Silvia Petri and Mirko Tavosanis

Building a Corpus of Italian Web Forums: Standard Encoding Issues and Linguistic Features

Abstract

This paper describes the creation of a reference corpus of nearly 1200 Web forum posts in Italian. The corpus was created evaluating and customizing a previous proposal for Xml standard encoding; a revised version of the relevant DTD is now proposed as reference for the structural features of Web forum posts and a set of correspondences, with little loss of information, is given for the TEI P5 encoding system. Preliminary results about syntactic features of the language of the posts are also included to sample the linguistic variability of this textual genre.

1 Overview

Web forums are, arguably, the most popular interactive textual genre on the web. Current Eurostat surveys show that up to 50% of the citizens of some states of the European Union have posted at least a Web message in the year preceding the interview: this posting is undoubtedly often a *forum* posting. However, few studies (such as Light and Rogers 1999) have dealt with the linguistic or textual features of the forums or even with more basic facts such as their diffusion.

Moreover, there is widespread variety even in the name of the genre. Describing Web forums, secondary literature calls them *message boards*, *discussion boards*, *discussion groups*, *conversations*, *chatgroups*, *newsgroups* and so on (as for classification issues for Web texts, see also Rehm et al. 2008). This variation also seems to push towards the grouping in the literature of many textual genres we find scarcely related from every point of view: Usenet newsgroups of the 1990s lack many of the features now common in Web forum interfaces, and so on. Typical of this attitude are syntheses such as Crystal (2006, pp. 134-177) where a single chapter devoted to “The language of chatgroups” describes both “asynchronous” and “synchronous groups”, with a few lines of specification.

In this paper we will instead deal with a very specific textual genre: asynchronous conversations collected in threads and managed by a particular kind of Web site. Many Web forums allow more than a single way of interaction and messages can be posted on them by various means (Web interface, e-mail and so on). We will see in §§ 4-5 that this seems to have linguistic consequences more relevant than those connected to, say, the topic of the forum, hinting to the need to make subtler distinction between subgenres.

We did not take into account, then, traditional newsgroups or mailing lists if they don’t have a web interface allowing users not only to read past conversations but also to write new messages. On the other hand, we did consider as possible sample material every kind of forum where:

- messages are archived on the Web in antichronological order, and,
- a Web interface allows users to compose posts on-line and to publish them on the system.

Our description of Web forums (or simply *forums*, in the following sections of the paper) will then be relative to this kind of publication. We feel that the genre features of Web forums are both specific enough to be described on their own a#nd varied enough to require in-depth discussion. In order to verify this feeling, we then decided to create a reference corpus of Italian forums.

2 Selection of Forums

Forums can undoubtedly be dedicated to different topics and can have very different features and very different language choices. Moreover, their features seem only partially standardized. In order to sample this variety, we choose to start the corpus construction with four Italian forums of different character and managed with different software tools, including the two packages which together seem to cover more than half of the market, phpBB and vBulletin:

Accademia della Crusca Formal discussions about Italian language (forum closed in 2005 and managed with phpBB software) (*http://forum.accademiadellacrusca.it/phpBB2/index.php*)

ADI – Associazione dei dottorandi e dottori di ricerca italiani Discussions about work issues of Ph.D,. students and Ph. D. holders (forum managed with phpBB software; messages can be sent to the forum both through a Web interface and a mailing address and are distributed both through the Web and a mailing list) (*http://www.dottorato.it/forum/index.php*)

HTML.it Technical forums about HTML and related languages (forum managed with house-developed software) (*http://forum.html.it/forum*)

NGI Social forum, gathering players of an online role-playing game (forum managed with vBulletin software) (*http://gaming.ngi.it/forum/forumdisplay.php?f=501*)

This selection is absolutely arbitrary: it is useful to stress that to our knowledge there are absolutely no reliable data about the distribution of forums to use as a starting point to build a "representative" corpus (see also Leech 2007 for a critical appraisal of the current corpus "representativeness"). Nor is the sheer number of forums known, even if there are many hundreds (or thousands) of them in the Italian Web alone. Random searches allow to find Italian forum sites with as many as 509,962 registered users, 21,838,712 messages and 1,374,969 threads each. As for the world scene, the BoardTracker site boasts the inclusion in its database of nearly 38,000 forums and of 59 millions of threads (Petri 2008, 6-8, 13), but actual figures could be much higher. The Gaia Online forum alone declares on its home page (19.9.2008) to host "**1,412,466,251** articles posted with **14,646,253** registered users". Anyway, at the moment there are also no reliable estimates about the most popular forum topics (games? general discussions?) or the most used languages in forum posts.

