SegWin: a Tool for Segmenting, Annotating, and Controlling the Creation of a Database of Italian Spoken Varieties

Mario Refice, Michelina Savino, Marco Altieri, Roberto Altieri

Department di Elettrotecnica ed Elettronica Politecnico di Bari Via Orabona, 4 - 70125 Bari - ITALY refice@poliba.it, esavino@poliba.it

Abstract

A number of actions have been recently proposed, aiming at filling the gap existing in the availability of speech annotated corpora of Italian regional varieties. A starting action is represented by the national project AVIP (Archivio delle Varietà di Italiano Parlato, Spoken Italian Varieties Archive), whose main challenge is a methodological one, namely finding annotation strategies and developing suitable software tools for coping with the inadequacy of linguistic models for Italian accent variations. Basically, these strategies consist in adopting an iterative process of labelling such that a description for each variety could be achieved by successive refinement stages without loosing intermediate stages information. To satisfy such requirements, a specific software system, called SegWin, has been developed by Politecnico di Bari, which:

- "guides" the human transcribers in the annotation phases by a sort of "scheduled procedure";
- allows incremental addition of information at any stage of the database creation;
- monitors/checks the consistency of the database during every stage of its creation

The system has been extensively used by all the partners of the project AVIP and is continuously updated to take into account the project needs. The main characteristics of SegWin are here described, in relation to the above mentioned aspects.

1. Introduction

The need of spoken corpora, as part of the general framework of "language resources", for a given language has been widely recognised, pointing out the advantages for both linguistic and more application-oriented research. The methodological and technical problems related to such an effort have been also described and suitable solutions have been found, even though room for further improvements is still largely available (Zampolli, 1998).

In all cases, the very basic question consists in defining what is the language the corpus intends to be representative of. In building a speech corpus, the usual language reference is assumed to be the standard variety, which is taken to be representative of the language spoken in a country. Yet unlike other languages like English or German, for which a standard version does exist, Italian language presents some peculiarities for which one definition is not possible.

For historical reasons, the process of defining a unique national model of the Italian language has been perceived and successfully achieved only for its written form, but not for the spoken one (De Mauro, 1963). The so-called "standard" Italian (derived from the "cultivated" Florentine variety, fiorentino colto) is an abstract reference used only by actors or professional speakers, being the regional varieties the ones actually spoken in the country. Despite that, traditionally most of the phonetic/phonological studies and technological applications have been dealt with the abstract model, whereas linguistic descriptions (both at segmental and suprasegmental levels) of regional varieties are still inadequate. As a matter of fact, all spoken corpora of Italian already available are of a strong applicationoriented nature, thus presenting the usual characteristics in terms of:

a) the speech material collected and analysed (consonant/vowel sequences in various combinations, isolated words, phrases, and short passages)

- b) the speaking style (read)
- c) the type of language ("standard" variety spoken by professional speakers, or some Northern varieties as an alternative),

where the reference to regional accent variation models is not a basic requirement. Only recently, some attention has been paid to regional accents, even though in the above mentioned terms and for the cited application fields - for example, COLLECT (developed by CSELT), SIVA (developed by FUB) - whereas corpora of spontaneous or semi-spontaneous dialogue-based speech of regional varieties - and consequently also the related descriptive models - are still missing. In the last years, a number of actions at a national level have been proposed aiming at filling such gap. A starting action is represented by the national project AVIP (Archivio delle Varietà di Italiano Parlato, Spoken Italian Varieties Archive), whose main challenge is a methodological one, namely finding annotation strategies and developing suitable software tools for coping with the above mentioned problems in building up a database of Italian accent variations.¹

2.The AVIP Project

The ongoing Italian project AVIP aims at building a dialogue-based spontaneous speech corpus of three spoken Italian regional varieties, namely those of Bari, Naples and Pisa².

The importance of having a corpus of this kind has been already pointed out also for other languages like German (Burger & Schiel, 1998), both from the side of linguistic

¹ Although reference is made to the project AVIP, the responsibility for the ideas expressed in this paper belongs to the authors only.

