Linguistic features of Italian blogs: literary language

Mirko Tavosanis

Dipartimento di Studi italianistici
Via del Collegio Ricci 10
I-56126 Pisa PI Italy
`tavosanis@ital.unipi.it`

Abstract

Preliminary surveys show that the language of blogs is not restricted to the more informal levels of expression. Instead blogs may include many kinds of written language: from simple personal notes to literary prose or poetry. The paper presents a sample of Italian blogs and comments on the results of the search of literary forms in two Web corpora using search engine queries.

1 Introduction

Close scrutiny of e-mails has revealed the presence of many different kinds of style in this medium (Baron 2000: 250-2; Pistolesi 2003: 178-184 for Italian). The same appears to be true for blogs. It is therefore difficult to determine specific linguistic features of blogs. Even occasional surveys however show that blogs are not limited to "a language reminiscent of brief notes, spoken asides, or short letters, rather than of essays or newsprint". Such language plays an important role in blogs, but accounts for only a small part of them. Many individual blogs aim instead at a true "literary" status and have a correspondingly high standard for word selection. Therefore, the linguistic equilibrium of this medium could be higher than expected. The paper will try to describe the general linguistic features of Italian blogs by contrasting them mainly with the language of newspapers, giving appropriate quantitative data.

2 Preliminary qualitative analysis: a textual sample

As a reference sample of blogs we can take ten blogs hosted by the Italian blog publishing site Splinder.com (arguably the most popular site of its kind). The sample was chosen by selecting the most recent blog appearing in the site list of the "Ultimi blog aggiornati" ('Most recently updated blogs'), and by selecting the first page of the postings published by the blog itself in November 2005 (if it had at least two postings in November). The selection was made at different times on one given day (29 December 2005) Some features of the selected blogs are described in Table 1.

A post taken from one of the less formal blogs in the sample (*di ritorno da...*) shows many of the linguistic features commonly ascribed to this kind of writing:

giornata più tranquilla..sarà che sono a casa mia..a fare la mia vita..parlo appunto di mia perchè la vita parallela che sto facendo a milano non mi appartiene..quindi non posso dire che sia mia..che discorso complesso però ci stava dentro bene nel contesto..

ho programmato il mio capodanno..dopo due anni consecutivi in una baita in montagna quest'anno lascio l'italia..pronta per 4 gg in scozia con tre amiche..

non vedo l'ora..partirò il 30 di dicembre..aspetto quel giorno e intanto mi preparo per quattro esami all'università..

ah...il viaggio al prezzo di 45 euro di volo e 90 di ostello..ultra risparmio...

The text describes the planning of a holday trip in Scotland for New Year's Eve and personal feeliings.The post shows also many unprofessional graphic choices (no capital letters for proper nouns or at the beginning of a sentence, no spacing after punctuation marks, *perchè* instead of *perché*) and the frequent use of three (or,

wrongly, two) suspension points. The latter feature, also illustrated in Table 1, is considered one of the most common features of blogs and in this case it is surely used to give a feeling of "spoken language" to the text.

The language of the other blogs in the sample is, instead, very different. Suspension points are used in the blog *SoleLuna* in order to create high-pitched literary texts. This is a kind of lyrical description of a problematic relationship:

Cancelli i tuoi passi nell'ombra di te stesso e scompari e compari quando e come più ti aggrada... Ed io mi lascio prendere dai pensieri e mi lascio intorpidire dai ricordi... Mi rivesto di te.. di noi.. Ho freddo.. Cerco di scaldarmi con il ricordo di un amore... Non ti amo. Amo il ricordo di quello che eravamo... E mi sfuggono via dalla mente le sensazioni e scivolano via gli odori.. si sbiadiscono i sapori... Tutto diventa la sfumatura del proprio colore.. la parodia, la beffa... Cerco di palpare le immagini e faccio attenzione a non sgualcirle.. più di quanto non lo sia io...

This kind of lyric language is heavily based upon the use of literary forms ("ti aggrada"), complex rhetorical constructions, "-*d* eufonica" and so on.

Midway between these two extremes we can find blogs like *Incontrista*. The posts of this blog are written in a language echoing newspaper editorials and brilliant prose. Significantly, they make no use of suspension points, as in this sectione (where the author contrasts the average psychological differences between female bloggers and female subscribers of dating sites):

E' quindi un fatto ancora che le splinderine sono mediamente delle ragazze migliori delle meetiche proprio per questo motivo. Hanno capacità, caratteristiche e aspetti che a me piacciono, come la voglia di esprimersi, di raccontare, di scrivere, di comunicare, di leggere, di scegliere, di assumere posizioni critiche. Le splinderine fanno parte della fascia più esperta e innovativa degli utenti internet, quelli che ne fanno un utilizzo più consapevole, e che hanno una cultura più elevata della media. Le splinderine sono donne che hanno qualcosa da dire e vogliono compagni di alto livello con cui cercano il confronto serrato e accettano anche lo scontro.