Forums 1 and 2 were then chosen, according to previous knowledge, as potential samples of a more formal language and style of communication; forums 3 and 4 as potential samples of less formal language. On the whole, we gathered from the four forums a grand total of 1186 posts (the size limit being imposed mainly by work constraints: see § 3), with different criteria.

From the first two forums were randomly selected four threads which were composed by 25 posts or more. The posts were encoded integrally (see § 2 for mark-up description) and in the original order. The thread encoding proceeds then from the newest to the oldest post.

Instead, from the HTML.it and NGI forums has been extracted a bigger sample: 50 threads of 10 posts each. This choice was based upon the hypothesis that these forums would have represented better the less formal language levels which seem more common on the Web even if, given the status of knowledge about this textual genre, there is no hard evidence to support this idea, which relies only on common sense (and can be disproved by future research).

Also in this case, the procedure of selection began from the first page of the forum site and has gone on turning back in the chronology of the forum to find new material: since this work required some weeks, every day the selection began again from the first page and from the "most updated" threads it hosted. For these forums too the threads were chosen according to the number of replies they included, discarding threads with less than 10 replies.

In the revision phase, a few of the posts were then discarded due to mistakes in selection. At the end of the work, the encoded posts were then 1186 instead of 1200.

As expected, interface features and style of writing seem to vary according to the forum. The selected posts of the forum of the Accademia della Crusca often offer high-handed discussions of language in a very formal tone. The NGI forum posts, instead, sport a variety of features of neostandard Italian (Berruto 1987). More precise descriptions, however, required the establishment of a frame of reference.

3 The Mark-up

Italian forums, as well as international ones, apparently are not gathered by standard sites hosting thousands of them, as is the case of blogs (Tavosanis 2007). Forums, instead, seem often created as additional features in many sites, and, even if the two products recalled in § 2 seem dominant, this creation is mediated by a wide variety of software (Wikipedia 2008). A widespread Html reference framework does not exist, and the source code of the forum pages follows no standard. Those features of forums pose of course many limits to the performance of search engines and crawling systems (Limanto et al. 2005, 978). In order to compare different forums we have then to reduce them to a common encoding for further processing.

This work has tried to verify the conditions for a standard Xml encoding. Its starting point has then been the standard proposed by Claudia Claridge (2007, p. 94-97). This standard was then modified according to the actual features of the forums encoded, as they emerged during this research.

3.1 Starting Point

To encode and interchange materials, an Xml standard of mark-up was proposed by Claridge (2007); we will refer from this point to this model simply as “Claridge”. It includes the following elements and attributes:

```
<forum> (root element)
<thread>
<message> with attributes “topic”, “no” and “ad”
<person> with attributes “gender” and “desc”
<mbinfo> (message board-related information) with attributes “joined”,
“posts”, “avday”, “greats” and “warnings”
<place> with attribute “desc”
<time>
<sig> (marks an automatic attachment to any message a given writer
sends; also called signature)
<body> (the content of the message)
<p> (the paragraphs in the body)
<quote> (marks the quotations)
<visual> (marks the presence of “graphic elements with emotive/interac-
tive meaning” (Claridge 2007, p.97)) with attribute “meaning”
<gap> (marks “purely additional, decora-
tive material” (Claridge 2007, p.97))
```