² Partners of the AVIP project are: Scuola Normale Superiore di Pisa (Coordinator), CIRASS-Univ. di Napoli "Federico II", Istituto Universitario Orientale di Napoli, Univ. del Piemonte Orientale, Politecnico di Bari.

research (to meet the need of regional varieties linguistic descriptions) and from the technology-oriented one; for example, all the speech recognisers already available for Italian (based on the standard abstract model) sometimes fail when dealing with unexpected regional pronunciations. Moreover, the availability of spontaneous dialogue-based speech data allows a better tuning of the dialogues systems to Italian accent variations.

AVIP corpus speech material consists in 22 kHz recorded semi-spontaneous dialogues. The dialogues were elicited using the Map Task technique (Anderson et al., 1991), involving verbal co-operation (via auditory channel only) between two participants, each having a map, with the aim of transferring as accurately as possible a given route from one map to the other. In order to reproduce typical real communication situations (e.g. misunderstandings), a number of discrepancies in placement and positioning of the landmarks on the maps are present. Some control of speech productions was exercised also on intonation contours, by selecting landmark names containing mainly sonorants, and being characterised by a variety of word stress patterns.

As pointed out also by EAGLES recommendations (Gibbon et al.,1998), in building and labelling a corpus it is fundamental to have in mind the kind of applications it is intended for. The AVIP corpus, however, is intended to cope with the needs of different kinds of applications. Therefore, the main corpus constituents are:

orthographic transcriptions of dialogues

- textual/discourse annotation
- word-by-word graphemic annotation
- word citation phonological annotation
- broad segmental annotation
- narrow segmental annotation
- prosodic annotation

where the unit of analysis is the speech turn.

The textual/discourse annotation has been carried out by one of the partners by making use of a specific tool, whose description is not included in this paper. All the remaining annotation levels have been performed by using SegWin, which also manages the whole database.

2.1 Annotation tiers

At the moment, only the following 6 tiers have been considered (for ease of reference, acronyms in capital letters throughout the paper refers to the annotation tiers):

- WRD orthographic transcription of both lexical and non-lexical speech events;
- PHM citation word phonemic transcription. It is a word-by-word segmentation which provides citation-form phonemic transcription by automatically applying a set of "standard" Italian grapheme-to-phoneme rules (including lexical stress assignment). This level of annotation is useful for two main reasons: 1) it plays the role of "reference model" in the database with respect to the "variety-oriented" labelling tiers. The linguistic purpose of the "citation form" tier is that of having a reference, common to all the considered varieties, which will allow the comparison of the different phonetic realisations in the same phonetic context, by means of some

suitable statistical analysis. 2) it is the only tier carrying information about lexical stress position.

- PHB broad-phonetic transcription. This is very similar to Barry & Fourcin's (1992) "broad-phonetic" description level, i.e. it is substantially phonological with the addition of a sub-set for the description of some variety-specific phenomena, like variation in the lexical distribution of open and closed vowel set (/e/ vs. /E/ and /o/ vs. /O/), variation in the acoustic realisation (voiced vs. unvoiced) of intervocalic /s/, and presence vs. absence of "raddoppiamento sintattico" (syntactic doubling), i.e. consonantal gemination across word boundaries. These phenomena have been considered so far as the very basic ones in distinguishing among regional varieties: it may be said that they are part of the intuitive knowledge (together with prosodic characteristics, of course) Italian native speakers seem to use in "recognising" the regional origin of Italian interlocutors. This annotation tier, then, contains a first-step, basic level description of the three Italian varieties under consideration;
- PHN narrow phonetic transcription, conceived as a successive refinement of PHB, where some phenomena are described in more detail, basing also on quantitative statistics performed on PHB annotated level;
- TON "phonetic-oriented" prosodic transcription, which is inspired to the intonation labelling system INTSINT (Hirst & Di Cristo, 1998);
- AUT "phonological-oriented" autosegmental-metrical (Ladd, 1996) ToBI-like prosodic transcription, with some language/variety specific adaptations with respect to the standard ToBI labelling conventions (Beckman & Ayers, 1993).

