VOLIP: a corpus of spoken Italian and a virtuous example of reuse of linguistic resources

Iolanda Alfano*, Francesco Cutugno°, Aurelio De Rosa*, Claudio Iacobini*, Renata Savy*,Miriam Voghera*[1]
*Dipartimento di Studi Umanistici – Università di Salerno
°Dipartimento di Ingegneria Elettrica e Tecnologie dell'Informazione – Università di Napoli Federico II
{ialfano, aderosa, ciacobini, rsavy, voghera}@unisa.it, cutugno@unina.it

Abstract

The corpus VoLIP (The Voice of LIP) is an Italian speech resource which associates the audio signals to the orthographic transcriptions of the LIP Corpus. The LIP Corpus was designed to represent diaphasic, diatopic and diamesic variation. The Corpus was collected in the early '90s to compile a frequency lexicon of spoken Italian 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 speech materials belong to five different text registers and they were collected in four different cities. Thanks to a modern technological approach VoLIP web service allows users to search the LIP corpus using IMDI metadata, lexical or morpho-syntactic entry keys, receiving as result the audio portions aligned to the corresponding required entry. The VoLIP corpus is freely available at the URL http://www.parlaritaliano.it.

Keywords: speech corpora, Italian, IMDI metadata, speech/text alignment

1. Introduction

In this paper we present VoLIP, a linguistic web resource based on a set of 60 hours of spoken Italian.

Corpus creation and annotation requires costs rarely affordable nowadays, and it is well known that resource creation is far from being a trivial achievement and, as a consequence, demand for reusable large corpora is on the rise. In this view, we approached the design of VoLIP, a speech corpus whose main aim was to fully employ and empower a previously available linguistic resource, not completely exploited in its former uses.

Starting from the audio files collected in the '90's for the LIP corpus, the VoLIP turned them into a queryable and structured speech database thanks to a modern technological approach. We provided VoLIP with new features and analysis tools added as web services (see also Voghera et al. 2013). Furthermore the dataset is open to the entire scientific community for free, with the unique constraint to be accessed through the *parlaritaliano* portal. This portal resumes most of the activity of our interdisciplinary group in the last decades in the field of speech corpora collection, classification and public (fully accessible) distribution. The materials available through the portal are thought both for basic research and for technological application. The activities of the *parlaritaliano* portal have been previously described in Voghera & Cutugno (2006) and in Voghera (2010). The present paper is divided into four parts.

In the first part, we will describe the process of the digitalization of the audio-tapes and the organization and cataloguing of the speech material according to the text typology criteria originally used in the LIP.

In the second part, we will illustrate the application we made of the IMDI metadata annotation (Broeder et al. 2001) on both text and speech materials available in the VoLIP corpus.

In the third part, we will describe two techniques that have been applied to the corpus in order to automatically align the speech and the text.

In the fourth part, the main linguistic features available in the web service will be treated.

As already said, the VoLIP corpus is freely available at the URL http://www.parlaritaliano.it. It can be accessed following the link at VoLIP in the left box in the home page.

2. From LIP to VoLIP

As we already said, this adventure starts with the collection of speech materials between 1990 and 1992. These audio recordings led to the production of a frequency lexicon of spoken Italian (*Lessico dell'Italiano Parlato* – LIP - De Mauro et al. 1993).

In order to collect speech as most naturally as possible, audio was analogically acquired using hidden tape recorders and speech material was considered at maximum degree of spontaneity. Recorded people's privacy and similar ethical problems were solved during the transcription process anonymizing all occurrence of names, personal data and any further similar sensible data.

As it is well known, the LIP lexicon has been worldwide considered one of the most important case of study for spoken language and it has been considered as a gold reference standard in most studies on spoken Italian in the last decades.

[1]Authors appear in strict alphabetic order

City	Face-to-Face Conversations	Telephone calls	Debates-Interviews	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

