Si ringrazia la «Associazione Amici della Accademia dei Lincei»
per la collaborazione offerta all edizione del presente volume

FINITO DI STAMPARE NEL MESE DI MAGGIO 2018
Antica Tipografia dal *1876* srl – Corso del Rinascimento 24, 00186 Roma
Azienda con Sistema Qualità certificato ISO 9001-14001 OHSAS 18001

ISSN: 0394-0705

ISBN: 978-88-218-1165-4

ACCADEMIA NAZIONALE DEI LINCEI

ANNO CDXV 2018

CONTRIBUTI DEL
CENTRO LINCEO INTERDISCIPLINARE
«BENIAMINO SEGRE»

N. 137

International Workshop

SPEECH AUDIO ARCHIVES: PRESERVATION, RESTORATION, ANNOTATION AIMED AT SUPPORTING THE LINGUISTIC ANALYSIS

(Rome, 18-19 May 2017)

Edited by

AMEDEO DE DOMINICIS

ROMA 2018
BARDI EDIZIONI
EDITORE COMMERCIALE

Indice

Organizing Committee

Brigitte Bigi
Silvia Calamai
Sergio Canazza
Amedeo De Dominicis
Antonio Rodà

Tito Orlandi*

PAROLE DI SALUTO

Mi è gradito portare ai partecipanti il saluto del Presidente dell'Accademia, prof. Alberto Quadrio Curzio, e del Direttore del Centro Linceo, prof. Mario Stefanini, impossibilitati ad intervenire, che hanno espresso vivo interesse per questo Convegno, organizzato nell'ambito delle attività del Centro. Come dice il suo Regolamento, "il Centro costituisce punto di riferimento e promozione per chi faccia ricerche che favoriscano l'interazione fra più discipline" e soprattutto fra le loro diverse metodologie. La sua attività è basata in primo luogo sulla collaborazione di dieci Professori Distaccati per tre anni presso il Centro dalle loro Università, e appunto dalle ricerche del Professore Distaccato Amedeo De Dominicis nasce questo Convegno, nell'ambito della Linea di ricerca detta "Discipline umanistiche e informatica" che ho personalmente curato nella mia qualità di Direttore del Centro fra il 2010 e il 2016.

La struttura e le finalità del Convegno saranno illustrate dal suo ideatore. Io desidero sottolineare come ci si trovi qui in uno dei molti campi in cui le moderne tecnologie e le scienze della natura possano aiutare in maniera estremamente efficace la conservazione, il restauro, e la diffusione della conoscenza dei beni culturali; ma soprattutto nel quale le procedure informatiche portino novità anche metodologiche nelle discipline designate in genere come discipline del testo. Già è rimarchevole notare come poco più di un secolo fa la costruzione di macchine che potevano conservare e riprodurre i suoni abbia mutato l'orizzonte della conservazione testuale, fino allora ristretta alla scrittura o in casi speciali alla memoria umana trasmessa direttamente. Oggi siamo in grado di sottoporre testi parlati a trattamenti tali che permettono, oltre che migliorare la qualità dei reperti, anche di effettuare analisi sofisticate di carattere linguistico. Esse potranno portare risultati nuovi in molti settori, p.es. nel rapporto fra lingua scritta e lingua parlata, nella determinazione dei caratteri dialettali, nella sociologia del linguaggio, e così via.

*Accademia Nazionale dei Lincei - Via della Lungara, 10 - 00165 Roma.

Vorrei tuttavia rimarcare come l'analisi linguistica ad alto livello con procedure informatiche richieda che si passi da uno stadio in cui il testo è conservato in maniera sostanzialmente analogica, in cui le strutture digitali servono solo a permettere la riproduzione del suono, ad uno stadio in cui i singoli elementi significativi, fonemi ed eventuali elementi paratestuali, sono in seconda istanza codificati in modo da ottenerne una rappresentazione paragonabile alla rappresentazione scritta. In questo modo l'analisi linguistica può avvalersi di tutte le strategie finora applicate ai testi scritti, oltretutto con nuove prospettive date dal particolare carattere dei testi in questione. Ma lo stesso passaggio di per sé porta ad interrogarsi sulla varietà di significati che gli elementi del testo hanno nei supporti tradizionali, e su come essi corrispondano a strati diversi di quello che si chiama testo ma in realtà rappresenta la coesistenza di sistemi od organismi concreti o astratti da considerare ciascuno per conto suo e poi nella loro relazione.

Convegni come questo da un lato indicano interessanti soluzioni per affrontare i problemi, ma dall'altro richiamano l'attenzione sulle novità anche metodologiche che stanno venendo dal mondo complesso dell'interdisciplinarità.

Amedeo De Dominicis*

INTRODUCTION

The papers collected in these Proceedings stem from the International Workshop on "Speech audio archives: preservation, restoration, annotation, aimed at supporting the linguistic analysis", held at the Accademia dei Lincei, in Rome (Italy) on May 18-19, 2017.

The Workshop was a considerable success, with eight talks by speakers from three countries.

As is well-known, out there in Europe and all over the world there are many audio corpora, witnessing an extraordinary cultural and linguistic richness awaiting research and cultural heritage exploitation. One needs not spend many words to underline the paramount importance of this assorted repository in terms of cultural heritage preservation. They are an inexhaustible wealth of linguistic variability even within one and the same politically homogeneous territory. It is equally well-known that these different corpora are based on a multiplicity of formats and stored in databases using incompatible standards, which make their interoperability impossible at the present stage. It is thus very important and urgent to develop intelligent tools in order to attain such a crucial goal. During the past 50 years, many resources have been collected and recorded by several national research centers (mostly dedicated to dialectological and sociolinguistic studies). They witness language usage by speakers differing in geographical origin, age and educational status. In quite a few cases, these materials constitute the living evidence of a now disappeared stage of the language.

In recent times, the European scientific community has become more deeply concerned with problems relating to the maintenance and exploitation of oral archives witnessing important cultural layers of language diversity. But if we do not act now this treasure may be lost forever. The reason is simple: a great deal of the most valuable materials has been collected some time back,

*University of Tuscia and Lincei (Italy).

and is still preserved on obsolete types of support, deemed to perish forever before long. This endangers their preservation and hampers their consultation.

To make these cultural assets available to a trans-national circulation, the speech materials have to be annotated, supplied with metadata for on-line searching queries, and – if necessary – restored. Therefore, an interdisciplinary approach developing strong synergies among many disciplines is needed. But the interaction of different skills and competences is often difficult. In fact, in every scientific field the most interesting part is always in the details, and the details are by definition comprehensible only to those who are competent. In order to encourage the interaction between researchers of different disciplines, we asked authors to work on a single, and shared speech corpus (called “Workshop Speech Corpus").

The speech corpora that already exist need to be made interoperable; and to this purpose a great effort must be made to make formats and annotation standards compatible. The same goes for new corpora. Moreover, sometimes the recorded signals are in need of restoration. And the restoration techniques are not good in every way: each technique has its advantages and its disadvantages. It depends on the action that each single corpus calls for. Therefore, to break down the barriers between different scientific approaches is what is actually required.

The papers concerning the creation and the organization of speech audio archives (session 1) are by Francesca Frontini and Silvia Calamai (*Speech audio archives and clarin metadata*), by Luca Lorenzetti, Paolo Milizia, and Giancarlo Schirru (*Collecting spoken data from upper-Southern Italian dialects*), by Emanuela Cresti, Massimo Moneglia, and Alessandro Panunzi (*The lablita Corpus & the language into act theory: analysis of Viterbo excerpts*).

The first paper regards the Italian digital speech resources that are accessible via VLO (*Virtual Language Observatory*) of CLARIN consortium.

The second paper concerns the speech archive of the *Centre for the documentation of dialects of Southern Lazio*. The third one presents data from the LABLITA corpus.

Session 2 (Linguistic annotations) includes the papers by Philippe Martin (*Prosodic annotation in adverse recording conditions*), Amedeo De Dominicis (*On repetitions*), and Brigitte Bigi (*Annotation representation and file*

conversion tool).

