An observatory on Spoken Italian linguistic resources and descriptive standards

***Miriam Voghera, ^Francesco Cutugno**

*Dept. of Linguistic and Literature, University of Salerno, Italy
^ Dept. of Physical Sciences, University "Federico II", Naples, Italy
*Via Ponte Don Melillo, 84084, Fisciano (SA)- Italy
^Complesso Universitario di Monte S.Angelo, Via Cinthia, 80126, Napoli, Italy
voghera@unisa.it , cutugno@na.infn.it

Abstract

We present the national project "Parlare italiano: osservatorio degli usi linguistici", funded by the Italian Ministry of Education, Scientific Research and University (PRIN 2004). Ten research groups participate to the project from various Italian universities. The project has four fundamental objectives: 1) to plan a national website that collects the most recent theoretical and applied results on spoken language; 2) to create an observatory of the linguistic usages of the Italian spoken language; 3) to delineate and implement standard and formalized methods and procedures for the study of spoken language; 4) to develop a training program for young researchers. The website will be accessible starting from November 2006.

1. Preliminaries

Most of the linguistic descriptions and analyses show that spoken language differs systematically from written language (Halliday, 1989). Spoken texts, even belonging to different diastratic and diaphasic registers, present similar regular features that make them very different from written texts (Voghera, 2000). Spoken language exhibits an 'intrinsic functional discontinuity': the on-line production of speech is physically continuous, but actually produces texts, which are deeply discontinuous (Sornicola, 1981; Biber, 1995; Voghera, 1992). A spoken text is the result of a multiparty activity to which both speaker and receiver contribute, so the speaker knows in advance that her/ his utterance can be interrupted, and that the initial textual strategy can dramatically be altered.

The increasing amount of investigations in the last decades has also shown that a deep understanding of the structure of spoken texts depends on the explanation of the entire physical process of transmission and reception of the speech signal. Fundamental data derive from researches on the physical features of phonic-acoustic signals and on the physical correlates of linguistic functions.

For all these reasons, studies of spoken language integrate many different areas of scientific interests: from physical aspects, such as transmission and reception features of speech signals, to structural verbal aspects, such as syntax or morphology, or to much more pragmatic aspects, such as communicative competence. In Italy attention to the epistemological significance of the spoken language was paid since the late 1960s, although it was only in the 1980s-1990s that the first empirical studies were carried out. Several studies have provided descriptions of contemporary spoken Italian, from phonetic to textual and pragmatic level. Nonetheless, a complete and systematic description is still lacking, because of the complexity the linguistic Italian situation. The expression 'spoken Italian' is not totally unambiguous, and can have different connotations, according to the different points of view adopted, e.g. geographical or social.

Moreover, not all the levels of linguistic structure have been investigated. For instance, the diatopic dimension, that has a central role in recent research on intonation, phonetics and the lexicon, has not been completely explored at the morpho-syntactic and the textual level. The diachronic consideration of speech phenomena has not been much developed by studies on contemporary spoken language, though such an analysis is possible by means of comparing contemporary spoken language with written documentation of the past (see for instance: D'Achille, 1990), and also by comparing different samples of contemporary spoken language belonging to corpora collected in different moments.

A 'national point of reference' is required even as far as the methodological procedures are concerned. In Italy the first corpus of spoken Italian, LIP, was published in 1993 and recently several projects (AVIP-COFIN 97, API-COFIN 99, CLIPS; IPAR-COFIN 01; LABLITA; C-ORAL-ROM) provided a great amount of spoken material. Interactive databases store different diatopic/diaphasic varieties of spoken Italian. These materials are phonetically, phonologically (according both to standard and regional variety phonology), lexically, prosodically, morpho-syntactically, textual-pragmatically annotated and labelled.

The more the amount of data grows in variety and dimension, the more the definition of a standard structure of catalogue descriptions and the structure of metadata descriptions become crucial. Metadata structure has been the object of *de facto* standardization (Gibbon et al., 1997). by the main association for corpora collection in Europe (ELRA) and in the US (LDC).

The studies on standards of representation and annotation have gained an increasing interest in Italian theoretical, applied and computational linguistics. Several projects explored the possibility of comprehensive multilevel systems of representation and annotation, and many research groups produced protocols and tools for spoken corpora. Nonetheless, we still lack an initiative, which, starting from the different points of view and proposals, could represent a national point of reference as far as Italian data are concerned.

