Traditional corpora and the Web as corpus: a comparison for Italian newspapers

1. Method and hypothesis

The great corpora created in the last 20 years are often based upon the assumption that objects or text types do influence heavily the linguistic phenomena (Summers 1991; Kennedy 1998: 62-66). Therefore, modern general corpora do not aim explicitly to be representative of language. They aim, however, to be representative of some subset of language, or of "objectively defined document and text types" (Summers 1991: 51). Many modern corpora are composed by the balanced aggregation of different subcorpora, linked to special text types.

The problem of balancing different subcorpora will not be taken into account here. I will try, however, to evaluate representativeness of subcorpora. This could be done, until recently, only through comparison of different corpora – a kind of self-contained proof. Indipendent and objective evaluation is now instead possible thanks to the growing strength of search engines. This strength allows now researchers to query, even if in much-defective way, a significant, and growing, percentage of the whole of a textual type.

Table 1 shows some of the most recent data (September 2004-August 2005) regarding diffusion of Italian newspapers according to the ADS certifying service (http://www.adsnotizie.it/). The table lists 63 newspapers. As a rule of thumb, we can then suppose that each of them generated in the 20 years taken into account by the corpus at least 380 Mw, since this is the total of tokens indexed for the years 1985-2000 by the *La Repubblica* corpus (see below). However, actual figures could be sensibly higher. Even if content-sharing is widespread and many local newspapers probably do publish less text than major ones, many Italian newspapers print each day more than 100,000 words, or 35Mw each year. We can tentatively assess total yearly output of Italian newspapers at 1,000Mw. Newspaper corpora do cover only a fraction of this. Are they representatives?

We can take as a working hypothesis that a particular subcorpus will be "representative" if it displays linguistic features in a way comparable to what we can find in the whole of the production of its textual type. In this paper I will therefore try to evaluate representativeness of newspaper corpora using as indicators some linguistic features of the so-called "neo-standard" Italian since.

2. The existing corpora and the Web

I will describe here the main corpora for Italian newspapers and their structure.

2.1 The Coris / Codis corpus

The Coris / Codis corpus is to date the largest general corpus for Italian language available through the Web (see http://corpus.cilta.unibo.it:8080/, including bibliographic references). It is available to researchers in twin form:

- The Coris corpus (*COrpus di Riferimento dell'Italiano Scritto*) is a corpus of written Italian, created with texts of the 1980-2000. It includes 100 millions of words, 38 of which are taken from the "press", according to a repartition in Sections ("newspapers, periodic, supplement") and Subsections ("national, local specialist, non specialist, connotated, non connotated").
- The Codis interface (*COrpus Dinamico dell'Italiano Scritto*) allows users to query pre-composed subcorpora of the Coris. The Codis subcorpus "Stampa / Press" can be queried in dimensions of 3, 5, 10 and 20 Mw. This includes also the data from periodic

press (weekly and monthly publication) and supplements, but it seems that at least three quarter of the texts are taken from newspapers.

2.2 The *La Repubblica* corpus

The *La Repubblica* corpus managed by the SSLMIT-DEV of Forlì includes to date 380Mw taken from a CD collection of articles published by this newspaper in the years 1985-2000. This corpus, "While arguably not ideal as a reference corpus – being mono-source – (...) is probably the largest freely accessible Italian corpus available to date" (Baroni et al. 2004). The description of the corpus does not state it explicitly, but it seems that the corpus does not include the whole of the articles published by the newspaper. The whole corpus has been POS-tagged and it is searchable through advanced interface.

2.3 The Italian online newspapers

Many Italian newspapers have now a significant presence on the Web. Some of them are publishing only the contents of the paper version (or a selection thereof). Many others are instead publishing special contents.[1] In the sites of the newspaper themselves we find often also texts created by readers (forums, blogs and so on). It is difficult to evaluate the percentage of "traditional" text face to less traditional ones.