The first paper describes WinPitch, the software for acoustic analysis that allows F0 and its annotation to be extracted in adverse recording conditions. The second paper analyses phonetically the repetitions. The third one points up to SPPAS, an automatic annotator that also converts from/to some existing file formats of various software.

Session 3 (Audio digital restoration/speech enhancement) consists of two papers: by Antonio Rodà, Alessandro Russo, Sergio Canazza, and Piero Cosi (*Audio documents restoration as a documentary source in the linguistic research: comparison of instruments*), and by George Brock-Nannestad (*The unrehearsed recording: source-critical optimisation of the sound that was captured*). The former is about a new algorithm to restore speech signals; the latter illustrates the problems of unrehearsed recordings.

Table of contents

SESSION 3 – *Audio digital restoration/speech enhancement*.

Antonio Rodà, Alessandro Russo, Sergio Canazza (Università di Padova) - Piero Cosi (CNR-ISTC, Padova): *Audio documents restoration as a documentary source in the linguistic research: comparison of instruments*.

George Brock-Nannestad (Patent Tactics, Gentofte): *The unrehearsed recording: source-critical optimisation of the sound that was captured*.

The "Workshop Speech Corpus" is available at:
`http://www.lincei.it/files/documenti/Corpus.zip`.

Francesca Frontini[*] - Silvia Calamai[**]

SPEECH AUDIO ARCHIVES AND CLARIN METADATA

Abstract. – The issue of metadata is crucially intertwined with interoperability, accessibility, and re-use of research data. On the other hand, such issue is particularly cumbersome in the case of speech and oral archives, which are often scattered among different public and private institutions. From this respect, the aim of the European Common Language Resources and Infrastructure for Social Sciences and Humanities (CLARIN) is to provide easy access to digital written, spoken, video, multimodal language data and at the same time to offer advanced tools to discover, annotate, and analyse them. The paper presents CLARIN and its Component MetaData Infrastructure (CMDI) and it further discusses the relevant profiles for the description of speech and oral archives. Since Italy has recently joined the CLARIN infrastructure consortium, it additionally analyses the accessibility of Italian corpora via the Virtual Language Observatory (VLO) and via the *Parlare Italiano* web portal.

1. Introduction

Nowadays, collecting speech archives appears to be less difficult than in the past, since the easy availability and accessibility of voice and video recording tools make it easier today to collect them. At the same time, old analogic collections are currently being digitized, to ensure their long-term preservation. Both the collecting of new original archives and the preservation of past archives imply some choices: data need to be described and metadata may vary in scope and format. Consequently, important issues arise as to the correct formats to be used for such data, as well as to the types of descriptors used to describe them and make them searchable. Speech data are being collected and used by several communities and for several purposes (from the development of speech technologies to the study of Oral History), and various standards and best practices coexist.

[*]Praxiling UMR 5267 CNRS - Université Paul-Valéry - Montpellier 3 - Route de Mende - 34090 Montpellier (France).

[**]DSFUCI - Università degli Studi di Siena - Viale L. Cittadini, 33 - 52100 Arezzo.

Terminological issues arise as well. For instance, "speech audio archives" is the label used in the title of the Lincei congress at which this paper was first presented. In the *rationale* of the International Workshop alternative labels also are used: "audio corpora" and "oral archives". In the relevant literature, *oral*, *speech*, and *audio* may be used more or less indifferently with the meaning of structured collection and repository of audio material; nevertheless the different labels refer *not only* to a lexical *question but also to different disciplinary traditions*. Among linguists, *speech archive* is the most frequent label, while among oral historians and anthropologists, *oral archive* is much more used. The label audio archive often covers a broader meaning, including also music. Likewise, *archive*, *corpus*, *collection* are sometimes used as synonyms and sometimes in a mutually exclusive way according to the relative disciplinary practices. Such differences crucially impinge on the metadata *côté*, in the sense that single sets of metadata used in the description of an archive collected within one branch of knowledge may be not fully applicable in case of re-use of the same archive by different scholars in different fields. Thus, the metadata and encoding schema heterogeneity – together with the lack of commonly agreed upon standards for the encoding and the description of recorded speech among diverse disciplinary perspectives – constitute a serious challenge to the accessibility and the re-use of this kind of resource (*e.g.*, [12]). Moreover, only a small part of the metadata is available online to researchers. Finally, a great deal of archives were born analogue and analogue recordings often had only minimal metadata on paper. The image displayed in fig. 1 is paradigmatic: the unique available metadatum of the compact cassette is the name of the Italian village ("Coniale") in which the fieldwork took place. The date of the interview, together with the personal data of the participants, may only be inferred [5].

Linguistically annotated archives are in the majority of cases born digital archives. The annotation of born analogue archives clearly represents a case of re-use of linguistic data which were collected for other purposes: a paradigmatic example is represented precisely by the task suggested in the Call for Papers of the Workshop, asking the participants to carry on "the analy-

Fig. 1. – Photographic documentation from Alto Mugello Archive (Gra.fo project). In Capital the code of the archive (Alto Mugello). The number refers to the progressive number of the cassettes in the archive, while "b" indicates the side of the cassette.

sis of (at least) one of the files in the Workshop Speech Corpus"[1]. The Italian audio files stem from the Bomarzo corpus, a born analogue oral archive collected in Tuscia region in the Seventies [8]: the entire corpus, although fully digitized and described, is not yet accessible to researchers. The Ayoreo (Zamucoan) audio file is born digital and comes from Luca Ciucci's personal archive, which is likewise inaccessible to others. On the contrary, the Occitan file is accessible via the Ortolang repository and has a persistent id hdl:11041/sldr000006 as well as a persistent URL `http://hdl.handle.net/1`

[1]"The authors may also add other corpora to their analysis objects, but it is compulsory to develop a detailed analysis of (at least) one of the audio files of the Workshop Corpus. This constraint concerning the use of at least one of the audio files of the Workshop Speech Corpus stems from our will to provide all authors with a common background, and specially to develop and foster a possible and highly desirable dialog among scholars coming from different fields of the scientific community (linguists, phoneticians, engineers, computer scientists, etc.) during the Workshop". `http://www.lincei.it/modules.php?name=Convegni&file=lista&func=Convegni_edit&Id=1350` (accessed 03.04.2017).

1041/sldr000006, where metadata and bibliographic references can be easily found[2]. Such differences are emblematic of the diffused heterogeneity characterising curation, description and accessibility of oral and speech archives.

2. CLARIN AND CMDI

CLARIN[3] is the Common Language Resource Infrastructure for Social Sciences and Humanities. Its goal is to provide easy and sustainable access to language resources to researchers in the humanities. More specifically, by language resources we intend digital language data, such as corpora, computational lexicons, formal grammars, statistical language models and natural language processing tools. In order to grant "easy and sustainable access" to such data, CLARIN created a network of centers located in the various European member states, and offering language resources repositories, namely digital catalogues and archives where resources can be searched for, accessed and if possible downloaded, prior online identification. In fig. 2 we show an example of a CLARIN repository, where data can be searched for via a faceted search facility, and using various parameters.

In the first place repositories provide access to the resources that are produced in the center that hosts them, but in some cases they also offer hosting facilities, allowing third party users to deposit their data[4]. All repositories are somewhat interconnected, since CLARIN also offers a metacatalogue, the CLARIN Virtual Language Observatory (VLO)[5] (fig. 3) that allows users to query for resources from any of the CLARIN centers at once.

[2]A similar example of best practices for what concerns minority and endangered languages is offered by the Pangloss (LACITO) project. The collection is not only available via its own project page http://lacito.vjf.cnrs.fr/pangloss/index.html, but also described and made searchable within the Cocoon repository. Permanent URLs are also assigned (just as in Ortolang and other similar repositories); a Pangloss entry in Cocoon can be accessed via the following url: http://cocoon.huma-num.fr/exist/crdo/meta/crdo-PHO_FERLUS_3_5_SOUND.