Table 1: Dimensions of the LIP corpus

The LIP Corpus was designed to represent diaphasic, diatopic and diamesic variations. It was collected in order to compile a frequency lexicon of spoken Italian and its size was tailored to be reliable for the first 3,000 lemmas.
As far as the diaphasic 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 programs.
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 in accordance with their geographical position as well as 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.
The LIP corpus consists of about 500,000 word tokens for 60 hours of recording. Table 1 shows the approximate number of words in each corpus portion according to a subdivision on both diatopic and diaphasic dimensions.
Since the LIP corpus was acquired in spontaneous conditions and the informants were unaware of being recorded, the acoustic quality varies, and in some cases is of a low level. Recordings were then transcribed, lemmatized, and, finally, the LIP (De Mauro et al. 1993) was published. LIP has then been considered as a reference as frequency lexicon for forms and lemmas of spontaneous Italian. The lexicon and the texts (transcriptions) used for its extraction were made digitally available together with the publication of the related volume (De Mauro et al.1993) while the audio recorded were put apart in sight of a future reuse.
A first reuse of LIP was made by the University of Graz (BADIP: http://badip.uni-graz.at/it/) who designed a web service delivering an access to the LIP both as a lexicon consultation system and as a concordance estimator. Both services were based exclusively on the transcriptions originally published in the book, once again, at that time, there was no attempt to recuperate the audio recording.
In 2009, thanks to a Research Program (FIRB) funded by Italian Ministry for Education, University and Research, it has been finally decided to access the analog tapes and to digitize them as the first step of a project that would have led us nowadays to offer to the scientific community a web service:

- that allows a parallel access to acoustic and textual information;
- improves the searchability of texts and comparability with other corpora, thanks to a new set of metadata in IMDI format (Broeder et al.2001).

Moreover, since the word tokens in the corpus were POS tagged, we designed a new tool of interrogation based on part of speech and text strings searching.

3. Modern and Digital

The process of accessing the original recordings and retrieving them for our purposes was very delicate. Original corpus collection produced an huge amount of recordings. Only a part of this material was effectively used to build the LIP lexicon. After some year the effective correspondence between the employed portion and the larger number of effectively recorded files has been lost. All analogic recordings were transferred on DAT tapes and then digitized for the first time in 1993. To create VoLIP we converted the DAT tapes into wav and mp3 files using a digital to digital converter device. We double-checked the audio compared to the original transcriptions, and in some cases the repeated listening of the speech material led us to revise some of the original transcriptions. Since most of the former linguistic computations and statistics were based on the original transcriptions we decided to offer both versions: when a reformulated transcription is available, the text appears in red bold fonts: positioning the mouse on that portion of the text allows the opening of a pop-up window containing the revised transcription. An XML file set stores logs for these different versions for all the corpus subparts.

4. Metadata

After an accurate study aiming at comparing different metadata standards, IMDI resulted as the most convenient for our purposes: in the process of creating metadata for the VoLIP corpus, the concept of Session, considered as the complete set of resources associated with a speech event, resulted as central. To meet the needs of the project, several sessions have been created in order to have the diatopic distribution reflected into the metadata schema. Thus, we could imagine the metadata system of our corpus as a tree structure whose symbolic root node generates four sub-corpora, one for each different session corresponding to each of the four cities: Milan, Florence, Rome, Naples. From these session nodes further branches lead to leaves containing audio and transcription metadata sets. The tree then continues its development with a set of branches imposed by the standard and according to our policies for annotation which we will discuss further on.
A team of annotators assigned metadata at each corpus file. Each file corresponds to a recording session as originally fixed by the former corpus designers. The main tag classes defined in IMDI and expressly used in our labelling process are:

- Town: for diatopic descriptions;
- Genre: for a preliminary diamesic classification (conversation vs TV/Radio features, etc.);
- Planning type and social context for diaphasic classification;
- Event structure
- Channel: for a finer diamesic classification.

Annotators recurred to the IMDI Editor tool freely available at http://www.mpi.nl/IMDI/ designed and distributed under open license by the IMDI group.

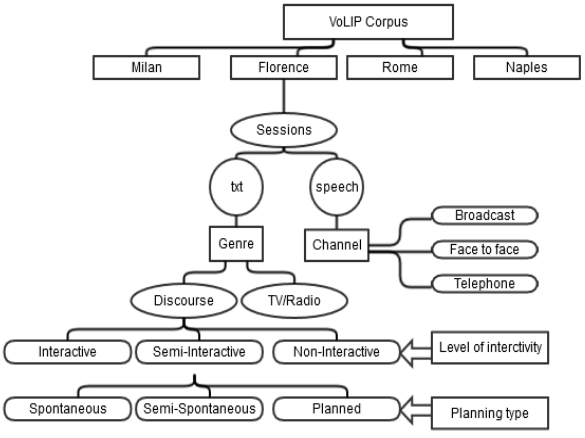

Figure 1: Main metadata annotation schema

In order to establish the inter-annotator agreement and to validate the tagging process, using a blind work distribution process, when assigning workload to every collaborator, we replicated a small portion of data to be annotated to each possible couple of annotators. For each couple we counted the number of tags identically assigned by the two members, and this did not go below 98%. We consider this percentage as reliable for our purposes.
The use of IMDI has direct advantages on the management, the access to the data, and the diffusion of the corpora in use. The IMDI metadata format has been chosen also to increase the comparability with other corpora. VoLIP can be presently queried on the base of its metadata schema directly referring to the XML IMDI file structure.