This structure was applied to the Html of the threads selected according to the procedure described in § 2. Of course, this required a heavy work of cleaning, since the typical Html code of an average post looks in this way (the example is drawn from the third selected thread of the HTML.it forum):

```
<TD style=”MIN-HEIGHT: 280px; WIDTH: 200px” vAlign=top bgColor=#f0f0f0
wrap><A name=post11412788></A><SPAN class=big><B>Myaku</B></
SPAN><BR><SPAN class=little>Utente di HTML.it</SPAN><BR><BR><IMG
alt=”” src=”Thread3 - Iframe_file/hetfield2.jpg” border=0><BR><BR><SPAN
class=little><P class=postbit_dati>Registrato il: Nov 2006</P><P
class=postbit_dati>Provenienza: </P><P class=postbit_dati>Messaggi:
2850</P><BR><P class=postbit_dati>ICQ : </P><P class=postbit_
dati>MSN : chiedere in privato</P><P class=postbit_dati>Skype : <A
```

```
href=”skype:chiedere%20in%20privato?chat”>chiedere in privato</
A></P></SPAN><BR></TD><TD style=”MIN-HEIGHT: 280px; WIDTH: 712px”
vAlign=top bgColor=#f0f0f0><SPAN class=little><B>Re: Iframe</B></
SPAN> <DIV class=corpopost onfocus=this.blur();><DIV class=head_
citazione>Citazione:</DIV> <DIV class=citazione><STRONG>Originariamen
te inviato da mayorca </STRONG><BR>Ho utilizzato Dreamweaver CS3, non
vorrei che avesse inserito qualche tag che mi blocca le pagine.<BR></
DIV><BR><BR>se non ci fai vedere il codice, mi sa che anche a tir-
are a indovinare la vedo dura <IMG alt=”” src=”Thread3 - Iframe_
file/smile.gif” border=0><BR><BR>ps. consiglio: evita i frames/
iframes se puoi<BR><BR><IMG alt=”” src=”Thread3 - Iframe_file/ciao.
gif” border=0></DIV><P><SPAN class=norm><BR>__________________<BR><FO
NT style=”LINE-HEIGHT: 150%” size=1><STRONG><A href=”http://www.sysu-
niverse.net/” target=_blank>grafica &amp; web</A> | <A href=”http://
www.thinksy.altervista.org/” target=_blank>blog</A> -- no pvt tecni-
ci -- </STRONG><BR>AVVISO: l’aiuto è gratuito, la “pappa pronta” si for-
nisce solo su preventivo.<BR><FONT style=”COLOR: red”>...USATE IL TAG
# quando vi chiedo di postare il codice!!!!</FONT></FONT></SPAN></P>
```

The differences in encoding made unavoidable to clean and to re-encode the corpus largely by hand. We are aware of current automatic cleaning practices such as those documented in the CleanEval competition, but these kind of approaches encountered many difficulties and we judged that manual encoding was the only way to have satisfactory results in the time frame available for the work. At the end of the process, it became evident that it was in fact possible to encode the posts using the Claridge structure, but also that some features could be represented in a more satisfactory way by making a few changes to the model.

3.2 Changes to the Original Structure and Final Mark-up

In our revision, the XML tree remains unchanged, with the root element <forum> that contains some elements <thread>, composed by elements <message>.

The main element <forum> has the new attributes “name” (the name of the corresponding forum), “section” (the section analyzed), “url” (including the web address of the forum), and “software” (indicating the software used for the creation and management of the forum).

The element <thread> has a “cod” attribte (to identify the thread univocally with two letters of the alphabet), and a “topic” with the argument of that thread.

The <message> element has now the following attributes:

- “title” (corresponding to “topic” in the original model)
- a progressive number (“no”)
- an indication allowing to correlate directly the message with another of which it includes quotations (“ad”).

In every message we encoded information about the writer: as in the Claridge mark-up, we have the <person> element, but we kept only the attribute "desc", a description of the rank reached from that sender in the forum, or (like in the NGI forum) a phrase or a title inserted from the sender to describe him- or herself, while the foreseen attribute "gender" has never found explicit correspondence, either in the forum or in the user profile.

Other data about the writer are fitted in the element <mbinfo> that includes the attributes "joined", "posts" (as in Claridge), "place" (that was an element by itself in the original mark-up but that has been inserted included here because it refers to the sender and not to the message), and other particular attributes such as "ICQ", "MSN" and "Skype".

In the corpus it has been found no information to instantiate the attributes "avday" ("the average posting rate per day" (Claridge, p. 96), present only in each user profile and of small interest for this kind of work), "greats" ("an evaluative category of "great" message" (Claridge, p. 96)) and "warnings" ("the number of warnings a sender has received for violating netiquette or forum rules" (Claridge, p. 96)), which were then dropped.