[3]www.clarin.eu.

[4]We shall not go into the details of the CLARIN infrastructure here, referring to comprehensive publications such as [11, 14].

[5]vlo.clarin.it (see also [15]).

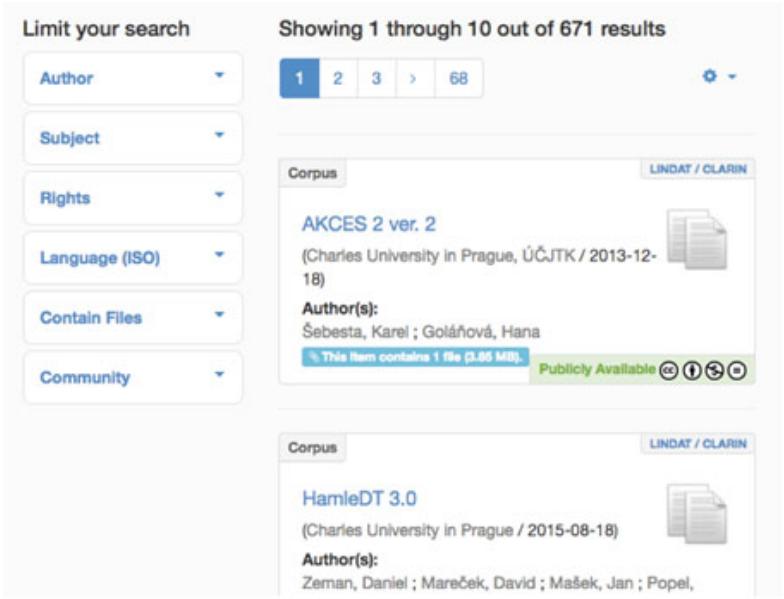

Fig. 2. – The LINDAT Clarin Repository.

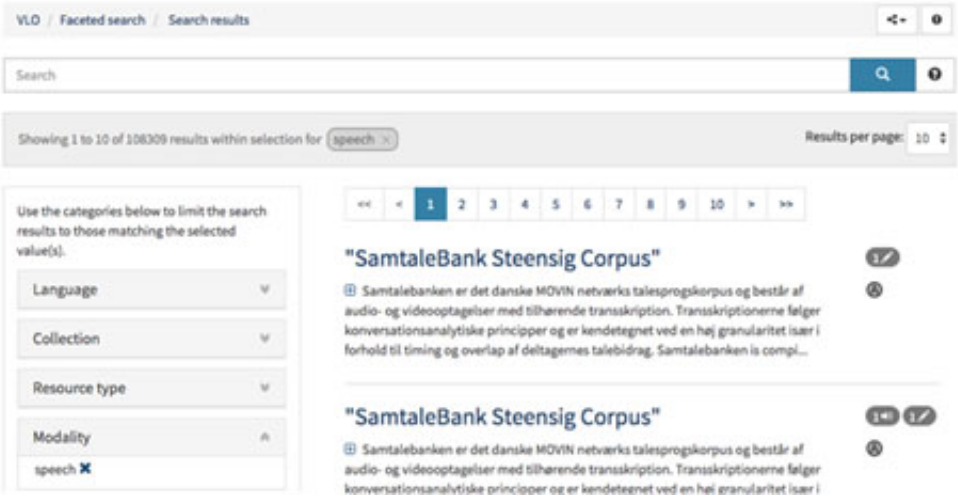

Fig. 3. – The VLO, CLARIN's meta-catalogue.

Because of this, metadata and metadata harmonisation constitute a crucial part of the CLARIN infrastructure. CLARIN promotes the use of a standard for metadata, namely CMDI, the Component Metadata Infrastructure [2–4] a format that allows for different metadata sets to be interoperable. CMDI is inspired by other metadata initiatives, such as Dublin Core (DC), OLAC[6], META-SHARE[7] and IMDI[8]; it is not itself a metadata scheme, but rather a framework, a common language, that several centres use to describe their own metadata in a way that allows the VLO to process them and make them searchable.

More specifically, CMDI defines 3 types of objects:

- *element*, an individual metadata description (*e.g.*, in a speech corpus an element may be the native language of the interviewer);
- *component*, a group of elements pertaining to a single aspect (*e.g.*, a component can describe all information on the license of a resource);
- *profile*, a complete metadata description for a given collection or type of resources, built by re-using different components.

Components are thus the building blocks of profiles, but single components may be re-used in different profiles to describe different types of resources. A "license" component for instance may be in principle used in all possible profiles. Profiles and components are defined in a common registry, the CLARIN Component Registry[9]; centres are encouraged to only create new profiles when existing ones cannot be used and even so, to re-use as many existing components as possible.

A feature of the CMDI standard that is important to understand for our purposes is that language resources may have a hierarchical structure. For instance a speech corpus may be constituted of several sessions, all grouped

[6]OLAC stands for Open Language Archive Community and its description can be found in [1].

[7]See [10].

[8]IMDI stands for Isle Metadata Initiative; this scheme is primarily used by the Language Archive at the Max Planck Institute in Nijmegen.

[9]`https://catalog.clarin.eu/ds/ComponentRegistry/`.

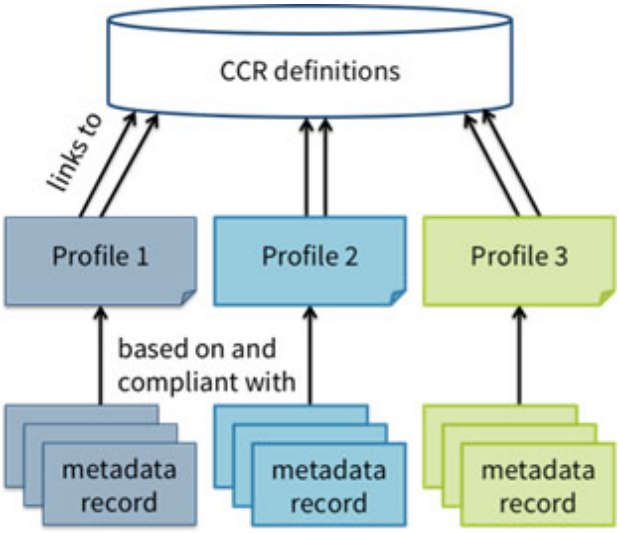

Fig. 4. – How CMDI works (image taken from www.clarin.eu/content/component-metadata).

into a single collection. Some of the medatada descriptors may apply to the collection as a whole (such as the contact person for the corpus, its license, the annotation levels ...), while other ones may be specific to a single session (such as the number of participants in a session, the languages spoken in it, the duration of the recording, ...). If the curators wish to introduce session specific metadata to describe parts of their corpus they will often use two profiles, a generic one to describe the collection and a specific one for the sessions. To preserve the hierarchical structure, the collection metadata file, constructed according to the more generic CMDI profile, will contain links to all the session specific metadata files by using a dedicated CMDI tag (*ResourceProxyList*).

A last important facility completes this picture, namely the Clarin Concept Registry (CCR), an ontology where the concepts in the domain of Language Resources are defined and provided with a persistent identifier. Each new element in the Component Registry is linked to a CCR concept clearly defining it (see fig. 4).

Let us take a concrete example. *AnnotatedCorpusProfile*, a profile which is used to describe annotated corpora, contains a component called *AnnotatedCorpusSpecific*, specifying information about the corpus. This in turn contains an element called *TemporalClassification* which is defined in the CCR[10]; moreover, its values are restricted to a list of possible choices (diachronic, synchronic, historical, modern, other, unknown) each of which is also linked to a concept in the CCR.

CCR links are crucial to allow for the meta-search in the VLO, since each facet of the meta-catalogue is linked to one or a list of CCR concepts. Thus different profiles, from different centers, may all provide access the same information under different elements, thanks to the shared concepts. In fact, it is the responsibility of the curators of each collection, with the help of the CLARIN centers and repositories, to apply, and if needed adjust or define CMDI profiles, to map them to the concepts and to verify that these concepts are associated to the right VLO facets, so that users may find their resources by making the appropriate queries.

