# Are blogs edited? A linguistic survey of Italian blogs using search engines

**Mirko Tavosanis**

Pisa University
Dipartimento di Studi Italianistici
Via del Collegio Ricci 10
I-56126 Pisa - Italy
phone: +39 050 2215065
email: tavosanis@ital.unipi.it

**Abstract**

Many blogs are written by people with no formal training in public writing; this could suggest a low level of editing and general correctness. A quantitative analysis of misspellings, however, shows that in their orthography Italian blogs are as at least as well revised as conventional Italian newspaper texts. On the other hand, their editing is more careful than the editing of the average of Italian web pages.

## Context: an empirical grid

The nature of the texts published on the Web is still poorly described from the linguistic viewpoint. References to the "informal" nature of all texts written for the Web can often still be found. This kind of view has been confuted in the past (see in particular Crystal 2001) but even in recent years descriptions of the linguistic traits of real Web writing are scarce, even for blogs.

In this survey we will take as a starting point for an orthographical analysis an empirical grid for the description of Web texts from the linguistic point of view presented in Tavosanis (2005). The grid is intended mainly as a compendium of words to enable linguists to speak more correctly about the texts published on the Web and to place particular phenomena in context. The whole classification, slightly different from current grids, is currently being tested on Italian Web pages; however, parts of it might also have wider implications and may be applied to other languages.

In the grid, four layers of description are directly related to the writing process:

1. time allowed for writing
2. writing tool
3. writing support
4. creation of writing

Two of them are related to the writer's attitude before or during writing:

5. text type
6. intended reader of the text

The first of these layers, "Time allowed for writing", is constructed upon four main categories, related to specific types of texts:

- fast unedited writing (forum postings and, occasionally, Web sites); includes text written without planning and without a second reading and/or correction.
- fast revised writing (forum postings and, occasionally, Web sites); includes text written with some degree of planning and/correction.
- conventional revised writing (Web sites and, occasionally, forum postings); includes text written within a process of planning and correction.
- writing designed for other kinds of publishing (Web sites and, occasionally, forum postings); includes text written for other media mechanically copied and published on the Web.

What kind of place can be allocated to blogs in this classification? A preliminary answer to this question will be provided in the Conclusion.

## Method of inquiry

The tendency of a single blogger is, of course, strictly personal and individual blogs will undoubtedly show huge variations in linguistic solutions. We can however try to individuate the slant of the textual genre by looking at a great mass of data. The study of entire blog sites, such as Blog.excite.it, is the simplest way of doing this.

Unedited writing should become apparent at many different levels: semantics, syntax and so forth. It is in the lexicon, however, that bad editing should be most evident. The frequency of writing errors such as misspellings is probably a good index of unedited writing.

We will therefore try to determine the correctness of blogs by measuring the proportion of correct and wrong forms of challenging Italian words. This ratio will be measured in three different situations: the whole Web, blog sites and newspaper sites. We must also remember that a text published on the Web may have been written on a more sophisticated instrument, such as a word processor with orthographical correction (third layer: writing support). This kind of tool would easily reduce the number of writing errors. This should happen often in newspaper texts but only occasionally in blogs.

As for the errors themselves, the most widely used Italian dictionary, "Zingarelli" (Zingarelli 2005), for a specific entry lists 103 "frequent" errors in written and spoken Italian. Most of these are errors of pronunciation and are irrelevant to a discussion of written language. Many others are errors in graphical accent and in a blog they might be difficult to avoid due to writing tool problems. The final aspect of an Italian text may be influenced by the use of particular tools, as discussed in the second layer of the above-mentioned empirical grid:

- keyboard or interface not suitable for the task
- standard keyboard or interface, suitable for the task
- professional tools

In the case of Italian orthography, Web publishing often encounters problems not just with text formatting (italics, bold etc.), as for other languages, but also with the correct representation of accented letters. Unwanted substitutions of characters are frequent: a writer may type an orthographically correct text only to discover that the publishing system being used cannot handle accented letters or text formatting.

The Italian forum *La meglio gioventù* published in 2004 on the Web site of the newspaper *La repubblica* provides many examples of this (in particular text formatting). Many Italians living abroad have taken part in the forum. Many orthographic errors can therefore be explained by the use of non-Italian keyboards (e.g. keyboards without accented characters) and not by the writers' lack of orthographic competence. The following is a quotation from a posting from England, where accented letters are replaced by the sequence letter + apex:

*Non ho potuto vedere il film perche' non ho accesso ai canali RAI in questi giorni e la*
*cosa mi rattrista molto.Penso che la mia meglio gioventu' sia legata al momento in cui*
*ho cominciato a decidere da sola. (...) E le lacrime l'ultimo giorno di campeggio,*
*perche' quella spensieratezza avrebbe dovuto aspettare un altro anno.*

For unequivocal use on the Web, excluding words using accents and apostrophes, the Zingarelli list should therefore be reduced to 24 simple errors, as shown in Table 1. We have to reduce it further because the huge numbers involved make it impractical to distinguish homographs. We will therefore avoid discussing the errors *avvallo* for *avallo* and *Macchiavelli* for *Machiavelli*, since the regular word *avvallo* (from the verb *avvallare*) and a common surname *Macchiavelli* do exist in Italian.

