PAISA:

A Creative Commons corpus

Marco Brunello
SMLC/CTS
University of Leeds

Before PAISA

- WaCky (Web-as-Corpus kool ynitiative)
  - http://wacky.sslmit.unibo.it
- Advantages given by the Internet to build corpora:
    - Very large amount of texts
    - Still in electronic format
    - Freely available
    - But...
- Are they freely available for real?
- The problem of **Copyright**

Copyright vs. Copyleft

All rights reserved

- The author creator of an original work has exclusive rights (e.g. copy, distribute, adapt) on the work
- Everybody else need to ask permission to the copyright owner
- Some limited exceptions: Fair use (United States) and Fair dealing (Commonwealth)

Some rights reserved

- The author of an original work decides to share some rights (e.g copy, distribute, adapt) on the work with users of that work
- No need to ask permission to the copyright owner
- Several projects (GNU, Creative Commons) found wide use, especially on the web

A possible solution

- Build web corpora using exclusively documents released under a copyleft regime
- *The creation of free linguistic corpora from the web*
  - Comparison between a “normal” web corpus and a corpus exclusively made with Creative Commons-licensed documents
- Results have shown that CC is employed enough throughout the web to consider this as a practicable road

The PAISA project

PAISA = Piattaforma per l'Apprendimento dell'Italiano Su corpora Annotati

(Platform for Corpus-Assisted Italian Language Learning)

Funded by Fondo per gli Investimenti della Ricerca di Base (FIRB)
Ministero dell'Università e della Ricerca

June 2009 to June 2012

Entities involved

- University of Bologna

  Sergio Scalise, Claudia Borghetti, Emiliano Guevara

- University of Trento

  Marco Baroni, Marco Brunello, Sara Castagnoli, Egon Stemle

- Institute of Computational Linguistics – CNR of Pisa

  Vito Pirrelli, Alessandro Lenci, Felice Dell'Orletta

- European Academy of Bozen/Bolzano

  Andrea Abel, Verena Lyding, Christopher Culy

Purposes

- Particularly targeted for Italian language and teaching to 2nd generation of Italian emigrants
- Build a large reference corpus of Italian language
  - From the Internet (web corpus)
  - Using non-copyrighted texts
  - Richly annotated (tagging, parsing...)
- Make the corpus freely available online
  - Development of a web-based interface to the corpus with advanced search and analysis options

Building the corpus

1. Selection of documents
2. Download corpus
3. Clean corpus
4. Further operations

Selection of documents

- Automatic procedure with search engine, using BootCaT with Yahoo!
  - Yahoo!'s option to select pages licensed with Creative Commons licenses, in particular
    - CC – Attribution
    - CC – Attribution – Share alike
    - CC – Attribution – Non commercial – Share alike
  - Using as seed pairs random combinations of words
    - Taken from Vocabolario di Base della lingua italiana (VdB) by Tullio De Mauro
    - 50,000 tuples

Download and clean the corpus

- Download
  - URL selected with Yahoo! downloaded as cleaned text with the http://krdwrd.org system
    - Boilerplate stripping with visual rendering
- Further cleaning
  - Removal of pages wrongly classified by the search engine as CC-documents
    - Basing on previous experiments which allowed to build a black list of sites wrongly recognized as CC by Yahoo!
  - Removal of empty and undersized files (<150 words)

Further operations

- Integration with other CC-licensed documents
  - Manually selected from
    - Wikipedia
    - Wikibooks
    - Wikinews
    - Wikisource
    - Wikiversity