Currently CLARIN repositories are required to export their metadata in a valid CMDI profile to the VLO. This means either converting their own metadata or using a curation system that automatically produces CMDI. LINDAT, the Czech CLARIN project, developed an adapted version of the Dspace repository[11], called CLARIN-Dspace[12], which implements its own CMDI profile. This type of repository is currently being adopted by various other national centers, including the CLARIN-IT. Moreover, two CMDI editors are currently available, namely COMEDI[13], developed by CLARINO and ABRIL[14].

From what we have just illustrated, the CMDI, Concept and Component Registry architecture may be seen as a masterpiece of diplomacy, achieving interoperability without forcing a unique set of metadata descriptors onto all

[10]Concept URI http://hdl.handle.net/11459/CCR_C-3823_21273bbe-3d22-38cd-9a9c-85cc8807d087.

[11]Dspace http://www.dspace.org/ is an open source software used to create open access repositories.

[12]The CLARIN-Dspace source code can be found here https://github.com/ufal/clarin-dspace.

[13]http://clarino.uib.no/comedi/page.

[14]https://tla.mpi.nl/tools/tla-tools/arbil.

centres, all disciplinary sub-communities and all types of resources. This is particularly important for legacy collections with an already established set of metadata, and for specific types of resources, such as is the case for speech audio archives and speech corpora in general.

2.1. *Existing profiles.*

In the present paragraph we shall present an overview of the existing profiles in the CLARIN Component Registry that may or have been used to describe speech audio archives. We are partly inspired by the work carried out by Freigang and Bergmann [9] for Multimodal corpora, though the situation has evolved since the publication of their paper.

At a very high level, speech archives, just as most corpus collections, may be described as a whole by using the very basic metadata description such as the one provided by the *collection* CMDI profile[15]; this profile only contains four basic components *Collection Info*; *License*; *Contact*; *Web Reference*. The elements in the *field Collection Info* contain relevant pieces of information: *Name*, *Title*, *Owner*, information about the language defined in terms of the standard *ISO639* codes; *Modality* (which, is defined with a list of possible values such as speech, writing, gestures, pointing-gestures, signs, eye-gaze, facial-expressions, emotional-state, haptic, song, instrumental music); *Time coverage*; *Description* (containing free text[16]). As for the component *License*, *DistributionType* (either public, academic, restricted, unspecified), *LicenseName*, *LicenceURL*, are the information required, and three further optional fields may further specify restrictions of use (*NonCommercialUseOnly*, *UsageReporRequired*, *ModificationsRequireRedeposition*). *Person*, *address*, *email*, *Organisation*, *Telephone*, *Website* refer to the Contact component, while *Website* and *Description* refer to *Web Reference component*.

Of course this very generic profile appears to be limited as to the specific

[15] https://catalog.clarin.eu/ds/ComponentRegistry#/?itemId=clarin.eu\%3Acr1\%3Ap_1345561703620&registrySpace=public.

[16] In fact the description component may be repeated to host more versions of the description in different languages each identified by the *LanguageId* attribute.

descriptors for speech data, apart from the *Modality* element and its possible values. Similar considerations hold for other widely used generic profiles such as the *OLAC-DcmiTerms* profile, based on the Dublin Core metadata with some OLAC-specific attributes, and used among other to import the Ortolang catalogue into the VLO[17]. Other, more specific profiles and components have thus been defined by curators of speech data. In table 1 we present a non comprehensive list of relevant profiles to be found in the CLARIN Component Registry.

Such profiles may be used for speech or multimodal resources, some describing a whole corpus and other more specific of a single session, or recording. A thorough description of each of these profiles, as well as of their use in the description of specific resources, would be beyond the scope of this paper. Here we shall describe some of the more important and re-used components, allowing for domain specific descriptions.

We have seen that some curators may use two sets of profiles, one for the entire corpus and one for the individual sessions. Such is the case of BAS (Bavarian Archive for Speech Signals) and HZSK (Hamburger Zentrum für Sprachkorpora) resources. In such case the more generic profile will contain a component describing the corpus as a whole; an example is *CMDISpeechCorpus*[18] component (fig. 5), used in many speech corpora profiles, and allowing curators to specify elements such as: *DurationOfEffectiveSpeech*, *DurationOfFullDatabase*, *NumberOfSpeakers*, *RecordingEnvironment*, *SpeakerDemographics*, *Quality*, *RecordingPlatformHardware*, *RecordingPlatformSoftware*. It also contains a sub-component, called *SpeechTechnicalMetadata*, for technical information about the recording (*SamplingFrequency*, ...).

In turn, the session specific profiles will contain components such as *mediasession*[19], storing information about the session itself such as *Location*, *Num-*

[17]For the CMDI metadata description of the file "Patois du Valjouffrey" see `https://vlo.clarin.eu/data/clarin/results/cmdi/Speech_and_Language_Data_Repository_SLDR_ORTOLANG_/oai_sldr_org_sldr000006.xml`.

[18]Component Identifier in the CLARIN Component `clarin.eu:cr1:c_1271859438144`.

[19]Component Identifier `clarin.eu:cr1:c_1324638957646`.

Table 1 – An overview of the principal profiles that are used for audio/speech/multimodal resources.

Profile name	**Profile ID**	**Examples of resources for which it is used ID**	**Remarks**
Collection	p_1345561703620	The Childes corpora collection	Generic profile, only allows for the specification of the modality
OLAC-DcmiTerms	p_1288172614026	Speech corpora esp. from French catalogues	Generic profile, DC + OLAC
Media-corpus-profile	2 versions: p_1324638957739 p_1387365569699	Corpora from BAS (Bavarian Archive for Speech Signals)	This profile is used to describe the whole corpus
Media-session-profile	p_1336550377513	Sessions from BAS (Bavarian Archive for Speech Signals)	This profile is used to specify individual sessions or recordings
SpeechCorpus	p_1357720977528	Used among others by the Dortmunder Chatkorpus 2.0	Defined by NaLiDa-Projekt
SpeechCorpus WithParticipants	p_1392642184799	Used only for *IMS GECO Datenbank*	Defined by IMS Stuttgart
SpokenCorpusProfile	p_1422885449343	Used by HZSK (Hamburger Zentrum für Sprachkorpora)	This profile is used to describe the whole corpus
CommunicationProfile	p_1427452477080	Used by HZSK (Hamburger Zentrum für Sprachkorpora)	This profile is used to specify individual sessions or recordings
OralHistoryInterviewDANS	p_1369752611610	DANS	Created for the INTERVIEWs corpus[a]

[a]For more information about the project see `www.clarin.nl/node/267`. In the VLO, the corpus is described only as a single entry, (*Living Oral History Workbench: Interviewproject Nederlandse Veteranen (IPNV)*), using a very generic profile.