Given the inadequacy of linguistic description of Italian varieties, expecially at prosodic level, it was decided to implement both the mentioned different descriptive models (relating to TON and AUT tiers), in order to allow testing and comparison between them and across varieties. For Bari Italian variety, in particular, since background work has been already carried out within the autosegmental-metrical framework by analysing Map Task dialogues speech material (Grice & Savino, 1995 and 1995a, Refice et al., 1997), level AUT has been annotating first.

The phonetic alphabet set used at the segmental level is SAMPA as suggested by EAGLES guidelines (Gibbon et al., 1998).

Further annotation tiers can be added without affecting labelling procedures and/or database structure taking advantage of the multi-files structure of the database produced by the SegWin system.

3. SegWin

The system described here is written in C and runs on PC Windows environment, taking advantage of its supporting

facilities.

As stated above, the inadequacy of descriptive linguistic models of Italian regional accents led us to adopt a strategy in labelling the corpus consisting basically in an iterative labelling process such that a description for each variety may be achieved by successive refinement stages without loosing intermediate stages information. One of the main characterictics of SegWin, then, is that of "guiding" the human transcriber in the annotation phases by a sort of "scheduled procedure", by which the cited iterative process is guaranteed.

Moreover, SegWin is not only a tool for assisting human trascribers in labelling procedures, but also for building the database. In this respect, the system monitors and checks the consistency of the database (at all levels) during every phase of its creation.

Having in mind future enlargement of the corpus both in terms of types of information (i.e. further annotation tiers and temporal information) and number of Italian varieties, SegWin architecture allows incrementing the number and types of information to be added at any stage of the database creation.

Details of all the above mentioned features are described in the following sections.

4. SegWin Basic Functionalities

The system basic functionalities are:

- a graphic interface for human interaction and labelling process;
- an automatic segmentation of the signal;
- some basic algorithms for computing the acoustic parameters which are useful for the human decision as to the correctness of the segmentation;
- an automatic generation of all the files containing the information related to the speech turn under examination (annotation tiers, acoustic parameters, atemporal information);
- a management support which guarantees the database integrity, checking the consistency of the data and not allowing actions which might corrupt it.

4.1 Graphic interface

The operator screen is divided into two main parts: one showing the signal with the related acoustic parameters, and the other one displaying the mentioned 6 tiers, namely WRD, PHM, PHB, PHN, TON and AUT. The display of these tiers is under operator control: they may be singularly activated or deactivated. Other common facilities, such as playing the entire file or parts of it, zooming, displaying of the computed acoustic values, positioning of the cursor (either as frame number and msec), duration of a selected portion of the signal, F0 values etc. are available and shown on the screen.

In each annotation tier, the operator is required to set the temporal marker related to the right boundary of a given segment, whereas the corresponding label is inserted through an automatically opened input window.

Figure 1 shows a snapshot of an almost completed annotated speech turn.

There is an almost fixed sequence for filling the annotation tiers, which is controlled by the system and shown by a schedule window on the screen. Details about such a procedures will be discussed in the following.

4.2 Semi-automatic segmentation of the signal

Upon reading in the signal, the system automatically computes all the acoustic parameters (i.e. spectrum, energy contour, zero-crossing, fundamental frequency, etc.) if they have not been computed already for the same speech file, records them into the corresponding files, and displays them on the screen.

According to a set of rules, the system makes a proposal about the possible segmental boundaries; the human operator can simply confirm the system proposals, insert new markers in different positions or delete the proposed markers. In Figure 1 the small dotted lines at the bottom of the screen represent the boundaries proposed by the system, while the continuous vertical lines represent those chosen by the human operator (the two lines are obviously ovelapped in cases of matched choice). The final choices are automatically recorded as temporal markers into the files containing the annotation information. In a second stage, they will be used also for adapting and/or improving the set of segmentation rules by an off-line procedure. The details of such a procedure as well as the structure of the rules are beyond the scope of the present paper and will not be discussed here.

4.3 Labelling schedule

A schedule window guides the operator through a given "labelling path". The operational sequence of labelling is shown in Figure 2. In this three levels scheme each bubble represents, by the name of the corresponding tier, the annotation operational procedure. The first level of labelling is the broad phonetic (PHB), while the PHN tier has to follow, being a refinement of the former.