5. A posteriori text and speech alignment and light POS tagging

5.1 Forced alignment

We will now describe two techniques that have been applied to the corpus in order to automatically align the speech and the text. We start by saying that, technically speaking, the evaluation of the performances of such types of processing is usually based on the comparison between positions of time markers automatically generated with those manually posed. The state of art of forced alignment systems is actually very good and performances are, in many cases, very high, thus guaranteeing a good automatic alignment up to segmental level.
On the other hand the time accuracy we aimed to reach with this operation was clearly lower: we just wished to give to users the possibility of listening to a selected portion of a required text (usually a word, or a list of words distributed over one or more texts). Consequently any query executed on the database, returns an audio portion including the target inserted into a context of two/three words before and after it.
We distinguished two cases:

a) material whose audio quality was above a minimum threshold of acceptability ,
b) material whose audio quality was under the threshold of acceptability because of the decreasing of the signal/noise ratio.

As far as the first case concerns, we used classical techniques for forced alignment typically used in Automatic Speech Recognition (ASR) procedures, with some useful specific corrections for our work conditions. In particular, we revised the grapheme-phoneme conversion in order to uniform it to the orthographic conventions originally adopted for LIP (i.e. accent mark format) and we added an exception list of acronyms to refine the adaptability of the ASR system to our purposes. As already stated before, we will not provide the reader with alignment accuracy results because it is almost evident that, in our case, errors in segmentation of words over time appear always compatible with the dimension of the listening window offered to the website users.
In about 35% of the digitized recordings, speech quality was very poor and it was not possible to use the forced alignment technique.
In this case, we wrote some scripts to ease the manual subdivision in blocks of a few seconds: an operator

listened to the audio files visually following their transcription on the monitor. Every time the listener decided (within an approximate span of about 2 seconds) to add a time marker on the text, she simply pressed the space bar and the system consequently produced an approximately two-seconds long block of speech.
The segmentation is then twofold, in some cases it works at word level granularity and, in other cases, at two-seconds blocks. This choice is compatible with our purposes as the results of both strategies allow the web service to address (or to download) the audio chunks within which the required portion of text (usually one or two words) is systematically included .

5.2 Light POS Tagging

In 1992, during the creation of the original LIP lexicon POS labels have been assigned by a tagger developed specifically for the project. Unfortunately, after more than 10 years, both the software and its outputs obtained giving the text file of the corpus as input got lost. As a consequence we were forced to choose between two alternatives:

1) to use a modern POS tagger off-the-shelf;
2) to try to retrieve a tagging for our texts as similar as possible to the original one.

The former solution, even if more simple and presumably more precise of any other *a posteriori* solution, had the drawback of generating counts and statistics that could significantly diverge from those originally produced. The latter one, on the other hand, if properly actuated, could produce, in the worst case, a limited divergence when previous analysis where compared to the newest. This happens because the procedure will tend to replicate the same errors made in the past without the addition of new sources of errors due to the introduction in the algorithm of new solving strategies. We choose to follow the second line. We just had a simple database in which all lemmas and forms occurring in the LIP lexicon were listed together with their POS label and frequency counts. We used these data to populate a database where, to each lemma, all the related forms were associated, and we afforded the POS tagging problem using a vocabulary based approach. Manual interventions were successively made to solve most of the inevitable ambiguities. A final control revealed that differences between counts made in the original work on LIP concerning a random sample of labels (for a list of these labels see Figure 2) remained systematically under a very low threshold of acceptability.

6. The Web Service

The VoLIP web service is hosted within the *parlaritaliano* web portal. This portal hosts the first national observatory devoted to the study of Italian speech. It presents a broad range of works in several areas of linguistic research concerning many levels of analysis and description; all of them are based on corpora collected in a variety of communicative situations and differently annotated. Data and Tools sections offer numerous resources for the study of spoken Italian.
The VoLIP web service can be queried using various units as index: lemmas, forms, words in audio context, part-of-speech units. All the queries are exposed as web services synthetically available in the same web page; high attention has been paid to the system usability with multiple guide web tools to the user, in order to make the query process simple for a wide range of visitors of different competences. The following two figures give a sketch of how the interface appears:

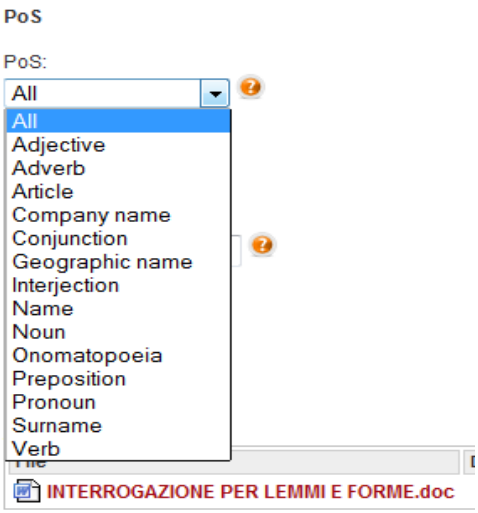

Figure 2: The list of POS tags as they appear in the web application

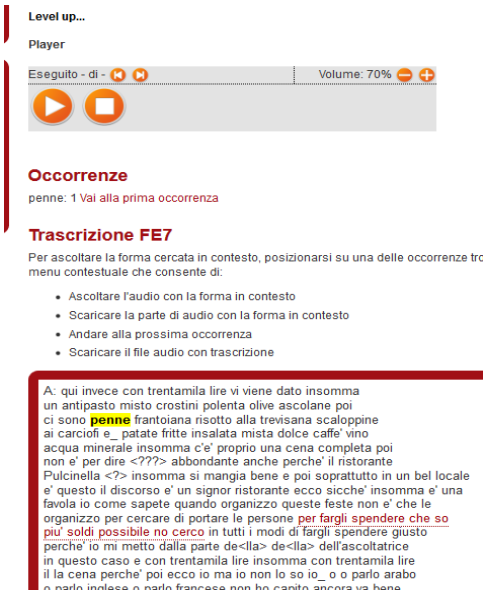

Figure 3 An example of query on our corpus. The request was to find all the occurrences of the form *penne* 'pens' belonging to the lemma *penna* 'pen'. Clicking on the yellow words allows listening and downloading of the selected audio portion.

7. Conclusion

We presented here an example of ecological reuse of a linguistic resource. Renewing its vitality, we modified the nature of the LIP corpus, tailored to extract a frequency lexicon of spoken Italian, formerly accompanied with a

very light POS labelling.
In its new form, VoLIP opens to phonetics and phonology, permits multilayered analysis; the new metadata tagging has enriched the kind of research on these data, concerning the diamesic, diaphasic and (some aspects of) pragmatic variation.
The VoLIP web service is online since summer 2013 and is constantly consulted by the Italian speech community.

Acknowledgements

This work has been supported by MIUR in the frame of the FIRB Project 2009 – 2012 (RBNE07WXMS): "*Universo Italiano. Perdita, mantenimento e recupero dello spazio linguistico e culturale nella II e III generazione di emigrati italiani nel mondo: lingua, lingue, identità. La lingua e cultura italiana come valore e patrimonio per nuove professionalità nelle comunità emigrate*"
Authors wish to thank Rossella Ianniciello, Annachiara Varriale and Debora Vena who took part in the annotation processes.

8. References

Broeder, D., Offenga, F., W. Don, Wittenburg, P., (2001). The IMDI Metadata set, its Tools and accessible Linguistic Databases, *Proceedings of the IRCS Workshop on Linguistic Databases*.

De Mauro, T., Mancini, F., Vedovelli, M., Voghera, M. (1993). *Lessico di frequenza dell'italiano parlato*, Milano, Etaslibri.

Voghera, M. (2010). *Parlare italiano: towards a multidimensional description and a multidisciplinary explanation*. In Pettorino, M., Giannini, A., Dovetto, F.M. (eds.), *La comunicazione parlata 3*, Napoli, Università degli Studi di Napoli L'Orientale, 603-617.

Voghera, M., Cutugno, F. (2006). An observatory on Spoken Italian linguistic resources and descriptive standards. In: *Proceedings of 5th int. conf. on Language Resources and Evaluation*. LREC2006 24-26 Maggio 2006. Genova, ELRA, 643-646.

Voghera, M., Iacobini, C., Cutugno, F.,Savy, R., Alfano, I., De Rosa, A.(2013). Il VoLIP una risorsa per lo studio della variazione nel parlato della lingua italiana. In *Atti del XXVII Congresso internazionale di linguistica e filologia romanza Nancy 15-20 luglio 20133 volumes. Strasbourg : Société de linguistique romane/ÉLiPhi.*