The PAISA' Project

Marco Baroni

HLT for Italian Workshop

PAISA’

Coming soon at `www.corpusitaliano.it`

- Piattaforma per l’Apprendimento dell’Italiano Su Corpora Annotati
- MIUR FIRB project 2009-2012
- Partners:
  - University of Bologna: Sergio Scalise, Claudia Borghetti, Emiliano Guevara
  - University of Trento: Marco Baroni, Marco Brunello, Sara Castagnoli, Egon Stemle
  - ILC, Pisa: Vito Pirrelli, Alessandro Lenci, Felice Dell’Orletta
  - EURAC, Bolzano: Andrea Abel, Verena Lyding, Christopher Culy

Goals

- Build and annotate Web-derived “reference” corpus of Italian language
- Innovative corpus visualization and exploration tools
- Applications in Italian language and culture teaching

The WaCky initiative

http://wacky.sslmit.unibo.it

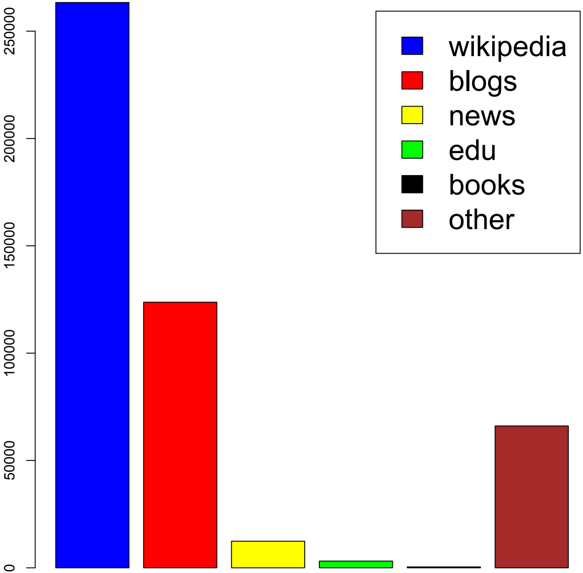
- Very large Web-crawled, automatically cleaned and annotated corpora for English, French, German and Italian (2005-2009)
- Lessons learnt:
  - Copyright
  - Corpus composition
  - Cleaning
  - Annotating Web text

Copyright

- PAISA’ corpus constructed by crawling Web pages published under CreativeCommons License only:
  - Attribution
  - Derived works OK
  - Non-commercial derived works only
- Same licensing propagates to PAISA’ corpus

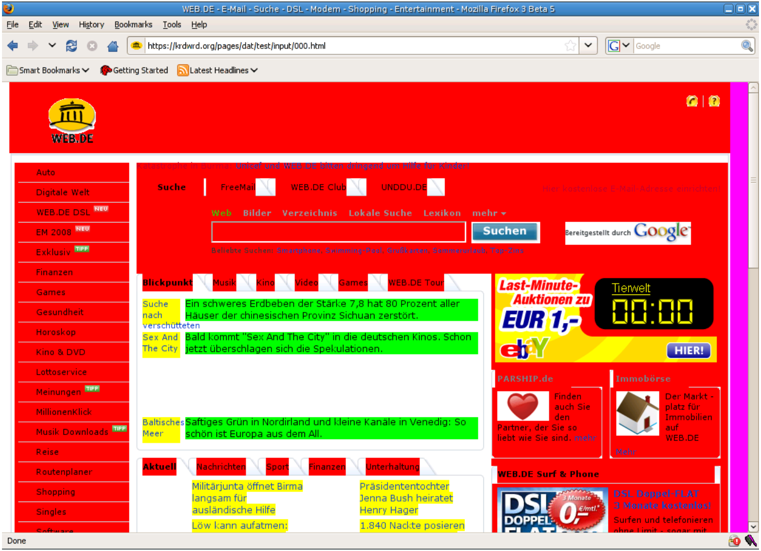
Corpus composition

470K documents



Web cleaning:

The boilerplate problem



Katastrophe in Burma: Unicef und WEB.DE bitten dringend um Hilfe für Kinder!

Hier kostenlose E-Mail-Adresse einrichten!

Beliebte Suchen: Smartphone, Swimming-Pool, Grußkarten, Sommerurlaub, Top-Zins

Suche nach verschütteten

Ein schweres Erdbeben der Stärke 7,8 hat 80 Prozent aller Häuser der chinesischen Provinz Sichuan zerstört.

Sex And The City

Bald kommt "Sex And The City" in die deutschen Kinos. Schon jetzt überschlagen sich die Spekulationen.

Baltisches Meer

Saftiges Grün in Nordirland und kleine Kanäle in Venedig: So schön ist Europa aus dem All.

Militärjunta öffnet Birma langsam für ausländische Hilfe

Löw kann aufatmen:

Präsidententochter Jenna Bush heiratet Henry Hager

1.840 Nackte posieren

Web cleaning on steroids: krdwrd

Cues from text

Main Index Page

General Ratings Page

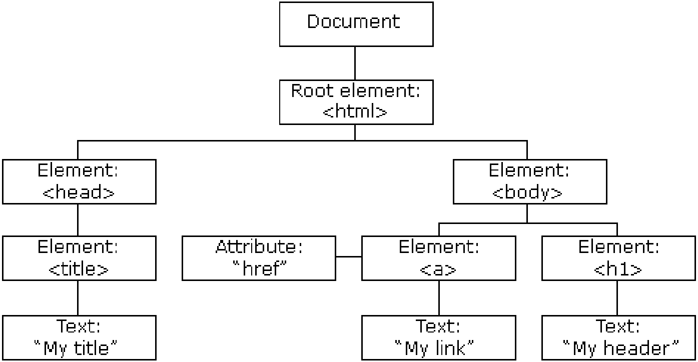
One true sign of a truly great band is when said band ardently defies categorisation, that is, when for every "well, they sound like this reggae-influenced heavy metal band playing avantgarde bebop" remark you can have yourself a "funny, I thought they were this raw punk outfit doing acoustic folk" counterproposal. And I don’t simply mean "being diverse" here, I mean "being different". Blazing off every colour of the spectrum. Baring one’s soul in all of its existing aspects. That sort of thing.