In this paper we present the project "Parlare Italiano: osservatorio degli usi linguistici", funded by the Italian Ministry of Education, Scientific Research and University (PRIN 2004). Ten research groups participate to the project (see §5 for complete list).

The project has four fundamental objectives:

1) to plan a national website that collects the most recent theoretical and applied developments of the researches on spoken language;

2) to create an observatory of the linguistic usages of the Italian spoken language;

3) to delineate and to implement standard and formalized methods and procedures for the study of spoken language;

4) to develop a training program for young researchers.

In order to guarantee the greatest transparency and verifiability of the research, the project has as objective to respect the following criteria:

a) publicity of reference corpora;

b) publicity of methods and analysis procedures;

c) publicity of results.

The project "Parlare Italiano" presents different scientific and applied points of view, because it involves scholars working in several fields of research (Linguistics, Computer Science, Speech and Hearing Sciences), who have great research experience in the study of spoken language both in national and international projects (AVIP, API, IPar, CLIPS, Coral-Rom). This project will develop theoretical and applied instruments in the following thematic areas: Phonetics and Phonology; Prosody; Morphology; Lexicon; Syntax; Semantics; Discourse and Conversational Analysis; Pragmatics; Diachrony of spoken language; Italian as second language; Spoken language and Mass-media; Speech and Language Disorders; Computational Linguistics; Speech Technology.

Each thematic area developed in the project will include sections devoted to: corpora; protocols for the standardization of representation formats; tagging and annotation tools; bibliographical references.

The website parlaritaliano.it, whose structure is an important part of the research, will be the first national initiative which has as objective not only to study and to describe the spoken Italian in all its theoretical and applied aspects, but also to validate protocols for the collection and the analysis of spoken materials, which guarantees the control of data and results.

2. Project syllabus, phases and expected results

The comprehensive program of research is articulated in different areas of research, which constitute the 'thematic sections' of the web site parlaritaliano.it

In many cases, however, sub-projects within the general frame contemplate some cross-area multilevel analyses of spoken language structures and investigations on level correlations and interfaces. On the one side, this makes problematic a rigid partition of research subjects, but, on the other side, it contributes to create and to design the hypercategorial structure of the whole project. We took in consideration such intersections in defining the programmatic lines of the research phases.

2.1. Planning of web site

In the first year of the project we set up a permanent observatory of the Italian spoken language through the creation of the reference framework within which the contributes of all the research units will be published.

We planned the logical architecture of a web-site defining the relationships among the contents, their representation and their structural organization, the relationships among linguistic structures concerning navigation and fruition modalities of the web-site, and the links to other sites within it.

Such activities constitute an important part of the research, because of the great and challenging theoretical and methodological implications. All research units contributed at the site plan, defining criteria for data and metadata collection of Italian spoken language corpora, for their use in the net and, above all, delivering surveys of the existing information available in the various areas that are present in the site.

At the same time we are selecting and organizing, from a conceptual point of view, the contents that will be included in the different sections of the website, that is corpora, tools, analysis and representation protocols.

2.1.1. Corpora

The catalogue of the Italian corpora, accessible through the net, is an essential infrastructure in order to guarantee the access and the utilization of linguistic resources. This is particularly important if we take into account the present tendency to integrate the linguistic resources in multilingual structured European catalogues, which is supported by an European network in the field of Humanities and Computing. Spoken language corpora will be accessible directly and via links to the main websites of Italian and foreign corporations and universities.

2.1.2. Tools, analysis and representation protocols

The realization of the electronic format of the catalogue will conform to the international standards of representation of linguistic resources. This step is fundamental to guarantee the circulation in the net of the catalogue and the full exploitation of the specific resources by the final users.

Tools which guarantee the application of the standard are distributed without restraints in the net and allow both the utilization of the metadata in the widest international dominion, and, when allowed by the holder, the direct access to the resources. A survey of the main existing formats and of their features is in course of preparation to contribute to the development of application for the planning of the site.

All the units collaborate providing protocols and tools of analysis and codification (like software for multilevel annotations, instruments of automatic and semi-automatic analysis of signals) expressly realized for other projects and designing the map of the links to resources already in the net.