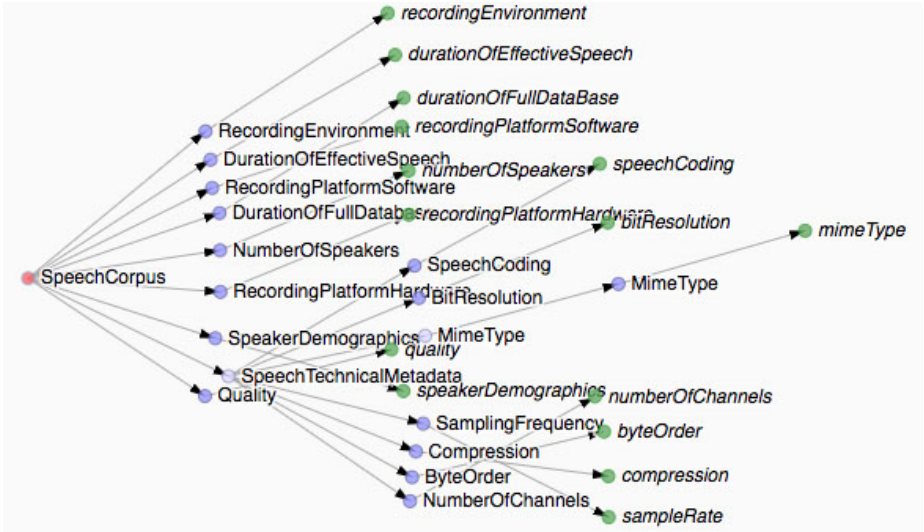

Fig. 5. – The SpeechCorpus component (on the left) with its subcomponents, elements and values. This graph may be obtained using the Curation Module made available by CLARIN-AT `https://clarin.oeaw.ac.at/curate/#!SMC%20Browser`.

berOfSpeakers, and the *MediaSessionActor* sub-component dedicated to detailed information on each participant in the session ...; another important sub-component is the *media-annotation-bundle*[20] allowing for a precise indication of the annotation levels. Other components may be used to specify the type of elicitation (*CollectionMethod*[21], *ElicitationMethod*[22]).

A final but important consideration needs to be made about the search facilities that allow users to query for such metadata. Indeed each CLARIN repository may have a search interface that is specifically designed for its hosted resources[23], or even no search interface at all[24]. At the same time, when

[20]Component Identifier `clarin.eu:cr1:c_1336550377515`.

[21]Component Identifier `clarin.eu:cr1:c_1307535113334`.

[22]Component Identifier `clarin.eu:cr1:c_1302702320411`.

[23]See for instance the NaLIDA faceted browser `http://www.sfs.uni-tuebingen.de/nalida/katalog/app/nalida/_design/nalida/index.html`.

[24]The BAS catalogue simply lists its own corpora and the set of metadata; see for instance this page referring to the Alcohol Language Corpus `http://www.bas.uni-muenchen.de/forsch`

each repository is harvested by the VLO, the metadata descriptions will be searchable by means of a single set of facets[25] (the mapping between elements and facets being a matter of agreement, as was described in the previous paragraph). So for instance the metadata of some sets of recordings might contain indication as to the sex of the participants, and yet it may be hard to systematically search for all recordings of female informants in the VLO, since no search facility for this information is available.

The panorama of speech resources within CLARIN is currently very active. Discussions are on-going on how to make such resources more visible and more easily identifiable as such in the VLO. The discussion that follows below may be seen as a first contribution of this debate, limited to the Italian scenario.

3. The Italian scenario

Only a small part of Italian speech archives is accessible via VLO. Several corpora of child language are present, both from normal and disordered children, and from bilinguals, uttering different styles of speech (*e.g.*, CHILDES[26], Bilingual ProjectS Corpus). Italian people with aphasia (*e.g.*, AphasiaBank Italian Protocol corpus; AphasiaBank Italian NonProtocol CAP Corpus); second language acquisition data (*e.g.*, SLABank ESF German-Italian corpus; SLABank ESF English-Italian corpus; P-MoLL Modalität in Lernervarietäten im Längsschnitt; e-LiLT corpus[27]; HABLA - Hamburg Adult Bilingual Language); and political talk shows transcribed according the conversational analysis framework are documented (*e.g.*, CABank MOVIN Corpus). Italian data stemming from single research projects are also accessible (*e.g.*, Evolu-

ung/Bas/BasALCeng.html.

[25]Such facets are: Language, Collection, Resource type, Modality, Format, Keyword, Genre, Subject, Country, Organisation, National project, Data provider.

[26]CHILDES is the child language component of the TalkBank system. TalkBank is a system for sharing and studying conversational interactions.

[27]The corpus was created within the ESRC funded project "The Ups and Downs of Learner Intonation: A Cross-Language and Longitudinal Investigation of the Intonation Systems of L2 Learners".

tion of Semantic Systems (EoSS) project)[28], together with several linguistic biographies and map tasks, mostly from Sud Tyrol, but also from the Italian national Project CLIPS (see *infra*)[29]. As far as we know, very few available resources stem from regional and dialectal speech: the BAS corpus AsiCa contains documentation of Calabrese dialect for the purpose of syntactical analysis[30].

This review (updated on 28/12/2016) proves how VLO is not yet an ideal access point to Italian speech and oral archives. The main virtual access to Italian speech corpora is the web portal *parlaritaliano.it*, which is presented as "an observatory of language uses". It contains the great majority of the resources created by Italian University research groups in the last decades. It is important to underline that the spoken resources virtually accessible via this web portal are all corpora and not archives, being all born digital and being designed for specific purposes. For our purposes, two sections of the portal are worthy mentioning: projects and corpora areas. Let us concentrate on the corpora section, where the following audio resources referring to the Italian linguistic scenario are available for download: Corpus AN.ANA.S. Multilingue (AN.ANA.S._MT); Corpus AVIP-API; Corpus bilanciato consultabile LABLITA; Corpus CLIPS; Corpus DILS (Dialoghi in Italiano Lingua Straniera); Corpus LIPS; Corpus PENELOPE; Corpus PraTiD; Corpus PraTiD nelle lingue europee; Selezione dal "Corpus di parlato telegiornalistico. Anni Sessanta vs. 2005"; SpIt-MDb (Spoken Italian - Multilevel Database). Notwithstanding, *parlareitaliano.it* is not the unique access point to Italian collection. In the volume of Cresti and Panunzi [7] a detailed description of different web access point to Italian and regional speech corpora and archi-

[28]A research initiative of the Max Planck Institute for Psycholinguistics, in which elicitation data for colour, body parts, containers and spatial relations are presented.

[29]Interestingly, via Bavarian Archive for Speech Signals (BAS).

[30]The aims of the AsiCa-Corpus were the analysis of syntactical structures and their geolinguistic mapping in form of interactive, web-based cartography. Around sixty speakers of Calabrese (one half of whom having migration experience in Germany, the other half almost always having stayed in Calabria), equally balanced regarding gender, age and geographical origin, were recorded. Spontaneous and controlled speech were elicited for almost all the informants. The results of syntactical analysis (maps and text) can be found on the project's website at `http://www.asica.gwi.uni-muenchen.de`.

ves is presented: their review points out a dynamic and industrious picture of several research groups and enterprises, together with a huge amount of audio data, which clearly suffer for fragmentation and lack of communication. This is particularly true in the case of born analogue archives collected in the domain of oral history, dialectology, anthropology, sociology and other social sciences, scattered everywhere in the Italian Peninsula in research institutions, academies, libraries or owned by private citizens.

4. Conclusions

Given this scenario, at least two desiderata can be suggested. Firstly, it could be useful to make Italian digital speech resources accessible via VLO. In order to achieve this goal, and therefore have much more visibility and increase the possibility of proper re-use of Italian research data, knowledge of and participation to the CLARIN infrastructure should be more widespread among scholars working with oral and speech data. Italy became Full Member of CLARIN ERIC from 1st October 2015: the contribution of Italian research groups and single researchers is therefore important in order to create a community which goal is to foster the knowledge and dissemination of Italian speech corpora[31]. Several initiatives are on the agenda: for instance, a feasibility study has been pursued with respect to the Gra.fo project oral archives, in order to be accessed via VLO [6]. Secondly, a huge amount of work has still to be done as for the born analogue oral archives, from census to digitization to description. From this respect, the census of Italian oral collections has recently been updated and can now be accessed, together with the census of all other European countries, via the following URL `http://www.oralhistory.eu/collections/clarin-eric`. The Italian Association of Oral History (AISO) collaborated with S. Calamai on the census and – together with the Italian Association of Speech Sciences (AISV) – has recently been involved in the CLARIN Oral History and Technology project, whose main goal is to start an OH-transcription chain, helping OH-scholars to use

[31]The Italian consortium (headed by Monica Monachini of ILC-CNR and still under construction) aims to be representative of all communities producing and working with Language Resources. For more information visit `http://clarin-it.it/` and see [13].

technology in order to speed-up the transcription process (`oral-history.eu`). In this respect three different workshops have been organized: the CLARIN workshops on Oral History (OH) archives in Oxford (April 2016), in Utrecht (December 2016), and in Arezzo (May 2017). Dutch, English, and Italian are the languages chosen for testing the transcription process: thus, the contribution of the Italian research community – both linguists and speech technologists – will play an invaluable part.