Every message contains the element <time> (as in Claridge) that indicates the temporal coordinates of the message writing, with the attribute "edited", in the case of a later editing by the writer; in two cases there was an indication of the reason of the editing, marked in the attribute "reason".

For the central part of the message, the element <body>, the original HTML structure has been kept, with <div>, <span>, <br/>, <i>, <b> etc. The style element indicating the font color has been marked with the element <color>, while the HTML <a href> has been changed in the Xml <link url> in order to be more understandable. A new element, named <object>, has been introduced to mark external objects in the message, such as video clips.

It has also been added an element <s> that is useful to the scope of the research allowing to encode the sentence boundaries of the text.

The element <quote> is kept as in the original encoding. The frequent element <sig> has been separated from the <body> of the message.

Regarding the encoding of graphic elements included in the message, this work follows Claridge's indications: the presence of images is marked with element <gap> while an emoticon is described in the element <visual> with attribute "desc" to encode its name and meaning.

At the end of the work, the post seen in § 3.1 is then encoded in this way:

```
<message title="Re: Iframe" no="02" ad="01">
<person desc="Utente di HTML.it">Myaku<gap/></person>
<mbinfo joined="Nov 2006" posts="2850" MSN="chiedere in privato"
Skype="chiedere in privato"/>
<time>18-02-2008 12:33</time>
<body>
<div class="corpopost"><quote>Citazione:<b>Originariamente invia-
to da mayorca </b><br />Ho utilizzato Dreamweaver CS3, non vor-
```

```
rei che avesse inserito qualche tag che mi blocca le pagine.<br /></
quote><br /><br /><s>se non ci fai vedere il codice, mi sa che anche
a tirare a indovinare la vedo dura <visual desc=”Smile”/></s><br /><br
/><s>ps. consiglio: evita i frames/iframes se puoi</s><br /><br /><vi-
sual desc=”ciao”/></div><br/>__________________<br/></body><sig><b><link
url=”http://www.sysuniverse.net”>grafica &amp; web</link> | <link
url=”http://www.thinksy.altervista.org”>blog</link> -- no pvt tecnici --
</b><br />
: l’aiuto è gratuito, la “pappa pronta” si fornisce solo su
preventivo.<br />
<color n=”red”>...USATE IL TAG # quando vi chiedo di postare il codi-
ce!!!!</color>
</sig>
</message>
```

The final version of the encoding was then archived for future work and published in our web pages. The current DTD is published as Appendix at the end of this paper.

3.3 TEI P5 Encoding

In order to make easier to reuse the corpus, we established also a set of correspondences of our Xml elements and attributes with TEI P5 elements and attributes. A few attributes considered of little use to researchers were deleted, while the new elements <byline> and <closer> were enclosed. This is the final list of correspondences:

Claridge revised	TEI P5
<forum> attributes: the values of “name”, “section”, “url” and “software” are enclosed as text in the <title> element (child of <sourceDesc>)	<teiCorpus>, enclosing <TEI> and <teiHeader>
<thread> attributes: “cod” values become values of “xml:id” of <div1>; “topic” values are enclosed as text in <head> (child of <div1>)	<div1 type=”thread”>
<message> attributes: “no” values become values of “n” of <div2>; “ad” values are suppressed; “title” values are enclosed as text in <head> (child of <div2>)	<div2 type=”message”>
<person> attributes: “desc” values are suppressed	<persName>

<mbinfo> attributes: “place” values are enclosed as text in <placeName>; “joined” values are enclosed as date in <date type=”jointime”>; “posts” values are enclosed as text in <num type=”msgnumber”>; attributes “MSN”, “ICQ” and “Skype” are suppressed	
<time> attributes: “edited” and “reason” values are suppressed	<time>
<body>	<p>
<s>	<s>
<div>	suppressed
<span>	suppressed
<td>	suppressed
<quote>	<seg type=”quotation”> enclosing <quote>
<sig>	<closer> enclosing <signed>
<link> attributes: “url” values become values of “target” of <ref>	<ref>
<u>	<emph rend=”underline”>
<i>	<emph rend=”italic”>
<b>	<emph rend=”bold”>
<color> attributes: “n” values are suppressed	<emph rend=”color”>
<br>	<lb>
<object> attributes: “desc” values are suppressed	<binaryObject>
<gap>	<graphic>
<visual> attributes: “desc” values are enclosed as text in <figDesc> (child of <figure>)	<seg type=“iconic“> enclosing <figure>