Initially, the content of PHN tier is simply a copy of PHB (as it is shown in Fig.1). Once the PHB tier has been completed and "locked"³, the operator can modify the PHN labels, as far as s/he is confident about a more detailed label to be assigned to any given segment. Such a decision may also result from statistical analyses previously performed on the fully annotated PHB level. Such procedure is an operational strategy for deriving hypotheses about Italian regional accent variation models. With this respect, this strategy follows the same principles adopted, for example, by Kohler et al (1995).

As an alternative labelling path, the operator can start with word-by-word segmentation (WRD) and related grapheme-to-phoneme conversion (PHM), the two being performed contemporarily.

In filling WRD and PHM levels, markers are allowed by the system to be positioned only on existing corresponding markers on PHB level (basing on the principle that a segment boundary must coincide with a word boundary). This is one of the features, which guarantees the consistency of the final database.

Once WRD and PHM tiers have been fully annotated, they have to be "locked" by the human transcriber in order to be able to go on with the prosodic labelling.

Even though the last mentioned labelling actions (of tiers TON and AUT respectively) can be performed, in principle, both after or before the WRD/PHM ones, the

³ "Locking" an annotation tier means that it cannot be modified any more, unless the related annotation file is deleted.

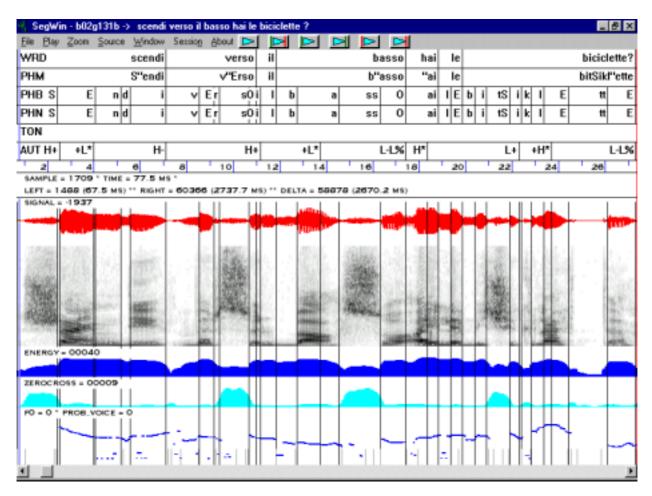


Figure 1: Snapshot of SegWin annotated speech turn

system requires them to be preceded by the word-by-word segmentation, in order to take advantage of information displayed on the WRD tier which can be useful for prosodic labelling (typically, word boundaries coinciding with prosodic edges). Of course, TON and AUT labellings are completely independent of each other.

Arrows in Figure 2 represent all the above mentioned operational dependencies.

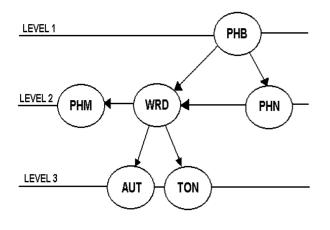


Figure 2: Operational sequence of labelling actions in SegWin

In order to facilitate the positioning of the temporal markers on the dependent tiers, and ensuring the perfect time alignment with the already existing (upper or lower levels) markers, the system provides a special "snap" functionality. When the operator puts the mouse cursor close to the target point on the screen, the system automatically "snaps" the time marker to the closest one.

Two different schemes of dependencies, concerning labels and time markers modifications respectively, are adopted.

4.3.1 Label modification dependency

Two cases of dependency in label modification procedures are considered: one between the PHB-PHN pair and the other between the WRD-PHM pair.

PHB 🗲 PHN

As already pointed out, we consider PHN as a refinement of the PHB tier, then the operator is allowed to modify PHN annotations both in terms of labels, and marker position. Every modification concerning existing temporal markers in PHN is then automatically recorded into the PHB tier. Of course, this is not applied to labels (and also to new time markers inserted only at PHN level).

The same dependency scheme applies to the WRD and PHM pair.