References

[1] S. Bird, G. Simons, *The OLAC Metadata Set and Controlled Vocabularies*. In: *Proceedings of the ACL Workshop on Sharing Tools and Resources for Research and Education*. Toulouse 2001, 7-18.

[2] D. Broeder, M. Kemps-Snijders, D. van Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg, C. Zinn, *A Data Category Registry – and Component – based Metadata Framework*. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta 2010.

[3] D. Broeder, O. Schonefeld, T. Trippel, D. van Uytvanck, A. Witt, *A pragmatic approach to XML interoperability – the Component Metadata Infrastructure (CMDI)*. Balisage Series on Markup Technologies, 7, 2011.

[4] D. Broeder, M. Windhouwer, D. van Uytvanck, T. Goosen, T. Trippel, *CMDI: a component metadata infrastructure*. In: *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme* (Workshop co-located with LREC 2012). Istanbul 2012, 1-4.

[5] S. Calamai, *La Romagna toscana nella percezione dei parlanti*. In: R. Bombi, V. Orioles (eds.), Atti del Convegno Internazionale della Società di Linguistica Italiana *Lingue in contatto / Contact Linguistics* (Udine 25-27 settembre 2014). Bulzoni, Roma 2016, 247-367.

[6] S. Calamai, F. Frontini, *Not quite your usual kind of resource. Gra.fo and the documentation of Oral Archives in CLARIN*. In: *Proceedings of the CLARIN Annual Conference 2016*. Aix-en-Provence 2016.

[7] E. Cresti, A. Panunzi, *Introduzione ai corpora dell'italiano*. Il Mulino, Bologna 2013.

[8] A. De Dominicis, P. Mattana, *Il Progetto Bomarzo*. In: L. Romito, V. Galatà, R. Lio (eds.), Atti del quarto Convegno Nazionale dell'Associazione Italiana di Scienze della Voce 2007 *La Fonetica Sperimentale: Metodo e Applicazioni*. Torriana 2009, 405-411.

[9] F. Freigang, K. Bergmann, *Towards Metadata Descriptions for Multimodal Corpora of Natural Communication Data*. In: *Proceedings of the Workshop on Multimodal Corpora 2013: Multimodal Corpora: Beyond Audio and Video.* Edinburgh 2013.

[10] M. Gavrilidou, P. Labropoulou, E. Desipri, S. Piperidis, H. Papageorgiou, M. Monachini, F. Frontini, T. Declerck, G. Francopoulo, V. Arranz, V. Mapelli, *The META-SHARE Metadata Schema for the Description of Language Resources*. In: *Proceedings of the eighth International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul 2012.

[11] E. Hinrichs, S. Krauwer, *The CLARIN Research Infrastructure: Resources and Tools for e-Humanities Scholars*. In: *Proceedings of the nineth International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavík 2014.

[12] E.A. Mazé, *Metadata: Best Practices for Oral History Access and Preservation*. In: D. Boyd, S. Cohen, B. Rakerd, D. Rehberger (eds.), *Oral history in the digital age.* Institute of Library and Museum Services 2012. Retrieved from `http://ohda.matrix.msu.edu/2012/06/metadata/`.

[13] M. Monachini, F. Frontini, *CLARIN, l'infrastruttura europea delle risorse linguistiche per le scienze umane e sociali e il suo network ita-*

liano CLARIN-IT. Italian Journal of Computational Linguistics, 2(2), 2016, 11-30.

[14] T. Váradi, P. Wittenburg, M. Wynne, S. Krauwer, K. Koskenniemi, *CLARIN: Common Language Resources and Technology Infrastructure*. In: *Proceedings of the sixth International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech 2008.

[15] D. Van Uytvanck, C. Zinn, D. Broeder, P. Wittenburg, M. Gardellini, *Virtual Language Observatory: The Portal to the Language Resources and Technology Universe.* In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*. Valletta 2010.

Luca Lorenzetti* - Paolo Milizia** - Giancarlo Schirru***

COLLECTING SPOKEN DATA FROM UPPER-SOUTHERN ITALIAN DIALECTS

Abstract. – In this paper we discuss some theoretical and practical problems of the collection and presentation of spoken data for dialectological purposes. We will pose questions connected with audio data recording, digital audio files storage, data transcription, transcription quality checking. Furthermore, we discuss pros and cons of the application of morphosyntactic tagging models to our data. More specifically, the paper describes the main activities and projects of the *Centre for the documentation of dialects of Southern Lazio*, set up by the authors at the University of Cassino.

1. The project

The *Centre for the documentation of dialects of Southern Lazio* was founded in the mid-2000s by the authors. The centre arose as a local project of the Laboratory of phonetics and linguistics of the University of Cassino, where all the three of us worked at the time. The project aimed:

a) to enhance linguistic studies of an exceptionally rich and interesting area;

b) to develop and test a protocol of data collection and analysis that would fit with the needs and level of BA or MA students;

c) to provide scientific and local communities with the results of this operation.

*DISTU - Università degli Studi della Tuscia - Via S. Carlo, 32 - 01100 Viterbo.

**Dipartimento di Scienze Umane, Sociali e della Salute - Università di Cassino - Loc. Folcara - 03043 Cassino.

***Dipartimento Asia, Africa e Mediterraneo - Università degli Studi di Napoli «L'Orientale» - Piazza S. Domenico Maggiore, 12 - 80134 Napoli.

The dialects of Lazio are well known for displaying many features that are crosslinguistically relevant at all levels, from phonology to syntax (see [20], for a detailed report of the scientific production on the topic; examples of analysis of specific phenomena in [21, 25, 26], Paolo Milizia, this paper). Moreover, language vitality is still very high in many areas of Lazio, so that there are very good chances to collect interesting data, both under a linguistic and an ethno-anthropological profile. It is worth noting that even if some local dialects have been already well described by distinguished scholars, between the end of the 19th and the beginning of 20th century, this is not an actual limit to the scientific interest of the area, an interest which is indeed increased by the possibility of comparative surveys between the contemporary and the older data, the latter often deserving, on the other hand, a critical rethinking in the light of new theoretical frameworks.

Let us come now to the second point: the need for a model of data gathering and analysis whose first steps could be feasible, even as part of a BA or MA thesis, by young researchers who underwent only a limited training. The involvement of university students in the work of data collection is crucial under several aspects. First of all, it fits well with an old desideratum of dialectological research, according to which the researcher should be also a native speaker or at least an insider in the investigated community, so that s/he could be able to collect texts of spontaneous speech through informal inquiries with a low degree of directivity. In addition, this involvement triggers a virtuous collaboration between the university, as a representative of the scientific community, and the local speakers of the investigated language varieties. Finally, working with university students could improve the economic sustainability of the project: but on this point we will come back soon.

One could also recall the widespread view that some of the best products of Italian dialectology stem from the reworking of Master, PhD or habilitation dissertations: works such as Campanelli 1986 (on the phonology of the dialect of Rieti, a thesis supervised by Ernesto Monaci), Maccarrone 1915 (written by the newly graduated author while holding his year of apprenticeship as a teacher in Cervaro), Bianconi 1962 (comparison between Viterbo and Orvieto, supervisor Arrigo Castellani), Ernst 1970 (on the tuscanization of Romanesco, a German PhD thesis), down to Bernhard 1998 (Language

Variation in Romanesco, German Habilitationschrift) and so on. Of course we have enough sense of humor to avoid comparisons with predecessors of that sort; moreover, it goes without saying that involving students in such a project would import problems not less than solutions, but again, on this point we will go back shortly in more detail.