READER COMMENTS SECTION

Return to the Index page! NOW!

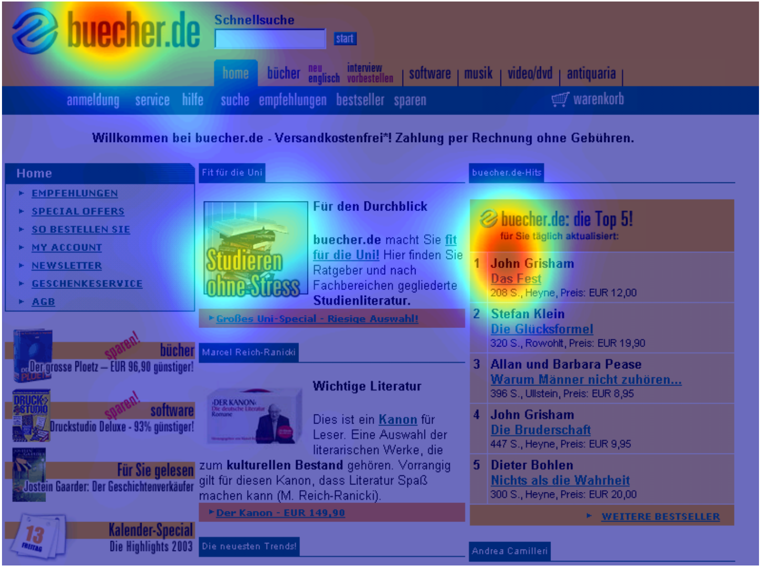
Web cleaning on steroids: krdwrd

Cues from the DOM tree (figure from http://www.w3schools.com/)



Web cleaning on steroids: `krdwrd`

Cues from visual rendering



Willkommen bei buecher.de - Versandkostenfrei*! Zahlung per Rechnung ohne Gebühren.

Für den Durchblick

buecher.de macht Sie fit für die Uni! Hier finden Sie Ratgeber und nach Fachbereichen gegliederte **Studienliteratur.**

Großes Uni-Special - Riesige Auswahl!

Wichtige Literatur

Dies ist ein Kanon für Leser. Eine Auswahl der literarischen Werke, die zum **kulturellen Bestand** gehören. Vorrangig gilt für diesen Kanon, dass Literatur Spaß machen kann (M. Reich-Ranicki).

Der Kanon - EUR 149,90

buecher.de: die Top 5!
für Sie täglich aktualisiert:

1 **John Grisham**
Das Fest
208 S., Heyne, Preis: EUR 12,00

2 **Stefan Klein**
Die Glücksformel
320 S., Rowohlt, Preis: EUR 19,90

3 **Allan und Barbara Pease**
Warum Männer nicht zuhören...
396 S., Ullstein, Preis: EUR 8,95

4 **John Grisham**
Die Bruderschaft
447 S., Heyne, Preis: EUR 9,95

5 **Dieter Bohlen**
Nichts als die Wahrheit
300 S., Heyne, Preis: EUR 20,00

WEITERE BESTSELLER

Web cleaning on steroids

Cross-validation on gold standard

	features	precision	recall	F
wacky heuristics	NA	80%	99%	88%
text cues only	21	92%	93%	92%
DOM cues only	13	89%	91%	90%
visual cues only	8	90%	93%	91%
full krdwrd	42	93%	92%	92%

Corpus comparison

- ITWAC: 1.9B tokens
- PAISA’: 700M tokens
- La Repubblica: 380M tokens

PAISA’ vs. La Repubblica

nouns		verbs		adjs	
PAISA’	*Rep.*	*PAISA’*	*Rep.*	*PAISA’*	*Rep.*
punto	miliardo	registrare	dire	informatico	scorso
blog	anno	pubblicare	chiedere	simile	politico
commento	presidente	leggere	andare	esterno	pubblico
settembre	governo	segnare	spiegare	hi-tech	primo
guida	ministro	inserire	arrivare	opzionale	stesso
articolo	lira	riservare	restare	famoso	grande
galleria	paese	inviare	parlare	successivo	generale
video	giorno	usare	sembrare	gay	finanziario
set	partito	contattare	cominciare	digitale	americano
pubblicità	stato	mostrare	mettere	bello	lungo

PAISA' vs. ITWAC

nouns		verbs		adjs	
PAISA'	*ITWAC*	*PAISA'*	*ITWAC*	*PAISA'*	*ITWAC*
punto	numero	registrare	svolgere	ultimo	pubblico
blog	articolo	pubblicare	chiedere	pubblicitario	regionale
commento	legge	segnare	ritenere	hi-tech	presente
guida	comma	leggere	presentare	simile	sociale
settembre	servizio	riservare	dire	opzionale	previsto
galleria	lavoro	commentare	tenere	esterno	nazionale
set	presidente	mostrare	riguardare	informatico	necessario
video	decreto	contattare	andare	famoso	generale
pubblicità	commissione	continuare	consentire	nuovo	seguente
e-mail	caso	deselezionare	prevedere	grande	legislativo

Ongoing work

- Annotation (POS tagging, dependency parsing. . .)
- Classification by genre and topic
- Visualization, user-friendly interface
- E-mail me (`marco.baroni@unitn.it`) if you want a copy of the raw corpus!