As for the relevant Schema, it was generated through the online tool Roma (http://www.tei-c.org/Roma/) adding to the standard set of TEI modules (Core, Tei, Header, Textstructure) the following optional modules:

- Analysis
- Corpus
- Namesdates
- Figures
- Linking

The corpus was then transformed using an XSL-T stylesheet. The output was a valid TEI P5 file. In this version, the post seen in § 3.1 is encoded in this way:

```
<div2 type=”message” n=”02”>         <head>Re: Iframe</head>       <by-
line>
        <persName>Myaku</persName>
        <time>18-02-2008 12:33</time>
        <placeName></placeName>
        <date type=”jointime”>Nov 2006</date>
        <num type=”msgnumber”>2850</num>
        </byline>
        <p><seg type=”quotation”>
        <quote>Citazione:<emph rend=”bold”>Originariamente inviato da
mayorca </emph><lb/>
Ho utilizzato Dreamweaver CS3, non vorrei che avesse inserito qual-
che tag che mi blocca le pagine.<lb/></quote></seg><lb/><lb/><s>se non
ci fai vedere il codice, mi sa che anche a tirare a indovinare la vedo
dura <seg type=”iconic”> <figure> <figDesc>Smile</figDesc> </figure></
seg></s><lb/><lb/><s>ps. consiglio: evita i frames/iframes se puoi</
s><lb/><lb/><s><seg type=”iconic”><figure>
<figDesc>ciao</figDesc></figure></seg></s><lb/>__________________<lb/></
p>
        <closer><signed><emph rend=”bold”><ref target=”http://www.
sysuniverse.net”>grafica &amp; web</ref> | <ref target=”http://
www.thinksy.altervista.org”>blog</ref> -- no pvt tecnici -- </
emph><lb/>AVVISO: l’aiuto è gratuito, la “pappa pronta” si fornisce solo
su preventivo.<lb/><emph rend=”color”>...USATE IL TAG # quando vi chiedo
di postare il codice!!!!</emph></signed> </closer>
<div2>
```

4 Linguistic Data Extraction

In addition to the validation of the structural elements, we used the corpus to evaluate some linguistic features of the texts. The starting numbers are:

- Number of posts: 1186
- Number of word forms: 150115
- Average number of word forms per message: 125

The number of words per message seems of particular interest:

Forum	*Words / message ratio*
Accademia della Crusca	111
ADI	470
HTML.it	108
NGI	75
Total	125

A second step was the encoding and evaluation of sentences. Many of the messages included non-standard constructions, especially from the punctuation point of view. In particular, ellipses (three or more full stops) and line breaks are used both for in-sentence pauses and as end-of-sentence marks. A sample of this can be seen in the following section, where the HTML tag <br/> is used to mark the limits of two sentences:

```
<s>evita JS finché puoi</s><br />
<s>evita frame e iframe finché puoi</s><br />
<s>usa le inclusioni asp o php per le parti co-
muni (testata, menu, footer...)</s>
```

This situation makes strict human supervision mandatory in order to have reliable sentence counts. At the end of the process, then, the grand total for the whole corpus was:

Sentences: 3435
Average number of sentences per message: 2.9

Those numbers can now be compared with those given by the few relevant studies for other electronic text types. Moreover, we can start clarifying the differences between different textual subgenres.

The average number of sentences in a message, in fact, seems more or less stable between the different forums:

Forum	*Sentence / message ratio*
Accademia della Crusca	3.1
ADI	4.9
HTML.it	2.9
NGI	2.4

The same is true for the word / sentence ratio (which is also surprisingly high for Italian prose standards):

Forum	Word / sentence ratio
Accademia della Crusca	35
ADI	100
HTML.it	37
NGI	31
Total	43

From the linguistic and textual point of view it is a little surprising to note that highformality forums such as those of the Accademia della Crusca are only slightly more articulated from the syntactic point of view than supposedly low-key forums such as those dealing with on-line gaming.