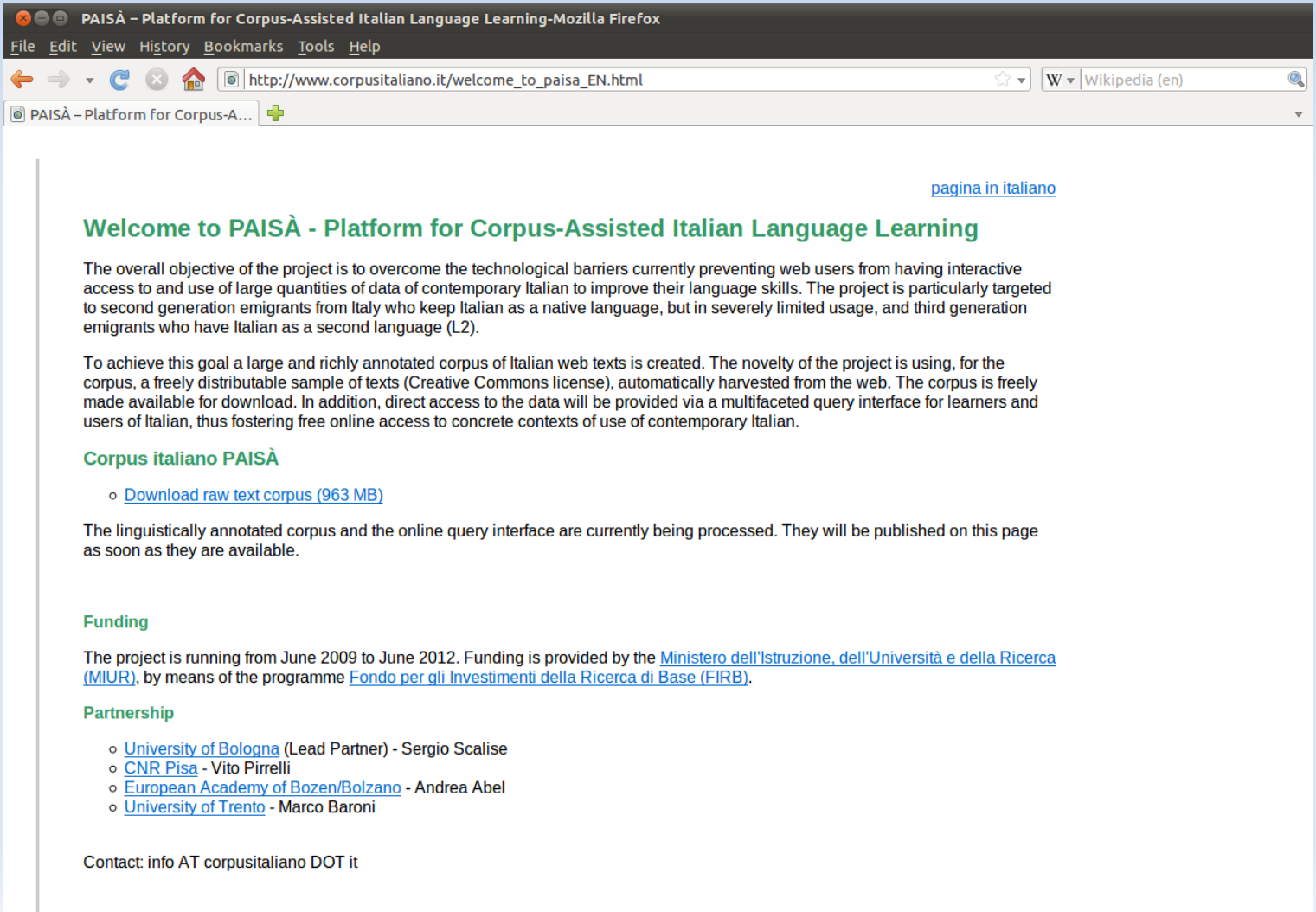
    using official dumps of Wikimedia Foundation: http://download.wikimedia.org/backup-index.html
  - Removal of empty and undersized files as before

Details

- Documents
  - BootCaT: 200,521
  - Wikis: 268,567
  - Total: 469,088
- Tokens
  - BootCaT: 506,638,863
  - Wikis: 193,393,072
  - Total: 700,031,935

PAISA now

Available in raw text version at http://www.corpusitaliano.it



Ongoing and future works

- Classification (Bologna, Trento)
  - Topic domain, genre type and communicative intention with a machine learning approach
  - Clustering
- Linguistic annotation (Pisa)
  - POS-tagging accuracy at 96.03%
  - Parsing
- User-friendly interface (Bozen/Bolzano)
  - Freely available online
  - With advanced exploration options