WRD 🗲 PHM

The WRD labels are automatically extracted from the turn orthographic transcription file. The operator can simply confirm the selection or choose another word by using a special browser. Modification of existing words is also allowed. Once the graphemic word has been selected, the corresponding citation form phonemic transcription is automatically provided by the system, by applying a set of rules. Each label modification in the WRD tier is therefore automatically applied to the PHM, whereas any modification in the PHM tier does not affect the WRD one. Changes in PHM tier are allowed in cases where the grapheme-to-phoneme conversion routines fail (for example, in cases like exceptions in stress assignment, which are quite frequent in Italian, etc.).

4.3.2 Time markers modification dependency

A different dependency scheme is implemented for the positioning of segments boundaries.

The obvious reason for this different strategy is that during the refinement phases it may be necessary, for example, to adjust a time marker position in a more precise way. All modifications made on time markers at a lower level of annotation have a direct influence on the time alignments of all the higher level labels.

Once the PHN annotation tier has been initially created by the system (as a copy of the already filled PHB), it becomes the hierarchical lowest level of labelling, as far as time markers modifications are concerned. This means that every time a temporal marker in the PHN tier is moved (the system does not allow the new position overcross the neighbouring ones) or deleted, such modification is automatically recorded into all the remaining tiers (the number of them may vary according to the "labelling path" followed). Then, if such marker refers to a word boundary the WRD and PHM tiers (and possibly TON and AUT, if such boundary coincides with one of the related boundaries) are also affected. Otherwise only PHB annotation file is automatically updated.

In PHN level it is possible to insert new temporal markers (i.e. new labels) for identifying segmental sub-units with respect to the ones in PHB. In such cases, the above described strict dependency scheme is not applied.

4.4 The Generated Files

The general structure of the information to be recorded in a corpus is still a debatable matter. Roughly speaking, the main question concerns the choice between the generation of a unique file containing all the required information, or of several files each containing a specific part of the total information. Both solutions have advantages and disadvantages: having a unique file helps in maintaining the consistency of data, while several files may cause possible discrepancies between them, due to the difficulty of maintaining the consistency among a big set of different files. Moreover, the unique file technology facilitates the physical transfer of data between different sites. On the other hand, having several files does not necessarily require the use of a DBMS for analysing the data and running statistics on them. Two examples, among others, can be cited as representative of these two different approaches, namely: the SAM Project (several files) (Fourcin, 1993) and the Kiel Corpus of Spontaneous Speech (one file) (Kohler et al., 1995).

Since the purpose of this project is to meet the urgent needs of as large scientific community as possible, any simple statistical package, or even any special-purpose home made program, running on any platform, can easily deal with such a data organisation. Also for these reasons, it has been decided to adopt multi-file data organisation, very similar to SAM files structure (Gibbon et al., 1998). All file names are automatically constructed by the system, according to a hierarchical scheme, that is represented in Fig.3. The first level recalls the specific map used in eliciting the dialogue (four different maps have been used so far); the second level numbers the recording session (since several sessions may be performed using the same map); the third level specifies the speaker's role (either instruction Giver or instruction Follower in a Map Task session, whereas L stands for reading sessions); the forth level numbers the speech turn, while the fifth codes the Italian regional variety (Bari, Naples, and Pisa).

A mapping file provides the correspondence between the coded file names and a more mnemonic file names convention.

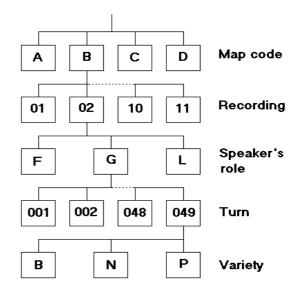


Figure 3: Hierarchical file names coding

Information related to annotation tiers are recorded in separeted files of type .WRD, .PHM, .PHB, .PHN, .TON, .AUT. Additional files, managed by the system but not directly related to the annotation tiers, contain some extra information, like orthographic transcription of each turn, atemporal data related to each recording session (header file), temporal information about beginning-end of each turn within the entire dialogue, etc.