It seems that none of those kinds of language could today claim to be the main model for blog writing. Free expression, literary prose and newspaper writing seem to co-exist without a clear dominant position. Of course, however, those impressions can be given substance only through extension of the field of search to a large set of blogs. This is partly made possible by the use of modern search engines.

3 Quantitative analysis of large corpora using search engines

Some linguistic features of blogs can now be probed and measured using search engines. For instance, preliminary searches show that from the orthographical point of view Italian blogs are much more correct than the average of the Italian web and that they are edited at least as well as online newspapers (Tavosanis in print b). Other indicators related to the use of "neo-standard" Italian forms (Berruto 1987) in the field of personal pronouns and demonstratives suggest a kinship between blogs and newspapers (Tavosanis in print a).

According to those searches, the main differences between blog posts and newspaper articles are not linked to writing accuracy or to different morphological choices. We can therefore assume as a working hypothesis that the main differences between blogs and newspapers in fact relate to lexicon and syntax.

The syntactic status of many blogs is probably well represented by the textual samples chosen above (widespread use of suspension points being the most conspicuous feature). However, close survey of this level can probably be obtained only through the encoding of a wide corpus with syntactic tagging.

The lexical features of blogs can instead be studied through simple search engine analysis (see again Tavosanis in print a and b for details of this method). Newspaper editing in Italy, enforced by a strong tradition and dedicated staff, excludes words considered too expressive (apart from those acknowledged by the same tradition: Bonomi 2002). Blogs, on the other hand, can include forms taken from every level of linguistic use. We can therefore expect that both literary and low forms are more used in blogs than in newspapers.

Two Web corpora were then selected: the web site of the newspaper *La Repubblica*, indexed and queried through the Google interface (= R), and the whole of the blogs indexed in the beta version of Blogsearch.google.com (= B). Of course, no exact data are available on the consistency of the two collections and the number of tokens indexed. The two corpora seem however roughly equal in size: the search of a common word like *questo* gives 427,000 occurrences in R

and 467,000 in B; the search of *quello* 209,000 (R) and 257,608 occurrences (B); the search of *lui* 118,000 (R) and 159,970 occurrences (B); and so on. Of course, since word frequency is strongly correlated with the style and topic of the texts (for the Italian situation see Bortolini 1971: XIV-XV; Voghera 1993), this assessment cannot be considered an exact estimate. It does however give a preliminary quantitative estimate.

The highest frequency of vulgar words in the B corpus is of course undisputed, since newspaper editing is a strong barrier against this kind of language, and it needs no particular demonstration, e.g., we can find 30,310 occurrences of the word *cazzo* in B against 278 in R, and so on.

It is more difficult to demonstrate the highest frequency of literary language, which in the Italian tradition has a wide and varied lexicon. The abundance of synonyms and dispersion of forms lead one to focus searches on large groups of "weak" words instead of a limited set of "strong" words.

Next the list of "literary" verbs beginning with the letters *b*, *e* and *v* in the De Mauro (2000) dictionary was selected for analysis. The chosen verbs were 31 (*b-*), 47 (*e-*) and 49 (*v-*). Many of them also had non-literary uses and/or coincided with other Italian words: therefore only the words without homographs were used for the search, where every meaning recorded in the dictionary was marked at least as "obsolete" (code OB), "literary" (LE) or "bureaucratic" (BU). This left 23 (*b-*), 28 (*e-*) and 21 (*v-*) verbs. The two corpora were then searched for the infinitive forms of the verbs. Many of them did not appear at all: *baiare*, *balbuzzire*, *ballonzare* (1 occurrence in a text written in the dialect of Naples), *basciare*, *benedicere* (2 occurrences in two texts written in the dialect of Naples), *biancicare*, *biastemiare*, *blasmare*, *bombire*, *botare*, *botarsi*, *bravare*, *buccinare*, *bulicare*, *ebere*, *ecclissare*, *educere*, *enfiare*, *enfiarsi*, *escomunicare*, *escuotere*, *escusare*, *esinanire*, *espedire*, *esseguire*, *esterminare*, *estollere*, *estollersi*, *estorre*, *estruere*, *estruare*, *esturbare*, *esurire*, *evellere*, *evenire*, *vagheggiarsi*, *vanare*, *vengiare*, *vengiarsi*, *verberare*, *verdicare*, *verdire*, *vernare*, *verzicare*, *vilificare* and *vincire*.

