Linguistic variability of Web Italian: a working empirical grid

Mirko Tavosanis

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

Presentation

The main constituents of a text, including word frequency, are strongly correlated with the style and topic of the text itself. This kind of variation was detected, a number of decades ago, in the first statistical analysis of corpora of written or spoken language (for the Italian situation see Bortolini 1971: XIV-XV; Voghera 1993). Such variations in style and topic are, however, often documented only in texts whose linguistic features are already acknowledged. In the field of Web linguistics there is little evidence of study and classification of the new kinds of text with reference to their linguistic individuality. Moreover, this dearth of studies often makes it difficult to recognize as such even widely-read kinds of texts.

In these circumstances, when dealing with the language of the Web even the most widely accepted methods of linguistic analysis often seem at a loss. It is interesting, for instance, to note that one of the few studies for Italian in this field limits itself to newspaper-like texts published on the Web (Prada 2003).

An empirical grid

An easily recognizable language of the Web does not exist in any known national language (Crystal 2001). It is probably possible, however, to trace a series of relevant linguistic phenomena to real categories of diaphasic and diamesic variation and to demonstrate a series of linguistic correlations.

This paper will demonstrate some linguistic applications of a classification of Web texts based upon six empirical layers. The layers derive from different analyses of specific texts and were partially elaborated for a research program co-ordinated by Mirko Tavoni and centred upon the blog *La meglio gioventù*. The layers in question are extracted from experience and tentative definitions are rooted in knowledge of the editorial situation of the Web (Nielsen 2004).

Four layers relate directly to the writing process:

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

Two of them relate to the attitude of writer before or during writing:

5. text type
6. intended reader of the text

The latter two layers have significant parallels with more traditional classifications of text. When applied to the language of the Web they have also peculiarities that will be outlined below.

The whole classification, slightly different from current grids, is currently tested on Italian Web pages. Some of its parts also have wider implications and can be applied to other languages. However, at this stage of the work it is intended mainly as a compendium of vocabulary to allow Italian linguists to speak more correctly about the texts published on the Web and to place particular phenomena in context.

The six layers

First layer: time allowed for writing

We can empirically classify at least four main categories, related to specific types of texts:

- fast and unrevised 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.

In e-mail and chat, both in Italian and in other languages, there is a strong bias towards speed of writing at the expense of revision (see Pistolesi 2004). It should be straightforward to assume that forum postings are written more quickly than traditional Web pages and are often published without a second reading. This common sense conclusion, however, is only partially supported by the data (see case study # 2).

Second layer: writing tool

The second layer is effectively low-level encoding and it is strictly linked to the peculiarities of the Italian orthographic system. However, a reduced comprehension of its role seems to bring real-world problems in the interpretation of data during the research work.

The final aspect of an Italian text can be influenced by the use of particular tools:

- keyboard or interface not adapted to the task (forum postings)
- standard keyboard or interface, adapted to the task (Web pages, forum postings)
- professional tools (Web pages)

In the case of Italian orthography, Web publishing often encounters problems not only with text formatting (italics, bold and so on), as for other languages, but also with 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 forum *La meglio gioventù* published in 2004 by the Web site of the newspaper *La repubblica* displays many examples of this (in particular for text formatting). The forum, however, has seen a wide participation by Italians living abroad. Many orthographic deficiencies can therefore be explained by the use of non-Italian keyboards (e.g. keyboards without accented characters) and not by lacunas in the orthographic competencies of the writers. The following is a quotation from a posting coming 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.

The top level of competence is, in this framework, the correct encoding of the text for multi-platform use (e.g. Unicode).