4.5 Database management support

Also the file format is very similar to the one adopted in SAM, including the labels used within each file. Yet differently from SAM, each file includes also the name of the file itself, and a special field containing a control checksum value. In every session, the system makes the same computation on each file and compares its result with that contained in the file itself. Any difference may only be caused by a modification made on the file outside the system environment. In this case, the operator has to delete the corrupted file and to go on with the labelling procedure again, according to the allowed schedule. Thanks to the multi-files organisation, the work to be recovered in these cases does not exceed a single speech

turn. A supporting tool, which runs offline but in the same environment of SegWin, is also available. Such tool allows reconstructing the correct checksum value, in case of emergency.

A set of other software tools provides additional off-line capabilities for cross checking and spotting any possible inconsistency among the files belonging to the same speech turn. Some of them run interactively, that is, once a possible discrepancy is detected, the system asks a human operator to confirm the needed modification to the related files. Just for an example, a specific tool checks the consistency of the automatically derived citation-form phonemic transcription with the one actually inserted in the PHM tier. As a matter of fact, the operator might have modified some of the labels in this tier, with respect to the automatically converted version. In case of mismatch, the tool asks for a confirmation and records this occurrence. Even though such a procedure may be extremely timeconsuming, and cannot ensure a perfect uniformity since different operators may make different choices, it is useful for spotting also the cases in which the graphemeto-phoneme conversion routines need to be improved.

5. Summary

The software system SegWin has been designed and developed by Politecnico di Bari for segmenting, annotating and controlling the construction of a database of Italian regional accent variations, starting from the specific needs of such an effort. It runs on PC Windows environment and has been extensively used within the ongoing national project AVIP for building a semispontaneous dialogue-based database of Bari, Naples and Pisa varieties of Italian. Since the system is continuously updated to satisfy the evolving requirements of the project partners, at the moment it cannot be considered available outside the project. As soon as such requirements will be completely satisfied, and the system considered stable, we hope to be able to make it available to a larger community.

6. References

Anderson A.H., M.Bader, E.G.Bard, E. Boyle, G. Doherty, S.Garrod, S.Isard, J.Kowtko, J.McAllister, J. Miller, C. Sotillo, H. Thompson and R.Weinert (1991). The HCRC Map Task Corpus. Language and Speech, 34 (4), 351--366.

- Barras C., E. Geoffrois, Z. Wu & M. Liberman (1998). Transcriber : a free tool for segmenting, labeling and trascribing speech. Proc. LREC (pp.1373--1376). Granada : ELRA.
- Barry W.J. & A.J. Fourcin (1992). Levels of labelling. Computer Speech and Language, 6,1--14.
- Beckman M.& G. Ayers (1993). Guidelines for ToBI Labelling. Ohio State University.
- Burger S. & F. Schiel (1998). RVG1 A database for Regional Variants of Contemporary German. Proc. LREC pp 1083—1087. Granada: ELRA.
- De Mauro, T. (1963). Storia linguistica dell'Italia unita, Bari: Laterza.
- Fourcin A. (1993). The SAM project. Chichester: Ellis Horwood.
- Gibbon D., R. Moore, R. Winski (1998). Handbook of Standards and Resources for Spoken Language Systems. Berlin: Mouton de Gruyter.
- Grice, M. & M. Savino, (1995). Intonation and communicative function in a regional variety of Italian. Phonus, 1, 19--32.
- Grice M. & M. Savino (1995a). Low tone vs 'sag' in Bari Italian intonation; a perceptual experiment. Proc. ICPhS 95 (pp 658-661). Stockholm:ESCA.
- Hirst D. & A. Di Cristo (1998). A survay of intonation systems. In D. Hirst & A. Di Cristo (Eds.). Intonation systems (pp. 1-44). Cambridge: Cambridge Un. Press.
- Kohler K., A. Paetzold, A. Simpson (1995). From scenario to segment. The controlled elicitation, transcription, segmentation and labelling of spontaneous speech. AIPUK, 29.
- Ladd R.D. (1996). Intonational phonology. Cambridge: Cambridge University Press.
- Refice, M. M. Savino, M. Grice (1997). A contribution to the estimation of naturalness in the intonation of Italian spontaneous speech. in Proc. EUROSPEECH 97 (pp.783-786). Rhodes: ESCA.
- Zampolli A. (1998). Introduction to the First International Conference on Language Resources and Evaluation. In Proc. LREC (pp xv-xxv). Granada: ELRA.