The Audipanel Nielsen NetRatings statistical survey includes eight newspapers sites (Table 2) among the 166 Italian sites monitored for October 2005. I will not include in the analysis the news sites connected to TV news; however, two of the sites I include lack of a one-to-one connection with a traditional newspaper: http://www.quotidianiespresso.it/ includes all of the local newspapers of the Espresso group; http://www.quotidiano.net/ is instead a newspaper-like information site without corresponding paper publication. Apart from those inclusions, the number of viewed pages from newspaper sites creates a rank order which closely match the rank order of the traditional newspaper. However:

- *L'Unità* occupies the 25th position among traditional newspapers, but it ranks 5th among newspaper sites, exceeding *La stampa* in page views (even if not in unique audience numbers).
- the relative strength of the traditional newspapers is different from the relative strength of newspaper sites even when their rank is the same.

Indexing of those sites by search engines is not homogeneous. In many cases only recent articles can be queried. This accounts for part of the huge differences in figures we will see for the indicators.

3. The query

Queries on the Coris / Codis and on the *La Repubblica* corpus were conduced using the public interface of both corpora. Web queries were conduced through the most efficient public search engine to date, Google.

Now, even if commercial search engines can still be exploited in a significant way (Kilgarriff and Grefenstette 2003:342; Calishain and Dornfest 2003; Maxwell 2004; Davis 2005), their unsophisticated linguistic functions lack of many features typical of corpus querying systems. The trickiest problem with Google is probably the fact that, when huge numbers of occurrences are

1 Online editions:
Full list by Ipse.com http://www.ipse.com/quotit.html

involved, the engine provides only the approximate number of *pages* where a given token occurs, instead of the exact number of occurrences of the token. The figures provided in the following tables are therefore referred to the "number of pages" or to the approximate number of occurrences found by the search engine. Due to this behavior, and to the secret nature of the computing mechanism, values and figures given by Google cannot provide reliable absolute values similar to those of conventional corpora. They can however be compared among themselves, allowing researchers to obtain comparable data in the form of ratios.

Of course, search engines do contain also an unavoidable percentage of "dirt". Spurious data can be filtered up to a certain extent by restricting the query to forms used only (or mainly) in the Italian language and restricting the search to the pages written in Italian (using the option of language restriction offered by the engine).

Even if the *La Repubblica* corpus has been POS-tagged, neither the Coris / Codis nor the Google indexes share this status. The queries have been therefore limited to forms instead of lemmas..

The searches described below were conduced in November-December 2005. Part of them was conduced using an experimental interface to the Google APIs created under the supervision of Prof. Paolo Ferragina at the Department of Computer Science of Pisa University.

4. Differences between the "Stampa" Codis subcorpus and other corpora

For the purpose to trace basic differences and to evaluate the composition of a newspaper corpus I propose to use two sets of indicators:

1. relative frequency of the demonstrative pronouns and adjectives *questo*, *codesto* and *quello*;
2. relative frequency of the subject pronouns of 3rd singular person (also in Bonomi 2003).

Both of them relate to facts of contemporary use well described from a qualitative point of view (Sabatini 1985, Berruto 1987). The second indicator is also included in a list of 13 linguistic tracts considered by Ilaria Bonomi as useful to describe differences in the Italian used by newspapers in 1991 and in 2001 (Bonomi 2003: 192-193; the 12 remaining tracts are difficult to study with general search engines such as Google).

As for the first indicator, the *La Repubblica* corpus, being POS-tagged, allows to discriminate between pronouns and adjectives. Both the Coris / Codis and Google, however, offer no way of doing this. I will therefore talk about "demonstratives" in a general way without trying to see in further detail the composition of this set.

The results of the comparison are shown in Table 3. We can consider an high ratio of *questo* as a sign of "innovative" tendencies; the form *codesto* is instead typical of Tuscan regional use and of literary or bureucratic use. As one could foresee, specialist newspapers shows their nature even through such indicators. The site of *La Gazzetta dello Sport* accounts for no "codesto"; the site of the economic newspaper *Il sole – 24 ore* shows a much higher percentage of *codesto*.

