA\S anche\ANCHE\CC / che	\CHE\CS	l'\LO\PER ha\AVERE\Vs3ip dovuto\DOVUTO\S fare\FARE\Vf

21478 parag. | 3221 matches | 9,7 s | Version: Demo | File: C:\Documents and Settings\Isabella Chiari\Documenti\MATERIALI LINGUISTICA\CORPORA\CORALROM\Textual_Corpu



Retrieval pros and cons

- *Contextes* by Jean Véronis for concordance
- Query syntax accepts:
  - Word types, lemmas, pos tagging, multiple word queries at a time, regular expression support
- Direct access to full text
- Export of concordance and frequency lists of selected words
- No general usage/frequency and dispersion data is given for sections of the corpus
- No data extraction for sequences of words
- Size is small and not reliable for medium frequency words

Italian electronic dictionaries

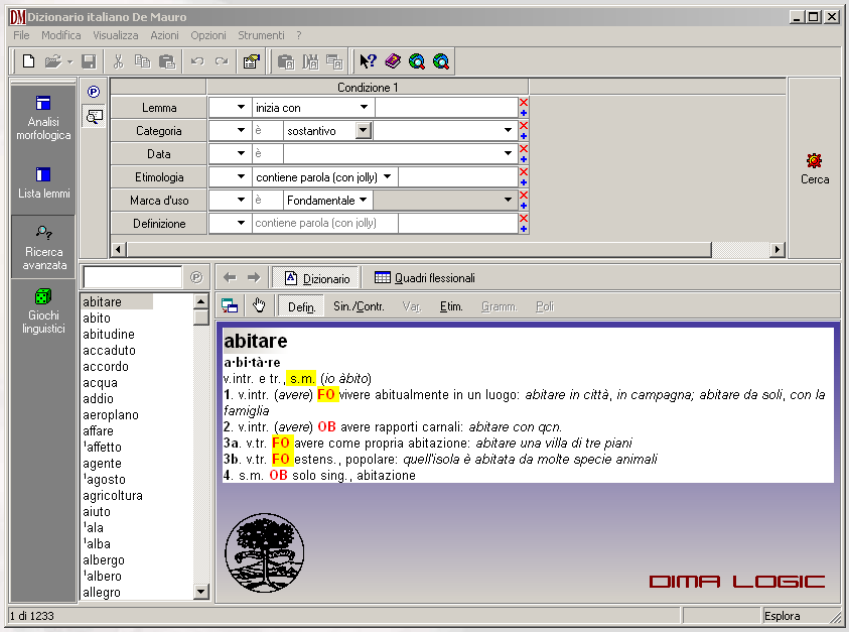
- No electronic learners' dictionary for the Italian language available
- no a fully corpus-based dictionary
- aimed at native speakers
- 4 dictionaries have been analyzed:
  1) T. De Mauro: *Dizionario della lingua italiana*. Paravia 2000
  2) Lo Zingarelli 2006: *Vocabolario della lingua italiana*. Zanichelli 2005
  3) Sabatini-Coletti: *DISC - Dizionario della Lingua Italiana*. Rizzoli Larousse 2006
  4) De Mauro: GRADIT - *Grande dizionario italiano dell'uso*. UTET 1999-2000 (cd 2003 con nuove parole)

Variation through usage labels

- Frequency of usage labels
  - fundamental (2,000), high frequency (3,000), high availability (2,000), common, low frequency (DMP; GRADIT): *core vocabulary*
  - General high frequency (10,000), not entirely based on statistics or experimental data (ZING, DISC)
- Genre/type labels
  - literary, specialized, poetry (DMP; GRADIT, ZING, DISC)
- Register
  - familiar, jargon, ironic, joking, vulgar, popular (ZING, DISC), children talk (ZING)
- Style/usage
  - emphatic, euphemistic , figurative, negative, positive (ZING)

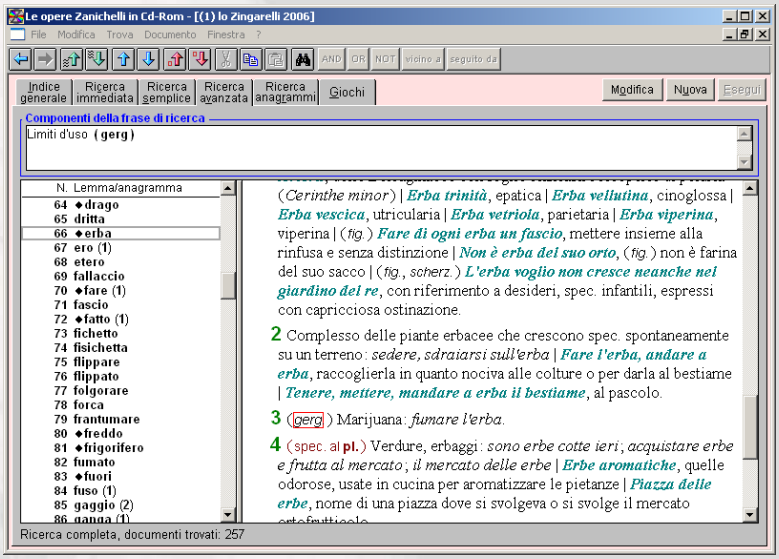
Other dimensions

- Domain labels (label set)
  - arts, anatomy, bibliography, etc. (GRADIT, DISC, ZING)
- Diatopic labels
  - regional, dialect, loanword*, geographic area (label set)* (DMP, GRADIT, DISC, ZING)
- Diachronic labels
  - Obsolete / archaic, *Date* (label set) (DMP, GRADIT, DISC, ZING)



41

jargon usage of *erba*



42

Using corpora…

- Although corpora show inconsistencies in classification and labelling of variation, they still contribute to direct experience in typology comparison, and in linguistic features observation through retrieval tools.
- Still the teacher does not always know what is inside a corpus, and what to expect in texts included in a corpus.
- When spoken corpora offer access to audiowaves in a aligned form the user can observe actual spoken realizations of different registers (phonetic and prosodic features).
- Written and spoken corpora cannot be **compared** reliably, since size, design, language processing and text tools are completely different.

And using dictionaries…

- Usage labels generally give an insight on language variation
  - On frequency, genre/typology, register, style, domain, geographic variation, and diachronic variation
  - Prototypical examples can be used as keys for further corpus activities
  - No specific mention is ever made over differences in written and spoken usage
- Different dictionaries differently code this information

Variation in corpora and dictionaries: what to do?

- curriculum design in the selecting and sequencing process
- language materials and resources design:
  - activities of variation discovery procedures, etc.
  - activities dictionary > corpus > dictionary
- In development and grading of linguistic activity exposure
- in classroom work methodology
- in assessment procedures

Thank you!

slides will be available on
http://www.alphabit.net/
Docente/Pubblicazioni.htm