Third layer: writing support
Text published on the Web can be stored in different formats:

- Html (displayed by the browser without plug-in software)
- image (graphic files displaying text)
- non-html file (.doc files, .pdf files)

It is worth noting the existence of particular graphic images designed to filter human beings from other "readers" of Web pages (see sixth layer). To avoid automatic registration procedures, many sites now ask readers to copy the text written in an image and deliberately distorted. Human readers should be able to decrypt the correct reading, OCR software should not (in another electronic medium, many spam systems for e-mail use tricks of this kind to allure human readers and avoid the attentions of spam filters). Here is an example from the Yahoo Italia site (Figure 1):

Figure 1 – A deliberately distorted text image from the Yahoo.it site

More relevantly, many texts not planned for the Web are now routinely "published" as non-html files. This state of affairs is described in more detail in Case Study #1.

Fourth layer: creation of writing
The text displayed on the Web can be created, alternatively:

- completely by human beings (forum, blogs, Web pages)
- partially by database combination (Web pages)

The database combination is widespread. In its most straightforward form it simply publishes data following a carefully designed output outline. In the form of search engines reports to queries, however, it creates a kind of text without any correspondence in the world of paper, created by partially semantic or non-semantic cutting of texts (Tavosanis 2004 and Figure 2).

Fig. 2 – Display of search results on the google.it site

Fifth layer: text type
Some text structures usually created for hard copy (like the novel or short story) have little diffusion on the Web, while others (e.g. newspaper articles, announcements and so on) play a substantial role in the reading habits of Web users. Among such text types, there are relatively few "text structures" original to the Web and known to the casual reader, but they represent a good percentage of the texts read on the Web. The most frequent texts types include:

- forum posting
- blog entry
- content of generic sites (infotainment, product description...)
- content of news sites or of portals (news, short informative texts...)
- search engines (search interface, display of results)

Moreover, the microcontents must also be taken into account: window names, menu names and so forth. Microcontents are present in every text structure and are of paramount importance in the navigation and orienteering of readers.

Sixth layer: intended reader of the text
Web writing is subject to the traditional rhetorical constraints: it can be directly aimed at a particular reading target, a specific goal and so on. These constraints are described in many Web writing manuals (for the Italian language, see especially Carrada 2000). However, Web writing must also take into account a non-traditional form of reading: the reading done by non-human readers, e.g. search engines.

On many Web pages the text written for search engines is not directly visible on the page, but it is frequently present as non displayed metadata. Such metadata often take the form of long lists of keywords, usually written by human beings to ameliorate the position of the site in the query output of search engines. An empirical survey of the metadata shows wide discrepancies in their use and writing style.

The keywords on the home page of the *La Repubblica* news site take this (unusually long) form:

<meta name="keywords" content="La Repubblica, notizie internazionale, giornaliere, nazionale, politics, scienze, business, affari, finanza, sport, cronaca, international news, daily newspaper, national, politics, science, business, your money, breaking news, business technology, technology, circuits, navigator, sports, editorial, forum, discussioni, sondaggi, calendari modelle, moda, bellezza, fashion, glamour, oroscopo, concorsi, lavoro, finanza, borsa in diretta, Piazza Affari, Mibtel, Wall Street, ricerca e annunci di lavoro, assicurazioni online, scuola, università, gallerie fotografiche e immagini, Webcam, sms, vignette, commenti, motori, polizze auto e moto, listini prezzi, salute, terme, farmaci, medicine, previsioni meteo, programmi tv, programmazione cinematografica, radio, canzoni, testi, Mp3, shareware, freeware, cellulari, programmi, audio, video, giochi, lotto, totocalcio, enalotto, superenalotto, estrazioni. Tutti gli approfondimenti: sport, calcio, gol, marcatori, classifica, coppe europee, basket, Formula 1, Ferrari e Schumacher, Fantacalcio, Parlamento, leggi, elezioni, deputati, senatori, Forza Italia, Ds, An, Lega, Margherita, Udc, Udeur, Rifondazione comunista, manifestazioni, cortei, scontri, sindacati, governo Berlusconi, Confindustria, Rai, Mediaset, satira, terrorismo, giustizia, giudici, processi, mafia, cronaca, rapine, violenze sessuali, omicidi, pedofilia, terremoti, incendi, maltempo, previsioni meteo, turismo, viaggi, week end, mare, montagna, laghi, alberghi, voli aerei, crociere, videogiochi, casa, mutui, computer, pc, Microsoft, Apple, Ibm, Sony, Nintendo, Playstation, informatica"/>

It is interesting to note that together with simple listing of words the metadata also contain syntactic structures (e.g. “ricerca e annunci di lavoro”, “Tutti gli approfondimenti”). Perhaps this bears traces of a paste-and-copy elaboration, since the harvesting of keywords by search engine is usually done one word at a time. Also noteworthy is the presence of English keywords in a Web site written almost completely in Italian.

