VoLIP: a linguistic resource for the study of variation in the Italian language

Miriam Voghera, Francesco Cutugno^,Claudio Iacobini*, Renata Savy**
*Università di Salerno ^Università di Napoli "Federico II"

VoLIP (Voce del LIP) is a linguistic resource which matches the audio signal files with the orthographic transcriptions of the samples of the LIP Corpus and allows the search of the corpus according to sociolinguistic as well as lexical and morpho-syntactic criteria.

The LIP Corpus

The LIP Corpus was collected in the early 1990s to compile a frequency lexicon of spoken Italian (T. De Mauro, F. Mancini, M., Vedovelli, M. Voghera, *Lessico di frequenza dell'italiano parlato*, Milano Etaslibri 1993) and its size was tailored to produce a reliable frequency lexicon for the first 3,000 lemmas. Therefore, it consists of about 500,000 word tokens for 60 hours of recording.

The corpus represents diafasic, diatopic and diamesic variation.

As far as the diafasic variation is concerned, texts are divided in five groups: A) face-to-face conversations; B) telephone conversations; C) bidirectional communicative exchanges with constrained turn-talking alternation, such as interviews, debates, classroom interactions, oral exams, etc.; D) monologues, such as lectures, sermons, speeches, etc.; E) radio and television programmes. The texts in groups A and B belong both to formal and informal registers, while C, D, E texts are mainly recorded in public contexts, which select formal registers.

As far as the diatopic variation is concerned, the texts were collected in Milan, Rome, Naples and Florence. The first three cities were chosen according to their geographical position as well as to the number of inhabitants, as Rome, Naples and Milan are the most populated Italian cities. Florence was chosen because of its great relevance in the linguistic history of the Italian language.

While the number of samples is variable, the corpus presents a balanced total number of words per city and per diaphasic situation, as reported in Table 1.

	Face-to-face conversations	Telephone conversations	Interviewes, debates, meetings	Monologues	Radio/TV	Total
Milan	25,000	25,000	25,000	25,000	25,000	125,000
Florence	25,000	25,000	25,000	25,000	25,000	125,000
Rome	25,000	25,000	25,000	25,000	25,000	125,000
Naples	25,000	25,000	25,000	25,000	25,000	125,000
Total	100,000	100,000	100,000	100,000	100,000	500,000

Since the corpus was originally collected for lexical purposes, the recording conditions and the acoustic quality of the sessions differ. The quality scale extends from high levels of clarity of signal to low levels.

The structure of VoLIP

The VoLIP provides all the samples of the LIP corpus in wav files (Windows PCM, 22050Hz. 16 bit) in addition to:

1. session metadata in IMDI format;

2. the original orthographic transcription, already published in De Mauro et al. 1993, in TXT files.

The queries

Two kinds of queries are possible: A) by textual and register variables, as registered in metadata annotation; B) by lexical and morpho-syntactic criteria, as derived from both the frequency lexicon and the parts of speech parsing. The two kinds of queries can be crossed.

The metadata entries are the following: city, actor sex, genre, subgenre, subject, interactivity, planning type, social context, event structure, channel.

All the queries have as output the orthographic transcriptions matched with audio-files.

1. Metadata search results in all the texts presenting the requested features; metadata queries can be crossed with lexical and morpho-syntactic queries.

2. Lexical and morpho-syntactic search results in all the texts presenting the requested item (word form or lexeme) and the specific item within a preceding and subsequent portion of time. Each requested lexeme, word form or part of speech is provided with the frequency of occurrence per city and per register.