Web2Corpus_it: a balanced pilot corpus of conversational computer mediated communication
NEGOTIATING MEANING IN CMC

Sapienza Università di Roma

Coord. I. Chiari

Research team: Federico Albano Leoni, Sergio Bolasco, Francesco De Renzo, Sabine Koesters Gensini, Ilaria Tani
Giulia Petitta, Alessio Canzonetti, Cristina Antonetti, Matteo Servilio, Marta Di Cola, Vincenzo Vasco, Giovanni De Gasperis
Web2Corpus_IT (http://glottoweb.org/web2corpus/)
Web2Corpus_IT: features

CMC typologies
- Interactivity and dialogue (web 2.0)
- Public environments
- Only written communication

Corpus balance
- Web2Corpus_IT is a small pilot corpus (1 million occurrences)
  - Balanced by genres (blog, forum, newsgroup, chat, social network)
  - Document annotation and processing fully checked manually
  - Make qualitative analysis with different tools possible (not only online querying)
- Manual collection (not opted for crawling or focused crawling)
  - Strict selection criteria to allow linguistic research on known data

Other existing CMC corpora of Italian language
- ItWac (Baroni et alii 2009) – web (all)
- NUNC (Newsgroups UseNet Corpora) – Univ. Turin
- Paisà (Borghetti et alii 2012) – blog and wiki
Selection criteria for data collection

- **Maximum variation in**
  - Contexts, registers, participants, domains for each subcorpus
    - E.g. for newsgroups thread from all different subject areas were collected

- **Maximum accessibility**
  - Selection prefers most accessed and popular websites using public lists of access.

- **Document integrity**
  - Full thread (from first post to up to collection day)

- **Size (medium size threads)**
  - Based on the average length of each source
  - For chat sessions were recorded for 10 minutes each on public activities

- **Layout**
  - **Boilerplate** (macro) exclusion
  - Layout preservation with graphics (odt)
Web2Corpus_IT composition
Workflow (sorry it’s in Italian!)

- blog
- forum
- newsgroup
- chat
- social network
- anonimizzazione manuale
- testi anonimizzati
- anonimizzazione automatica
- DTD CMC
- Annotazione XML manuale
- Annotazione XML automatica
- Annotazione linguistica documento
- Annotazione XML documento
- metadati
Anonymising users

**CHAT example**

- senderA: senderB guà che ti farei sorridere di più di quel freddo telecomando |004
- [server] * il topic può essere modificato solo da un operatore
- [server] * gli utenti non presenti nel canale non possono inviare messaggi
- senderA: senderD |001)
- senderB: |016
- senderB: mi provoca!
- senderA: auhsuhsushushu
- senderC: senderA ma senderB lo vedo molto preso
- senderD: ciao ciao senderA^ 
- senderA: no senderB, ti faccio il solletico |010

Same procedure for **URL's** (if referring to personal pages) and **e-mails**

**AFTER XML ANNOTATION**

---

**Anonymisation table**

<table>
<thead>
<tr>
<th>FILE</th>
<th>USERNAME</th>
<th>ANON.CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>b_chat_1.txt</td>
<td>EmY</td>
<td>senderA</td>
</tr>
<tr>
<td>b_chat_1.txt</td>
<td>Lilly</td>
<td>senderAA</td>
</tr>
<tr>
<td>b_chat_1.txt</td>
<td>^CATGUT^</td>
<td>senderAB</td>
</tr>
<tr>
<td>b_chat_1.txt</td>
<td>MassY^</td>
<td>senderAC</td>
</tr>
<tr>
<td>b_chat_1.txt</td>
<td>Mattw27</td>
<td>senderAD</td>
</tr>
<tr>
<td>b_chat_1.txt</td>
<td>tatuatore</td>
<td>senderAE</td>
</tr>
<tr>
<td>b_chat_1.txt</td>
<td>ehi</td>
<td>senderAF</td>
</tr>
<tr>
<td>b_chat_1.txt</td>
<td>jalo</td>
<td>senderAG</td>
</tr>
<tr>
<td>b_chat_1.txt</td>
<td>j0ker</td>
<td>senderB</td>
</tr>
<tr>
<td>b_chat_1.txt</td>
<td>so`Rakkia</td>
<td>senderC</td>
</tr>
<tr>
<td>b_chat_1.txt</td>
<td>henryyy</td>
<td>senderD</td>
</tr>
</tbody>
</table>

**MANUAL ANONYMISATION**
- Blog
- Forum

**AUTOMATIC ANONYMISATION**
- Chat
- Newsgroup
- Social Network

> fully checked manually for identities slipped..
Corpus annotation

Metadata (header):
typology, website name, URL,
length, post number, number of participants,
keywords, extra annotation

Document DTD
-very simple dtd! (bottom-up)

• GUI Annotation Tool: WebAnnotator
  • http://perso.limsi.fr/Individu/xtannier/en/WebAnnotator
• WA code cleaning and HTML formatting cleaning
Il testo

- **body**: testi autori.
- **gap**: elementi non inseriti dall'autore ma dal sistema
  - *webaction* (pulsanti ed elementi interattivi—invia, modifica, rispondi, etc.)
  - *other* (elementi fissi: inviato da, data, XX ha lasciato la conversazione)
  - *dynamic other* (elementi dinamici derivati da altre fonti (es. profili forum: luogo, età, stato, etc.).

- **quote**:
Corpus linguistic processing (POS tagging)

1. Corpus pre-processing
   - Emoticon conversion and standardisation (not to interfere with pos tagging)
     - Emoticon standardisation (e.g. :-) , :), :o) :
     - Conversion into 'dummy words' (table)
     - Conversion of graphic emoticons into dummy
       - e.g. :-) > *smile*

2. Preliminary POS tagging (TreeTagger - Stuttgart)
   - table of misprints, unknown words, etc.
     - Lexicon (for second POS tagging phase)
       - E.g.
         - FORM POS LEMMA
         - emotcon N emoticon
         - intrepreta V interpretare
Corpus linguistic processing (POS tagging) –cont.

• 3. POS tagging (TT) with additional lexicon
  – (of XML annotated corpus)
• 4. Full manual post-correction
  – with tool Analog (Marie-Hélène et François Lay, Univ. Poitiers)
• 5. Indexing and online querying with CQPweb
Objectives and requirements

• **Research objectives** (on corpus processing):
  – Identify external parameters of conversation (in metadata)
  – Define with a bottom-up procedure prototypical genres
  – Keep separate genre elements (external, formal, informative) from linguistic properties of text

• **Requirements** (for corpus use and querying)
  – Combining metadata, textual criteria (from XML annotation) and linguistic (POS tagging) features in online querying
  – Filter searches on specific subcorpora or to specific textual properties (titles, replies, quotes)
  – Allow querying not only on concordances but on entire text fragments identified for example as posts, or sequences of posts
  – In order to make conversation analysis procedures possible not only in the downloadable texts but also in the online querying system
  – Allow query results export in different formats

• **Release of corpus versions**
  – Raw anonymized odt (with original formatting)
  – Annotated XML
  – Online (and offline) lemmatized corpus (also useful for further training)
provisional website
http://glottoweb.org/web2corpus/

isabella.chiari@uniroma1.it

THANK YOU