Testing the grid through search engine queries

Traditional linguistic investigation has to limit itself to a comparatively small subset of the language. We can take as an example the Coris / Codis corpus of written Italian: it includes 60 million words from the books published in the years 1980-2000, but this figure represents only 1/666 of the words actually published (Italian book publishing issues approximately 2 billion words every year: ca. 50,000 new yearly titles with an average length of 40,000 words).

The Web, however, can be examined in a completely different way. Estimates concerning the percentage of the Web effectively indexed by search engines like Google indicate up to 50% percentage of indexing. The data tracked by Google can be treated as if they defined the *whole* writing of the Web, more than a restricted specimen.

Naturally, commercial search engines lack many features typical of corpus querying systems. But even their rough linguistic functions can then be exploited in a significant way (Calishain 2003; Maxwell 2004; Davis 2005), especially when applied to lexical searches. The trickiest problem, in this field, is probably the fact that the most efficient search engine, Google, gives only the approximate number of pages where a given token occurs, instead of the sheer number of occurrences of the token. Due to this behaviour of the engine, values and figures can be compared only with one another and cannot give reliable absolute values.

The searches cited in the following case studies were conduced with Google in the winter of 2004 and in the spring of 2005. All of the searches were restricted to the pages in Italian (through the related Google function); the figures given always refer to the “number of pages” only.

Case study #1: distribution of demonstratives according to file type

Simple search queries demonstrate the strong relationship that some linguistic forms bear to the variables described in section 2. For example, we can demonstrate a variability factor of approximatively 1 order of magnitude in the use of bureaucratic demonstrative forms in texts according to the different file formats. This difference is easily explained by the fact that bureaucratic texts are often published on the net simply by placing on the Web text files created for print and/or for internal office use.

Through the search engine we can then try to assess the percentages of use of the three Italian demonstrative pronouns and adjectives *questo*, *codesto* and *quello*. A Google search for the pronoun and adjective *codesto*, typical of spoken Tuscan and of bureaucratic use but not of common Italian, yields the following results:

Generic search: 35.900
search restricted to .pdf files: 6,050
search restricted to .doc files: 3,850
Taken together, .pdf and .doc files account for more than 27% of total hits.

If we perform the same search the pronoun and adjective *quello*, commonly used also in spoken

Italian, we have the following results:

Generic search: 8.690.000
Filetype .pdf: 552,000
Filetype .doc: 123,000
Taken together, .pdf and .doc files account for 7.8% of total hits.

The same search for the pronoun and adjective *questo*, widely used also in neo-standard Italian (Berruto 1987), gives the following results:

Generic search: 23,400,000
Filetype .pdf: 728,000
Filetype .doc: 143,000
Taken together, .pdf and .doc files account for only 3.4% of total hits.

Case study #2: typos in forums and blogs

Unrevised writing should made its nature evident in many aspects. One such feature should be the frequency of typos in writings like forum postings. However, we can find little evidence of this in the distribution of typos when examined through a search engine. The common typo *propio* instead of *proprio* in many forums appears with a frequency slightly lower than its percentage on the whole Web, 1.75%. The variation is not only less than in Case study 1: it is often in the wrong direction. On average, in 12 randomly selected forum sites the typo represents only 2.16% of the forms of the word (see Table 1).

This kind of search is of course subject to statistical error. Testing it with four of the most popular blog sites in Italian produces a percentage of typos of only 0.59% (with a maximum percentage of errors of 1.83% and a minimum of 0.31%). Testing it with four of the most popular newspaper sites in Italian yields a typo percentage of only 0.1%.