5 Use of Emoticons

On a different level, we investigated also the role of emoticons. Even in this case we found a surprisingly stable distribution:

Forum	% of emoticons to messages
Accademia della Crusca	48
ADI	13
HTML.it	48
NGI	46
Total	44

About the emoticons, we found out some characteristics that could indicate a first categorization of their functions and use:

1) Can be substituted by adverbs (13 cases)
- example from HTML.it: "<s> Allora ho provato ad usare il metodo get al posto di post… funziona <visual desc="Mmmm… strano… molto strano"/> </s>"
2) Can be substituted by verbal expressions (3 cases)
- example from NGI: "<s> avuti 3 rogue, 1 prete e ora ne sto rollando n'altro <visual desc="Love"/> </s>"
3) Emphasize graphically the meaning of the words surrounding them (143 cases)
- example from Accademia della Crusca: "<s> Allora continuiamo a indagare <visual desc="Rolling Eyes"/> </s>"
4) Express the attitude of the writer (270 cases)
- example from ADI: "<s> A tuo piacimento guarda…. tanto l'effetto non cambia, ahimè sono andata <br/> <visual desc="Smile"/> </s>". This is the original function of emoticons but, evidently, not the only one.
5) Can be substituted by imprecations (23 cases)
- example from HTML.it: "<s> Ho appena sentito il cliente, gli ho fatto fare dei test, non funziona <visual desc="Mannaggia li pescetti"/> </s>"
6) Form a phrase by themselves (82 cases)
- example from HTML.it: "<s> <visual desc="ciao"/> </s>"
7) Are used for their graphical appearance in the composition of a word (1 case)
- example from NGI: "<s> Eventualmente sono disponibile anche <visual desc=":v"/>ome riserva nel 3v3 o 5v5. </s>"

6 Data Evaluation

We suspect that the most conspicuous variation in our data, the highest complexity of the messages of the ADI forum and their relative lack of emoticons, is related to the nature of the writing medium. As anticipated in § 2, the ADI forum is the only source in our corpus including messages sent through a mailing list. The system does not allow to check if a message is composed with mail software or with forum interface; we suspect however that many messages are composed as e-mail messages and that the strong difference between the two mediums explains the difference in message and phrase length. Even the relative lack of emoticons could be easily explained in this way, since mail software usually does not display ready-to-use emoticons, while forum interfaces display them.

In order to sample the internal variety of forums, we also selected a single thread of the Accademia della Crusca forum. The thread ("Auguri") was completely dedicated to the exchange of well-wishing messages for Christmas and New Year's Eve and was then of a very different character front of linguistic discussion threads. The counts for this single thread are:

sentence / message ratio: 2.5
word / sentence ratio: 25
message length in words: 63
% of emoticons to messages: 68

The difference between those values and those of the Accademia della Crusca forum as a whole is noteworthy, and it is consistent with the general idea of how such a thread can be different from a scholar discussion. However, it is also useful to note that this difference is not particularly striking in quantitative terms.

7. Conclusions and future work

Even if the Xml encoding of forums for research purposes is rare, we feel that our work displays the feasibility and utility of such a practice. The corpus described is now available through our institutional web pages: we hope that the collected materials can be used as term of comparison for further analyses of forums, especially from the linguistic point of view.

Of particular interest seems the stability in phrase and message length across different topics and different social contexts. This hints to a dominant role of tool and genre, more than topic and situation, in the creation of Web texts. Further analyses could be able to better describe this situation.

Bibliography

Berruto, G. (1987). Sociolinguistica dell'italiano contemporaneo. Roma: Carocci.

Claridge, C. (2007). „Constructing a corpus from the web: message boards." In: Hundt, M., Nesselhauf, N. and Biewer, C. (eds.) (2007). Corpus Linguistics and the Web. Amsterdam; New York: Rodopi, 87-108.

Crystal, D. (2006). Language and the Internet. Second edition. Cambridge University Press: Cambridge (UK).

Leech, G. (2007). „New resources, or just better old ones? The Holy Grail of representativeness.” In: Hundt, M., Nesselhauf, N. and Biewer, C. (eds.) (2007). Corpus Linguistics and the Web. Amsterdam; New York: Rodopi, 133-150.