We should also note that some of the "wrong" forms in this list are included by other lexicographers in standard Italian. The Italian dictionary of Tullio De Mauro (2000) acknowledges as regular forms *efficenza* ("bureaucratic use"), *interpetare* ("regional form in Tuscany"), *peronospora* ("variant") and *scorazzare* ("variant" ). Moreover, this short list of errors is aimed at an high-level writing. It does not include, for instance, writing errors in commonplace words, such as *propio* instead of *proprio*, and so on. It includes instead many words of a technical or literary nature.

Of course, some misspellings are likely to be made only by experienced writers: rare words such as *collutazione* or *collutorio* may only be used by people with a good knowledge of the Italian language. Others errors, in commonly used words, are typical of inexperienced writers: in our list, *eccezionale* or *scenza* are the most conspicuous examples of this.

## Searching for errors using search engines

Commercial search engines lack many features typical of corpus querying systems. But even their unsophisticated linguistic functions can still be exploited in a significant way (Kilgarriff and Grefenstette 2003:342; Calishain and Dornfest 2003; Maxwell 2004; Davis 2005). The trickiest problem, in this field, is probably the fact that, when huge numbers of occurrences are involved, the most efficient search engine, Google, provides only the approximate number of pages where a given token occurs, instead of the exact number of occurrences of the token. Due to this engine behavior, values and figures can only be compared with one another and cannot provide reliable absolute values.

The searches presented below were done using Google in October 2005. All of the searches were restricted to the pages in Italian (using the option of language restriction offered by the engine); the figures provided refer to the "number of pages" or to the approximate number of occurrences found by the search engine. Moreover, we will

not take into account conjugated forms or differences between singular and plural, masculine and feminine.

From the point of view of corpus extension, we should note that Google enables to search the "whole web" but does not allow restricted searches like "all forums", "all newspapers" and so on. Searches were then restricted to single sites through the "site" function of the search engine. We should also note that not all blogs or newspapers, nor the " deep web", can be accessed using search engines. For example, the popular Italian blogging site Digilander is not indexed by Google and the contents of its blogs cannot then be retrieved in this way. The selection of blog and newspaper sites below was then created by trial on the most popular Italian sites of their kind.

Blog sites:
Blog.excite.it
Clarence.com
Splinder.it (contents are partially repeated in the following)
Splinder.com

Newspaper sites:
Corriere.it (*Corriere della sera*)
Ilmattino.caltanet.it (*Il mattino*)
Repubblica.it (*La repubblica*)
Unita.it (*L'Unità*)

## Search Results

The search results are summarized in Table 1. Before commenting upon blogs and newspapers it is important to note two facts regarding the use of the different words included in the list. Looking at the results for the whole Web we can see that:

- The balance of wrong / correct forms is highly differentiated. In the case of *anedottico* / *aneddotico* the wrong form is twice as common as the correct one. In the case of *interpetare* / *interpretare* the wrong form accounts for only .01% of the use of the word.
- The frequency of words is also highly differentiated: as for correct forms, it ranges from 6.480.000 occurrences of *scienza* to 114 of *aneddotico*; as for wrong forms, it ranges from 488.000 of *Caltanisetta* to 234 of *aneddotico*.

Restricting the search to different kind of sites we can then see that blogs show a much lower percentage of errors than the entire Web. In fact, the percentages are often similar to those of professionally edited texts such as Web newspapers. Most surprisingly, the overall total of the list shows a *lower* percentage of errors in blogs than in newspaper sites. Both kinds of sites are also consistently more accurate in their spelling than the average of the Italian Web pages.

In general, we should also note that blogs in rare or technical words have a significantly higher percentage of errors than newspapers, e.g. *collutazione* or *collutorio*. On the other hand they have consistently lower percentages of errors in common words like *eccezionale* or *coscienza*. However, a 22-word vocabulary is too small to draw conclusions from this observation. Other common errors are in fact more frequent in blogs than in newspapers. Searching for the misspelled word *propio* for *proprio* on the same blog sites of the queries discussed above, we obtain an error percentage of only 0.59% (with a maximum of 1.83% and a minimum of 0.31%). Testing it with the pre-examined newspaper sites yields an error percentage of only 0.1%. The percentage of errors on the entire Web is on the contrary considerably higher than both figures (4%).

## Conclusion

The quantity of data and the quality of the sample seem sufficient to draw a preliminary conclusion: on the orthographic level, Italian blogs are edited with the same care as materials published in Italian newspapers. There are of course differences but on average a rough index such as the total percentages of errors taken from a limited list gives blogs and newspapers equal rank while distancing them from the less edited mass of Web texts.