This suggest two conclusions:

1. that there are effectively measurable differences in text revision;
2. that the average of the Italian Web is approximately the same as the Italian forums.

We can further test this conclusion with another word often considered difficult to spell, *aeroplano*, for which many Italian dictionaries, such as the Zingarelli and the De Mauro, explicitly suggest avoiding the popular form *areoplano*. Searching the Web, however, we find in the first place that the most common spelling error is not *areoplano*, but *aereoplano*. The latter scores 27.2% of occurrences of the singular of the word in the Web (data extracted from seven of the most popular Italian blog sites are summarized in Table 2). It is interesting to note that this is a form unknown to the most widespread Italian dictionaries.

If we try to fix the form, we find that in blogs the percentage of occurrences is higher only for the popular form *areoplano*. Forums often have such a low frequency of this word as to be statistically irrelevant; newspaper sites have on average a much lower ratio of errors (in the *Repubblica* Web site we found only one *areoplano*, in a forum section, and 8 *aereoplano* as against 69 correct forms). The misspelled *aereoplano* is instead less common in blogs than in the whole of the Web.

We can then begin to glimpse a kind of linguistic average. Forums place themselves just below the line, blogs just above it. If these samplings are representative of the Web, further analyses should confirm this reconstruction and perhaps allow the creation of a more realistic portrait of current Web Italian.

Bibliography

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

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

[Calishain 2003] Calishain, T., and Dornfest, R. *Google Hacks*. O'Reilly, Beijing, etc.

[Carrada 2000] Carrada, L. *Scrivere per il web*. Lupetti, Milano.

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

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

[Maxwell 2004] Maxwell, M. "Resource Discovery for Low Density Languages: Internet Search". Abstract in ACH/ALLC 2004 - Goteborg University, Goteborg, pp. 88-89.

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

[Prada 2003] Prada, M. *La lingua del web*. In Bonomi, I., Masini, A. and Morgana, S. (eds) *La lingua italiana e i mass media*. Roma, Carocci: pp. 101-120.

[Tavosanis 2004] Tavosanis, M. *L'italiano in rete*, ICoN module, ICoN site.

[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, pp. 32-38.

Table 1
Occurrences of *propio* and *proprio* in different kinds of sites

Blog sites

(pages)	Splinder.it	Splinder.com	Blog.excite.it	Clarence.com	Sum
Proprio	50800	163000	129000	44600	387400
Propio	946	499	616	241	2302
Total	51746	163499	129616	44841	389702
% of irregular forms	1,83	0,31	0,48	0,54	0,59

Forum sites

	cicloweb.it	risorsehitech.it	metaforum.it	spazioforum.net	forum.html.it	forumfree.net
Proprio	529	190	1070	1480	86	12400
Propio	6	3	8	3	3	686
Total	535	193	1078	1483	89	13086
% of irregular forms	1,12	1,55	0,74	0,2	3,37	5,24

Web

	Groups	Italian web pages
Proprio	5250000	4770000
Propio	57800	84900
Total	5307800	4854900
% of irregular forms	1,09	1,75

Newspaper sites

(pagine)	repubblica.it	corriere.it	unita.it	ilmattino.caltanet.it	Sum
Proprio	55300	27900	13000	5690	101890
Propio	87	16	0	4	107
Total	55387	27916	13000	5694	101997
% of irregular forms	0,16	0,06	0	0,07	0,1

Table 2

Occurrences of *aeroplano* and of related typos ind blog sites

	Total	Groups	Pages in Italian
Aeroplano	969	6242	49400
Aereoplano	204	1100	18900
Areoplano	69	249	1180
Total	1242	7591	69480
Total of irregular forms	273	1349	20080
% of irregulars forms face to total	21,98	17,77	28,9
% more irregular form to total	5,56	3,28	1,7
% more irregular form to aeroplano	7,12	3,99	2,39
% less irregular form to total	16,43	14,49	27,2