2.2. Thematic areas

As already said, many sub-projects cover, in their intersections, a wide range of subjects of spoken language linguistics. They are articulated in different phases, lines and analysis procedures and have different intertwined objectives. However, it is possible to draw main lines aiming to achieve general, intermediate common objectives. All the research units have mostly attended to the development of preliminary activities to complete the

linguistic analyses and the treatment of linguistic data and in particular:

1. Extension and updating of the support and reference bibliography.
2. Planning, selection, design and acquisition of the corpora for the analyses.
3. Definition of protocols, representation systems, data and metadata codification and their elaborations.
4. Definition of analysis methodologies, classification and interpretation.
5. Definition of instruments and tools for data analyses.
6. Preliminary studies for the creation of integrated databases.

According to the progress of the preliminary activities, the research units are now developing the linguistic analyses of corpora and/or the testing of the analysis criteria, first elaborations and descriptions of the data and the verification of scientific and methodological hypotheses.

3. Linguistic framework and levels of analysis

Each research unit contribute to one or more areas and will develop many cross-area subjects:

1. Multilevel grammar analysis of Italian spoken language.
2. Applied and computational linguistics.
3. Methods and analysis procedures.

3.1. Multilevel grammar analysis of italian spoken language

The project is presently providing a wide range of linguistic analyses on many different levels of Italian spoken texts belonging to several diatopic, diachronic varieties and registers.

3.1.1. Phonetics and phonology

- analysis and classification of co-articulation and segmental reduction phenomena;
- analysis of voice and voice 'labels';
- analysis of segmental features of TV speech;
- text –to-speech alignment.

3.1.2. Prosody

- comparison of different schemes and models of prosodic notation;
- definition and validation of spoken language reference units in the interfaces between prosodic, pragmatic and morphosyntactic levels of linguistic annotation;
- analysis of pauses and hesitation phenomena;
- analysis of speech rate.

3.1.3. Morphology

- study of morphological component in relation with phonic reduction processes;
- analysis of morphological properties of syntactic heads of NP's and VP's;
- analysis of the most frequent suffixes in spoken Italian lexicon.

3.1.4. Syntax

- analysis of word order variation in regional varieties of spoken Italian;
- analysis of "microsyntax" in spoken Italian, conceived as the grammar of constituents below the phrase or as "subtle grammar";
- analysis of coordination and subordination relationships;
- analysis of argumental structure of nominal and verbal lexical items;
- syntactic description of the clause in different types of spoken texts;
- detection of standard measurements for the syntax of spoken texts.

3.1.5. Pragmatics

- topic/comment analysis;
- deixis analysis;
- anaphoric annotation of co-reference between all referring nominal and pronominal expressions.

3.1.6. Lexicon

- analysis of the lexicon used in political-parliamentary discourse;

3.1.7. Discourse and conversational analysis

- analysis of linguistic markers of cognitive and conversational processes;
- analysis of dialogic repetition;
- analysis of boundary phenomena between Grammar and Interaction;
- analysis of turn taking mechanisms;
- analysis of "framing" phenomena;
- analysis of communicative interaction and their role in the acquisition of Italian as foreign language.

3.1.8. Semantics

- analysis of semantic properties of syntactic heads of NP's and VP's.

3.1.9. Diachrony of spoken language:

- morphological, syntactic and textual features in diachronic varieties of spoken Italian.

3.1.10. Italian as foreign language

- analysis of the characteristics of Italian spoken by non-native speakers;
- analysis of the characteristics of Italian spoken by native speakers with non-native speakers;
- study of cognitive and semiotic parameters of the acquisition process of Italian by non-native speakers.

3.1.11. Spoken language and mass-media

- collection of an audio/video data base extracted from television news of the past years ('60s-'70s);
- collection of an audio/video corpus of news read by professional speakers;
- segmental, prosodic, and mimic-gestual analysis of the corpora;

3.1.12. Language disorders

- segmental and suprasegmental analysis of speech of hypoacoustic children with both traditional, digital hearing aids and cochlear implants;
- speech analysis in follow up of deaf children with cochlear implant;
- pragmatic and prosodic analysis of subjects with speech disorders.

3.2. Applied and computational linguistics

An important part of the project is devoted to validate or develop applications in the treatment of spoken language data, such as tagging and annotation software programs for the analysis of prosodic, morpho-syntactic, lexical and semantic levels, and for a functional and friendly access to database queries. This part of the project concerns the following scientific domains:

3.2.1. Computational linguistics

- morphological automatic tagging;
- automatic lemmatization;
- data mining analysis on linguistic data derived by various Italian spoken corpora;
- multi-modal and multi-file annotation, access and visualisation methods.

