

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>

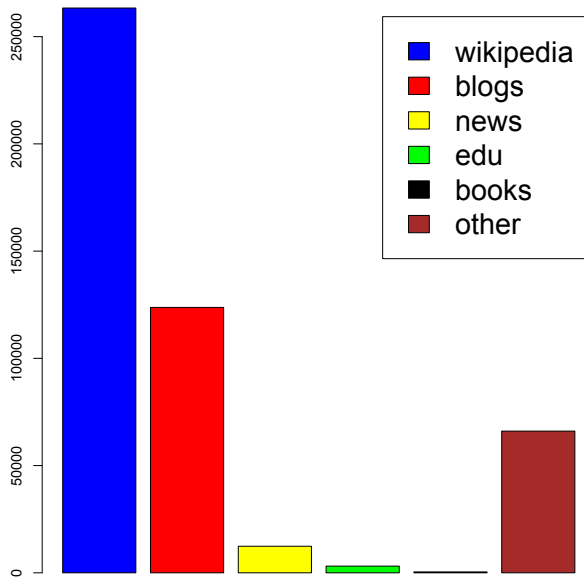
- ▶ Very large Web-crawled, automatically cleaned and annotated corpora for English, French, German and Italian (2005-2009)
- ▶ Lessons learnt:
 - ▶ Copyright
 - ▶ Corpus composition
 - ▶ Cleaning
 - ▶ Annotating Web text

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

WEB.DE - E-Mail - Suche - DSL - Modem - Shopping - Entertainment - Mozilla Firefox 3 Beta 5

File Edit View History Bookmarks Tools Help

https://krdwd.org/pages/dat/test/irput/000.html

Smart Bookmarks Getting Started Latest Headlines

WEB.DE

Auto
Digitale Welt
WEB.DE DSL
EM 2008
Exklusiv
Finanzen
Games
Gesundheit
Horoskop
Kino & DVD
Lottoservice
Meinungen
MillionenKlick
Musik Downloads
Reise
Routenplaner
Shopping
Single
Software

Suche | FreeMail | WEB.DE Club | UNDDU.DE

Web Bilder Verzeichnis Lokale Suche Lexikon mehr >

Suchen Bereitgestellt durch Google

Blickpunkt | Musik | Kino | Video | Games | WEB.DE Tour

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

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

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

Aktuell | Nachrichten | Sport | Finanzen | Unterhaltung

Militärjunta öffnet Burma langsam für ausländische Hilfe
Präsidententochter Jenna Bush heiratet Henry Hager
Löw kann aufatmen
1.840 Nädte posieren

Last-Minute-Auktionen zu EUR 1,-
Tierwelt 00:00
ebay HIER!

PARSHIP.de Finden auch Sie den Partner, der Sie so liebt wie Sie sind.
Immobilier Der Markt - platz für Immobilien auf WEB.DE

WEB.DE Surf & Phone
DSL Doppel-CLAR 3 Monate kostenlos! Surfen und telefonieren ohne Limit - sogar mit

Done

Web cleaning on steroids: krdwrdr

Cues from text

Main Index Page

General Ratings Page

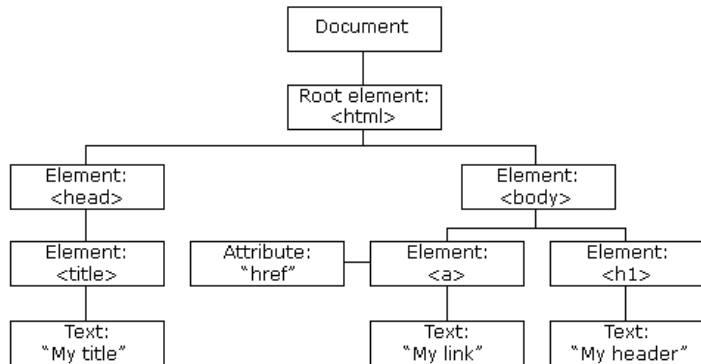
One true sign of a truly great band is when said band ardently defies categorisation, that is, when for every "well, they sound like this reggae-influenced heavy metal band playing avantgarde bebop" remark you can have yourself a "funny, I thought they were this raw punk outfit doing acoustic folk" counterproposal. 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: `krdwrdr`

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



Web cleaning on steroids: krdwrdr

Cues from visual rendering

buecher.de Schnellsuche

[home](#) [bücher](#) [neu](#) [englisch](#) [interview](#) [vorbestellen](#) [software](#) [musik](#) [video/dvd](#) [antiquaria](#)

[anmeldung](#) [service](#) [hilfe](#) [suche](#) [empfehlungen](#) [bestseller](#) [sparen](#)

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

Home

- ▶ EMPFEHLUNGEN
- ▶ SPECIAL OFFERS
- ▶ SO BESTELLEN SIE
- ▶ MY ACCOUNT
- ▶ NEWSLETTER
- ▶ GESCHENKESERVICE
- ▶ AGB

Fit für die Uni

Für den Durchblick

Studieren ohne Stress

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

▶ **Großes Uni-Special - Riesige Auswahl!**

buecher.de-Hits

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**

bücher sparen!
Der grosse Ploetz – EUR 96,90 günstiger!

software sparen!
Druckstudio Deluxe - 93% günstiger!

Für Sie gelesen
Jostein Gaarder: Der Geschichtenverkäufer

Kalender-Special
Die Highlights 2003

Der Kanon - EUR 149,90

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**

Die neuesten Trends!

Andrea Camilleri

Web cleaning on steroids

Cross-validation on gold standard

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

Corpus comparison

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

PAISA' vs. La Repubblica

nouns

<i>PAISA'</i>	<i>Rep.</i>
punto	miliardo
blog	anno
commento	presidente
settembre	governo
guida	ministro
articolo	lira
galleria	paese
video	giorno
set	partito
pubblicità	stato

verbs

<i>PAISA'</i>	<i>Rep.</i>
registrare	dire
pubblicare	chiedere
leggere	andare
segnare	spiegare
inserire	arrivare
riservare	restare
inviare	parlare
usare	sembrare
contattare	cominciare
mostrare	mettere

adjs

<i>PAISA'</i>	<i>Rep.</i>
informatico	scorso
simile	politico
esterno	pubblico
hi-tech	primo
opzionale	stesso
famoso	grande
successivo	generale
gay	finanziario
digitale	americano
bello	lungo

PAISA' vs. ITWAC

nouns

<i>PAISA'</i>	<i>ITWAC</i>
punto	numero
blog	articolo
commento	legge
guida	comma
settembre	servizio
galleria	lavoro
set	presidente
video	decreto
pubblicità	commissione
e-mail	caso

verbs

<i>PAISA'</i>	<i>ITWAC</i>
registrare	svolgere
pubblicare	chiedere
segnare	ritenere
leggere	presentare
riservare	dire
commentare	tenere
mostrare	riguardare
contattare	andare
continuare	consentire
deselezionare	prevedere

adjs

<i>PAISA'</i>	<i>ITWAC</i>
ultimo	pubblico
pubblicitario	regionale
hi-tech	presente
simile	sociale
opzionale	previsto
esterno	nazionale
informatico	necessario
famoso	generale
nuovo	seguito
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!