The substantial point is that we are talking about a basically unfunded project. Including the project within the current activities of university teaching could be a partial solution to the lack of funding, even if one with obvious limitations in the expected results.

It is clear indeed that the goals of CDLM are significantly different from those of other similar enterprises, of much larger scale (larger both for economic and scientific commitment), either successfully accomplished or still actively pursued in other areas of Italy. Projects of great value, such as that of ALT (*Atlante Lessicale Toscano*) or that of the VIVALDI (*Vivaio Acustico delle Lingue e dei Dialetti d'Italia*), give the user great wealth of information, which goes from the usual linguistic atlas query, sorted by questions (stimuli) or phenomena, to the access to the raw investigation materials. Nonetheless, this sort of projects cannot provide - by its own nature - the availability of spontaneous speech recordings for the user: the limits that this exclusion entails for analysis at the syntactic or phonetic and phonological level do not need to be highlighted here.

The expected results of CDLM can be summarized as follows. The project is making large texts of spontaneous spoken dialects available in the following forms:

1) audio streaming;

2) downloadable audio files (low fidelity);

3) wav audio files (high quality, on request);

4) IPA broad transcription;

5) simplified transcription;

6) Italian translation of the text;

Fig. 1. – From CDLM website, a screenshot of the incipit of a text from Boville Ernica (Frosinone) with broad IPA transcription, simplified transcription and Italian translation.

7) English translation of the text;

8) a full lexicon indexed by tokens and types.

Except for (7) and (8), all these features are already – more or less – available at `linguistica.unicas.it/dlm` for 21 investigation points corresponding to different traditional dialects of the area between Southern Lazio, Northern Campania and Molise. In fig. 1 a specimen of a text (from Boville Ernica) is displayed.

We can see the broad phonetic transcription on the left column, the simplified transcription in the centre and the Italian translation on the right. On the webpage, clicking on the upper buttons the user could also hear to the corresponding audio low quality streaming. Researchers, teachers, local scholars or whoever can of course ask for copies of high quality audio files by contacting the authors.

The characteristics of the sound signals, the different sound file formats that CDLM provides and the technical questions on the various types of transcription involved in the project are detailed in the following of this paper by Giancarlo Schirru, while Paolo Milizia will describe some promising examples of application to CDLM materials of morphological tagging techniques. I would just add here a few words about the quality of the transcription.

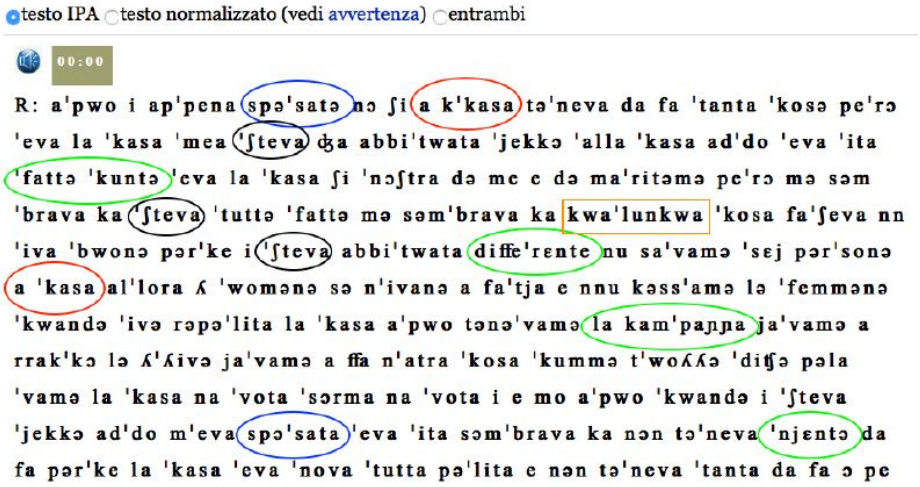

Fig. 2. – From CDLM website, a text from San Donato Val Comino with some errors displayed in the transcription.

As we said above, the bulk of the current transcriptions set on the webpage of our project has been made by junior researchers: BA, MA or PhD students in Linguistics. This poses a reliability question: how much should we trust this sort of transcriptions? A first control is actually performed by us as the teachers during the normal correction of the thesis: but as I said above, it would be impossible to check on the audio any single transcripted form. Normally we check the transcriptions in a somewhat naïf way. First we look for inconsistencies in phonetic ortography: for example, in the transcript from San Donato Val Comino displayed in fig. 2 we should not have passed without further checking the form [kwalunkwa] with coronal nasal, nor the doublet [a kˈkasa] vs. [a ˈkasa] with and without phonosyntactic doubling, nor even the voiceless vs voiced stops in the clusters with a preceding nasal [kuntə, kampaŋŋa, njɛnte]. But we actually did control the apparent variation between alveolar vs postalveolar clusters in [spəˈsatə, spəˈsata] vs [ʃteva], just because at the time one of us was studying that particular phenomenon.

Then we pick up a few random items along the whole transcript and check them on the audio. This way we get a general picture about the quality of

transcription, but it is surely not enough to present it to the public. For that reason we have planned, but not yet extensively done, a more effective control on audio samples taken out from the whole text: one full word out of fifty is picked up and checked on the audio, then its transcription gets possibly fixed. On the basis of this control a ranking of accuracy will be given to the transcriptions, ranging from 1 (excellent) to 5 (unreliable): the majority of our transcritptions are at the moment classified as "n.v.", that is "not yet checked". The scholar – or the simple reader – who is using a specific transcription will thus be able to know in advance whether or not the text s/he is reading is trustworthy, and whether s/he should better check transcriptions on the audio tracks. Since both naïve transcriptions and audio files are immediately available on the web, everyone who wants can make his own double-check on the materials, and that would allow us to give the users a "good-enough" service, if not definitely an optimal one.

2. Acquisition, storage, transcription of audio data

Each acquisition of audio data is oriented to a specific target: the recording of an interview as a document for the oral history requires different standard from the shooting of a documentary movie or a TV interview[1]. Similarly, the recording of a speaker as a document for a syntactic inquiry on a dialect needs different parameters from the acquisition of a speech signal for an acoustic phonetic analysis. The goals of speech signal capture are spread in a wide range, including medical investigation and analysis, human-to-machine communication, forensic applications.

Speech data records must be stored in order to be analyzed: short of the presence of particular constraints of protection, they remain available for further investigations. Therefore, they can be used even outside the contexts and the aims which motivated their collection.

The multiple functionality of data is one the advantages of the big classical collections of linguistic interest: for example, linguistic atlases (*e.g.* [7, 16])

[1] A discussion on the technical standards of the data recording for oral history is developed, for example, in [3, 29].

are currently employed in specific analysis, from the pioneering essays of Jules Gilliéron (see [8–11]) to present-day dialectometry (see [12, 14, 15]). Such investigations are not only dedicated to the lexicon, which is the main object of classical atlases, but involve phonology, morphology and syntax as well; moreover, they range in different theoretical perspectives.

The same aim is pursued by new-generation collections, through the availability of audio data on cd-rom or on dvd; see, for example, the *Atlante linguistico del ladino dolomitico* [13] and the *Atlas of North American English* [17].

These premises lead to the discussion of the technical standards of audio data: how can they be recorded, stored and transcribed?

2.1. *Audio recording.*

Audio recordings are the primary data for linguistic investigations: a first set of standards concern the process of their acquisition.

A digital recorder should be employed, producing an uncompressed audio file, a so-called PCM audio (the most common formats are WAV, AIFF and AU).

The parameters of the digitization can be a sampling rate at least of 22.500 Hz (or 20.000 Hz, or even 16.000 Hz), with 16-bits quantisation (for digital audio parameters see cap. 14 in [23]). Since the highest frequency recorded is one half of the sampling frequency, and the main linguistic phonetic features can be detected in the first 8.000 Hz, a frequency sampling lower than 16.000 Hz seems not to be adequate for acoustic phonetic analysis.