3.2.2 Speech technology

- automatic prosodic analysis.
- technologies for multilevel analysis of speech data;
- speech recognition and synthesis.

3.3. Methods and analysis procedures

This section is deeply intertwined with the previous sections, because it involves all the steps of spoken language analysis. The great increase of studies on spoken language data at national and international level created the need of standard methods and analyses to guarantee the widest comparability of results. The standardization process is concerned with both the accessibility issue and the representation issue. The main products of this section will be the creation of national protocols for collecting data and analysis procedures specifically designed for Italian spoken data according to international standard. All the research units will contribute to achieve this goal.

3.4 Training

An intensive training program has been undertaken. Most of the economic resources in the project has been devoted to research grants. One of the main aim of the project is in fact to create a national research task force of young graduates or postgraduates, who will have the opportunity to work with competent scholars in a comprehensive program of linguistic research.

4. Future developments

The site design will continue through the further project phases. An intensive testing of all site functionalities will be operated Links will be established between the site and other initiatives connected to resources already available in the national panorama. Tools, protocols, coding procedures and data analysis methodologies will be definitely refined and published in the web-site.

As far as all the scientific thematic areas listed in section §2, the following objectives should be achieved:

- corpora transcription and labelling;
- standardisation and validation of base research hypotheses and analysis methodologies;
- definition of standard for corpora measurements;
- validation of qualitative, quantitative and statistical analyses;
- comparison among different levels of grammatical analyses (i.e. syntax, pragmatics and prosody, lexicon and semantics, phonetics and morphology etc.);
- comparison among diachronic, diatopic and diaphasic analyses.

5. Acknowledgements

This paper and the whole project "Parlare Italiano" have been funded (PRIN2004) by the Italian Ministry for University and Scientific Research (MIUR). The project "Parlare Italiano" is carried on by following institutions:

Università di Salerno (national coord.) - Dipartimento di Studi Linguistici e Letterari (coord. Miriam Voghera). *Università di Firenze* - Dipartimento di Italianistica (coord. Emanuela Cresti). *Università di Napoli Federico II* - Dipartimento di Filologia moderna (coord. Francesca Dovetto); Dipartimento di Scienze Fisiche NLP group (coord. Francesco Cutugno); Dipartimento di Neuroscienze sez. Audiologia (coord. Elio Marciano). *Università di Napoli "L'Orientale"* - Dipartimento di Studi dell'Europa Orientale (coord. Antonella Giannini). *Università di Roma "La Sapienza"*- Dipartimento di studi filologici, linguistici e letterari (coord. Tullio De Mauro). *Università di RomaTre* - Dipartimento di Linguistica (coord. Raffaele Simone). *Università per Stranieri di Siena* - Dipartimento di Scienze Umane (coord. Massimo Vedovelli). *Università di Torino* - Dipartimento di Filosofia (coord. Carla Bazzanella).

6. References

Biber, D. (1995), *Dimension in Register variation. A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

CLIPS: http://www.clips.unina.it

Cresti, E., Moneglia, M. (2006) (eds), *C-ORAL-ROM. Integrated reference corpora for spoken romance languages*. Vol. 1+ DVD, Amsterdam: Benjamins.

Crocco, C., Savy, R., Cutugno, F. (2003) (eds), *API: Archivio del Parlato Italiano*. DVD-rom, Prodotto e distribuito da CIRASS-Università degli Studi di Napoli Federico "II".

D'Achille, P. (1990). Sintassi del parlato e tradizione scritta della lingua italiana, Roma, Bonacci, 1990.

Gibbon, D., Moore, R., Winski, R, (1997) (eds.), *The handbook of Standards and Resources for Spoken language Systems*, Berlin: Mouton de Gruyter.

Halliday, M.A.K. (1989). Spoken and Written Language, Oxford University Press.

Sornicola R. (1981), *Sul parlato*, Bologna: Il Mulino.

Voghera, M. (1992), *Sintassi e intonazione dell'italiano parlato*. Bologna: Il Mulino.

Voghera, M. (2000), Les théories linguistique et le parlé. In Emglebert, A., Pierrand, M., Rosier, L., Van Raemdonck, D. (eds), *Contacts interlinguistiques*, Niemeyer, vol. IX, pp. 419- 422.