As for the second indicator, the diffusion in the newspapers of the "innovative" forms *lui* for the masculine and *lei* for the feminine and the corresponding regression of the traditional and literary corresponding forms *egli* and *ella* is a well known fact. Bonomi (2003: 195) speaks in general terms of a more than 10:1 ratio in favour of the innovative masculine form in the newspapers of 2001, in front of a nearly 2:1 ratio in 1991 (data are drawn from a little corpus, possibly of less than 1Mw). Anyway, the Codis subcorpus data seems to indicate that the 10:1 ratio is more or less the average of the years between 1990 and 2000.

As for the feminine, the *ella* form does not appear in the Bonomi corpus (see again Bonomi 2003: 195). In the Codis subcorpus we can instead find one occurrence of *ella* every hundred *lei*. Broadening this comparison to the source already seen for the demonstratives we find the figures

shown in Table 4. In the sum of Web newspapers, in particular, we find an higher ratio of *lui*, but also an higher ratio of *ella*. Again, different newspapers show very different percentages. The ratio of the innovative *lei* in *La Repubblica* is 25 times higher than in *Il sole-24 ore*. The same *La Repubblica*, however, is markedly more conservative as for the masculine form.

Three facts seem constant.

1. Among online newspapers, there are significant differences not only between all-purpose and specialist publications (such as between, say, *La Repubblica* and *Il sole - 24 ore*), but also between different newspapers of the same category (such as *Il corriere della sera* and *L'Unità*).
2. Differences between the ratio of each tract shown by Codis subcorpus and the sum of the online newspapers never exceed 50% of the former figure.
3. Differences between the Codis subcorpus and non-newspaper kind of text are instead more marked, up to the 5000% increase we can find as for the difference in the *lei* / *ella* ratio between the subcorpus and the whole of the Web.

In the whole, differences between the Codis subcorpus and the reality of newspaper seem reduced.

5. Conclusions

Calculations and parameters shown in the paper are of course very rough. Clearly, both statistical tools and special monitoring software are needed to give to this kind of search more focus and more depth. Also, future searches must achieve a better understanding of the covering of search engines and should be based on more than a single engine. It would be useful, moreover, to identify and exploit other searchable indicators of the linguistic quality of a text.

On the other hand, even those data can give substance to a linguistic description of the Italian used by newspapers. Banishment of "literary" forms has been often noticed and it is confirmed by direct questioning of journalist and study of editing practices (see in particular Bonomi 2003). The Google queries suggest that this is a real fact and that the Coris / Codis samples do represent it in the right way.

Moreover, this confirm of the usefulness and representativeness of traditional corpora is significant. Commercial search engines will not have, for the foreseeable future, many of the functions of traditional corpora for selection, filtering of data and managing of the output. It is therefore interesting to know that data from traditional corpus are "representative" and can be used to make linguistic descriptions confirmed by facts.

Baroni et al. 2004 = Baroni, Marco, Bernardini, Silvia, Comastri, Federica, Piccioni, Lorenzo, Volpi, Alessandra, Aston, Guy and Mazzoleni, Marco. *Introducing the la Repubblica corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian*. *Proceedings of LREC 2004*, Lisbon: ELDA: 1771-1774.

Berruto 1987 = Berruto, Gaetano. *Sociolinguistica dell'italiano contemporaneo*. La Nuova Italia, Firenze.

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

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

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

Kennedy 1998 = Kennedy, Graeme. *An introduction to corpus linguistics*. London: Longman.

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

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

Sabatini 1985 = Sabatini, Francesco. *L'italiano dell'uso medio: una realtà tra le varietà linguistiche italiane*. In Holtus, G. & Radtke, E. (a cura di). *Gesprochenes Italienisch in Geschichte und Gegenwart.* Tübingen: Gunter Narr: 154-184.

Summers 1991 = Summers, Della. *Longman/Lancaster English language corpus: Criteria and design*. Technical Report, London: Longman.