The search also revealed that a verb marked in the dictionary as "literary" was instead widely used in both corpora: *vigilare*. While other forms occurred at most 94 times, in the corpora there are 644 occurrences of *vigilare*, evenly balanced (332 in B, 312 in R). It therefore appears more correct to consider this verb as a "common"

word, without literary connotations, and to exclude it from further analysis.

In a second phase, many forms were excluded from counts since they resulted simple typos or broken forms of different words (e.g., many occurrences of *ventare* are in fact occurrences of widely used verbs like *diventare* or *inventare*, with incorrect spacing). Only words where the possibilities of misspellings seemed low were therefore included In the counts.

After this sifting, the forms represented in the corpus occurred as described in Table 2:

Form	Occurrences in Blogs	Occurrences in *La Repubblica*
basire	1	0
bastarsi	12	1
beare	9	2
biasmare	0	1
biondeggiare	0	2
biscazzare	1	0
bruire	1	0
bruttare	0	1
bugiare	1	0
elicere	0	1
ergere	24	9
esondare	6	4
esperire	56	21
esplicare	79	15
estimare	2	0
evoluire	2	0
vacare	4	0
vagolare	7	1
vanire	1	0
vaticinare	17	6
ventare	2	0
vigoreggiare	2	0
villaneggiare	1	0
volvere	2	0
volversi	0	1
Total	**230**	**65**

Table 2: occurrences of literary forms

It seems therefore that some Italian blogs have in fact a higher proportion of a random selection of literary words than Italian newspapers. Further searches should be able to confirm or refute this finding.

4 Conclusion

Preliminary analyses of Italian blogs seem to confute the simple equivalence "blogs = informal text". Clearly, both statistical tools and special monitoring software are needed to give this kind of search more focus and more depth. Future searches must also achieve a better understanding of the coverage of search engines and should be based upon different search engines. It would be useful, moreover, to identify and exploit other searchable indicators of the linguistic quality of a text. Anyway in future researches adequate space should be allowed for the assessment of the presence of literary features in many blogs.

Reference

[Baron 2000] Baron, N. *Alphabet to Email. How written English evolved and where it's heading*. London-New York, Routledge.

[Berruto 1987] Berruto, G. *Sociolinguistica dell'italiano contemporaneo*. La Nuova Italia, Firenze

[Bonomi 2002] Bonomi, I. *L'italiano giornalistico. Dall'inizio del '900 ai quotidiani on line*. Franco Cesati Editore, Firenze.

[Bortolini 1971] Bortolini, U., Tagliavini, C. and Zampolli, A. *Lessico di frequenza della lingua italiana contemporanea*. IBM Italia, Milano.

[De Mauro 2000] De Mauro, T. *Il dizionario della lingua italiana*. Paravia, Milano.

[Pistolesi 2004] Pistolesi, E. *Il parlar spedito*. Esedra, Padova.

[Tavosanis in print a] Tavosanis, M. *Traditional corpora and the Web as corpus: the Italian newspapers case study*. Presented at CLIN 2005, Amsterdam, 16 December 2005.

[Tavosanis in print b] Tavosanis, M. *Are blogs edited? A linguistic survey of Italian blogs using search engines*. To be presented at AAAI-CAAW 2006, Stanford, 27-29 March 2006.

[Voghera 1993] Voghera, M. "Le variabili testuali e pragmatiche". In T. De Mauro, F. Mancini, M. Vedovelli, M. Voghera, *Lessico di frequenza dell'italiano parlato*. Etaslibri, Milano: 32-38.

Title	Description	Suspension dots	Other graphical highlightings	Images	Quoting	Smileys	Literary language, poetry	Alleged sex of main poster
cavi & gu	Personal exchanges by a couple of fiancées	x	x			x		M+F
CominciareDall'Inizio	Notes on her work by a high-school Physics teacher		x	x	x	x		F
di ritorno da_	A female student in Milan gives vent to her feelings about the big city	x		x	x			F
Incontrista	Newspaper-quality reports by a dedicated user of dating sites in Italy			x	x	x (character sequences)		M
NuVoLaBiAnCa In MeZzO a NuVoLe gRiGie	Personal rantings, poetry, reading notes	x	x	x	x	x	x	F
Pillola Rossa	Political rantings from a leftist position	x		x	x	x		M
53r3n4	Personal posts: feelings, song lyrics	x		x	x	x		F
se_lo_dici_è_vero	Collection of quotations	x	x	x	x	x		F
Servo del Signore	Prayers, thoughts about religion		x	x	x	x		M
SoleLuna	Poetry, recipes	x	x	x	x		x	F
VolObliquo	Poetry, both quoted and original		x	x	x		x	F