## References


Calishain, Tara, and Dornfest, Rael. 2003. *Google Hacks*. Beijing, etc.: O'Reilly.

Crystal, D. 2001. *Language and the Internet*. Cambridge: Cambridge University Press.

Davis, H. 2005. *Building Research Tools With Google for Dummies*. Hoboken: Wiley Publishing.

De Mauro, T. 2000. *Il dizionario della lingua italiana*. Torino: Paravia.

Kilgarriff, A. and Grefenstette, G. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics* 29(3):333-347.

Maxwell, M. 2004. Resource Discovery for Low Density Languages: Internet Internet Search, abstract in the abstract book of ACH/ALLC 2004 - Goteborg University, Goteborg: 88-89.

Tavosanis, M. 2005. *Linguistic Variability of Web Italian: a Working Empirical Grid*. Forthcoming (included in the ACH/ALLC 2005 program – Victoria University).

*Lo Zingarelli 2005*. 2005. Bologna: Zanichelli.

**Table 1**

| **Zanichelli list** | | **Whole Web** | | | **Blog sites** | | | **Newspaper sites** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Wrong form** | **Correct form** | **W** | **C** | **%** | **W** | **C** | **%** | **W** | **C** | **%** |
| accellerare | accelerare | 53200 | 912000 | 5.83 | 402 | 2202 | 18.26 | 51 | 1257 | 4.06 |
| anedottico | aneddotico | 234 | 114 | 205.26 | 7 | 66 | 10.61 | 1 | 10 | 10 |
| appropiato | appropriato | 775 | 1290000 | 0.06 | 89 | 12585 | 0.71 | 4 | 680 | 0.59 |
| avvallo | avallo | Not inc. | Not inc. | Not inc. | Not inc. | Not inc. | Not inc. | Not inc. | Not inc. | Not inc. |
| areoporto | aeroporto | 197000 | 1920000 | 10.26 | 510 | 32494 | 1.57 | 65 | 17403 | 0.37 |
| biricchino | birichino | 10300 | 41000 | 25.12 | 438 | 606 | 72.28 | 17 | 75 | 22.67 |
| Caltanisetta | Caltanissetta | 488000 | 2650000 | 18.42 | 65 | 615 | 10.57 | 45 | 901 | 4.99 |
| collutazione | colluttazione | 1370 | 35500 | 3.86 | 119 | 398 | 29.9 | 23 | 242 | 9.5 |
| colluttorio | collutorio | 11900 | 37300 | 31.9 | 285 | 151 | 188.74 | 8 | 13 | 61.54 |
| conoscienza | conoscenza | 35900 | 5860000 | 0,61 | 463 | 208400 | 0.22 | 81 | 28069 | 0.29 |
| coscenza | coscienza | 48000 | 2230000 | 2.15 | 1088 | 212700 | 0.51 | 163 | 16970 | 0.96 |
| eccezzionale | eccezionale | 98200 | 2270000 | 4.33 | 750 | 77024 | 0.97 | 41 | 1924 | 2.13 |
| efficenza | efficienza | 91000 | 2380000 | 3.82 | 300 | 11925 | 2.52 | 113 | 2175 | 5.2 |
| essicare | essiccare | 1510 | 54400 | 2.78 | 28 | 266 | 10.53 | 5 | 62 | 8.06 |
| esterefatto | esterrefatto | 12300 | 48700 | 25.26 | 326 | 731 | 44.6 | 42 | 189 | 22.22 |
| ingegniere | ingegnere | 585 | 2050000 | 0.03 | 68 | 22497 | 0.3 | 30 | 2009 | 1.49 |
| interpetare | interpretare | 363 | 2830000 | 0.01 | 39 | 42981 | 0.09 | 1 | 1526 | 0.07 |
| Macchiavelli | Machiavelli | Not inc. | Not inc. | Not inc. | Not inc. | Not inc. | Not inc. | Not inc. | Not inc. | Not inc. |
| Missisipi | Mississippi | 791 | 403000 | 0.2 | 86 | 1131 | 7.6 | 14 | 373 | 3.75 |
| metereologia | meteorologia | 143000 | 955000 | 14.97 | 287 | 511 | 56.16 | 35 | 666 | 5.26 |
| peronospera | peronospora | 797 | 27900 | 2.86 | 34 | 26 | 130.77 | 6 | 7 | 85.71 |
| pressocché | pressoché | 86600 | 2480000 | 3.49 | 789 | 26764 | 2.95 | 271 | 1600 | 16.94 |
| scenza | scienza | 89200 | 6480000 | 1.38 | 110 | 101573 | 0.11 | 54 | 41621 | 0.13 |
| scorazzare | scorrazzare | 25100 | 41100 | 61.07 | 462 | 758 | 60.95 | 69 | 117 | 58.97 |
| **Totals** | | 1341916 | 32793900 | 4.09 | 6745 | 756404 | 0.89 | 1139 | 117889 | 0.97 |