Light, A. and Rogers, Y. (1999). „Conversation as Publishing: the Role of News Forums on the Web.” In: Proceedings of the 32nd Hawaii International Conference on System Sciences – Journal of Computer-Mediated Communication, 4, 4.

Limanto, H. Y. et al. (2005) „An Information Extraction Engine for Web Discussion Forums.” In: International World Wide Web Conference archive. Special interest tracks and posters of the 14th international conference on World Wide Web table of contents. Chiba, Japan. New York: Association for Computing Machinery, 978-979.

Petri, S. (2008). „I forum italiani: analisi linguistica e problemi di codifica.” Master Degree thesis, Università di Pisa.

Rehm, G. et al. (2008). „Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems”. In: Proceedings of the LREC 2008.

Tavosanis, M. (2007). „Juvenile Netspeak and subgenre classification issues in Italian blogs.” In: Rehm, G. and Santini, M. (2007). Proceedings of the International Workshop Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing. Shoumen: Incoma, 37-44.

Wikipedia. (2008). Comparison of Internet forum software. Ad voc. (***http://www.wikipedia.org***).

Appendix: the full DTD

```
<?xml version=”1.0” encoding=”UTF-8”?>
<!ELEMENT forum (thread+)>
<!ATTLIST forum
                name CDATA #IMPLIED
                section CDATA #IMPLIED
                url CDATA #IMPLIED
 software CDATA #IMPLIED>
<!ELEMENT thread (message+)>
<!ATTLIST thread
                cod CDATA #IMPLIED
                topic CDATA #IMPLIED>
<!ELEMENT message (person, mbinfo,  time, body, sig?)>
<!ATTLIST message
                title CDATA #IMPLIED
                no NMTOKEN #IMPLIED
                ad NMTOKEN #IMPLIED>
<!ELEMENT person (#PCDATA | gap | link)*>
<!ATTLIST person
                desc CDATA #IMPLIED>
<!ELEMENT mbinfo EMPTY>
<!ATTLIST mbinfo
                 joined CDATA #IMPLIED
                 posts CDATA #IMPLIED
                 place CDATA #IMPLIED
                 bowling CDATA #IMPLIED
                 ICQ CDATA #IMPLIED
                 MSN CDATA #IMPLIED
                 Skype CDATA #IMPLIED>
<!ELEMENT time (#PCDATA)>
<!ATTLIST time
               edited CDATA #IMPLIED
               reason CDATA #IMPLIED>
<!ELEMENT body (#PCDATA | span | td | br | s | div | quote)*>
<!ELEMENT sig (#PCDATA | br | link | i | color | u | b | visual)*>
<!ELEMENT s (#PCDATA | color | br | i | gap | visu-
al | b | link | u | span | td | s |div | quote)*>
<!ELEMENT span (#PCDATA | b | i | u | br | s | vi-
sual | gap | color | link | quote)*>
<!ATTLIST span
                class CDATA #IMPLIED>
<!ELEMENT td (#PCDATA | br | i | visual | span | b | td | u)*>
<!ATTLIST td
                class CDATA #IMPLIED>
<!ELEMENT br EMPTY>
<!ELEMENT div (#PCDATA |  b | i | u | br | s | visu-
al | gap | link | div | span | object | quote)*>
<!ATTLIST div
               class CDATA #IMPLIED>
<!ELEMENT quote (#PCDATA | b | br | visual | link
| i | gap | span | quote | u | color)*>
<!ELEMENT link (#PCDATA | i | b | color | gap | u)*>
<!ATTLIST link
                url CDATA #IMPLIED>
<!ELEMENT i (#PCDATA | br | visual | color | u | b | link | div)*>
<!ELEMENT b (#PCDATA | link | color | br | i | u | visual)*>
<!ELEMENT u (#PCDATA | br | b)*>
<!ELEMENT color (#PCDATA | b | br | link | vi-
sual | color | u | i | quote)*>
<!ATTLIST color
                n CDATA #IMPLIED>
<!ELEMENT gap EMPTY>
<!ELEMENT visual (#PCDATA)>
<!ATTLIST visual
                desc CDATA #IMPLIED>
<!ELEMENT object EMPTY>
<!ATTLIST object
               desc CDATA #IMPLIED>
```