These are the suggested parameters for the new recordings. If a digital audio file is the result of the conversion of another file (for example, previous analogical recording), the analysable bandwidth is not represented by the sampling rate of the conversion process, but obviously by the original recording[2].

[2]As an example, all five files of the Workshop Speech Corpus are sampled, in the digitization process, at 44.100 samples for second (the audio CD standard), and allow a spectrogram bandwidth of 22.500 Hz. But only in one of them (Ayoreyo.wav) the entire bandwidth is analysable; in the other four, which are presumably converted from analogical recordings, the actual bandwidth of the spectrogram is largely lower: the file CG5.wav displays, among

A manual regulation of recording volume levels is preferable, in order to avoid the equalization effect of the automatic adjusting.

One of the main concerns of data recording is the presence of environment noises, may be unnoticed during the interview, but creating real gaps in the speech signal[3]. The best solution is a recording in a closed and silent room, without other people speaking inside it. If an open-air recording is necessary, the reduction of unavoidable noise can be operated by the microphone: a close-talking noise-cancelling microphone, held by a head mount a few centimetres to the side of a speaker's lips can deal with this situation [18: 33].

2.2. *Storage.*

Audio file can be stored in a digital memory support (for example a hard disk), from where they can be read and analyzed in a laboratory. The full format of the audio file is not easy to handle for remote transfer or streaming.

Our *Centre for the documentation of dialects of Southern Lazio* offers the online streaming of most of the speech signals. They are coherent with the time tagging within the text of the transcription and the translation. But in the case, the file streamed is not the full PCM audio file, but a compressed MP3 version of it, stored into the web server.

A similar possibility of on-line hearing of audio samples is offered, for example by the useful *Vivaio acustico delle lingue e dei dialetti d'Italia* (Vivaldi), directed by Roland Bauer and Dieter Kattenbusch [1].

The full audio file can be sent by means of a transfer, under request of remote users.

2.3. *Transcription.*

Although a broad phonetic transcription in International Phonetic Alphabet cannot satisfy all the needs of a linguistic investigation, and sometimes a more detailed transcription is requested (see De Dominicis' paper in this Workshop [4]), a phonemic oriented IPA transcription is able to deal, at least

other lacks, a visible spectrogram just for the first 4000 Hz.

[3] An example of it is offered by the file maiale.wav of the Workshop Speech Corpus.

as a first approximation, with most of the necessities of linguistic analysis, at all the levels: phonology, morphology, syntax, semantics, lexicon.

The main concern of the IPA transcription is represented by its obscurity for non-specialist readers. Again the claim for an availability of speech data outside the disciplinary field in which they were collected can be formulated. An IPA transcription is doubtless unfriendly for ethnologists, historians, psychologists, sociologists or, more simply, for people interested in dialectal data for personal reasons.

The solution we pursued in the web site of our *Centro di documentazione*, to get over such a difficulty, is the adoption of a double system of transcriptions. Beside the IPA transcription, a "normalized" transcription, based on the convention of Italian orthography, is available. The latter was illustrated by Glauco Sanga [24] in the first number of the "Rivista italiana di dialettologia": it is in italic, and contains the following among its few conventions:

- the letter <ə> represents the central vowel (IPA [ə]);
- the letter <j> represents the palatal glide (IPA [j]);
- the letter <z> represents a voiceless alveolar affricate (IPA [ts]); underdotted <ẓ> stands for voiced alveolar affricate (IPA [dz]);
- in front of a consonant, <s> can represent a post-alveolar fricative (IPA [ʃ]), instead that an alveolar fricative (IPA [s]);
- between vowels,<c(i)> can correspond to a phoneme realized with a post-alveolar fricative (IPA [ʃ]), instead of post-alveolar affricate (IPA [tʃ]);
- <é> and <ó> represent mid-high vowels (IPA [e, o]), <è> and <ò> mid-low ones (IPA [ɛ, ɔ]);
- digraphs and trigraphs <ch, gh, ci, gi, sc(i), gn, gl(i)> have the value that Italian orthography (respectively IPA [k, ɡ, tʃ, dʒ, ʃ, ɲ, ʎ]).

This “normalized” transcription is automatically produced from the IPA one, by the mean of a software. Therefore, it lacks apostrophes, capital letters and punctuation marks, and displays phono-syntactic phenomena.

Example of such a normalized transcription of the Italian files within the Workshop Corpus (turn-taking overlapped are rendered in sequence):

(1) file maiale.wav:
a. [ɛ ritorˈnato ˈdʒu … aŋˈkora ˈstavano ko … koˈlato ˈdʒu ko a katiˈnɛlːa e … koˈsi nu mːuˈriva l maˈjale]
b. *è ritornàto giù … ancóra stàvano co … colàto giù co a catinèlla e … così nu mmurìva l majàle*

(2) file oro.wav:
a. D [l ˈɔro njɛnte]
R [l ˈɔro]
D [alˈːora se faˈʃeva … koˈralːi]
R [pe ˈkːwelːo ˈɛkːo na ˈfila de koˈralːi o na katiˈnina … na ˈkosa koˈsi]
b. D *l òro niènte*
R *l òro*
D *allóra se facéva … coràlli*
R *pe kkwéllo ècco na fìla de coràlli o na catinìna … na còsa così*

(3) file vestiti.wav
a. D [ˈkwando vo … ve ˈsjete spoˈsati ˈvoi … ke kosˈtume portaˈvate ke vesˈtito portaˈvate]
R [vesˈtiti]
D [ɛ]
R [emˈbɛ vesˈtiti imˈpɔrta]
b. D *quàndo vo … ve sjéte sposàti vói … che costùme portavàte che vestìto portavàte*
R *vestìti*
D *è*
R *embè vestìti impòrta*

3. Annotation

An important task in order to facilitate the use of the recorded data in researches focusing on levels of analysis different from the phonetic/phonemic one is to provide transcriptions with morphosyntactic annotation. At present, this step of the project has not been realized yet in the *CDLM*, but here it is possible to present some preliminary considerations.

Among the different annotation systems that have been proposed in literature, the Universal Dependencies scheme (http://universaldependencies.org, [5]) has begun, in very recent years, to be considered as a *de facto* standard. A main goal of this system is to devise a tag-set which is language-independent, *i.e.* potentially applicable to every linguistic system. As is inevitable, its architecture implies the selection of several choices which involve equally admissible theoretical alternatives: *e.g.*, to mention only a couple of very fundamental options, the system is based on dependency representations rather than on constituency, and it assigns head roles to content words rather than function words. As for the file type, UD treebanks adopt tab-separated-values text files (utf-8 encoding) according the so-called CoNLL-U format.

Here follows a minimal example for UD-CoNLL-U annotation applied to a sentence in our corpus[4] (see fig. 3):

alvito_testi_1 sent_id to_be_determined

speaker: informant

text: təˈneva dːudəʃˈ anːə e jˈːava a tsːapˈːa kə tˈːata;

translation: avevo dodici anni e andavo a zappare con papà

[id form lemma part-of-speech features head dependency-relation]

1 təˈneva təˈne VERB Mood=Ind|Number=Sing|Person=1|Tense=Imp|VerbForm=Fin 0 root

2 dːudəʃ dːudəʃ NUM NumType=Card 3 nummod

[4]For a detailed explanation of the labels applied to morphosyntactic features and to dependency relations, see [5] and the UD website (`universaldependencies.org`).

3 ˈanːə ˈanːə NOUN Gender=Masc|Number=Plur 1 obj

4 e e CCONJ _ 5 cc

5 jˈːava ˈi VERB Mood=Ind|Number=Sing|Person=1|Tense=Imp|VerbForm=Fin 1 conj

6 a a ADP _ 7 mark

7 tsːapˈːa tsːapˈːa VERB VerbForm=Inf 5 xcomp

8 kə kə ADP _ 9 case

9 tˈːata tˈːata NOUN Gender=Masc|Number=Sing 7 obl

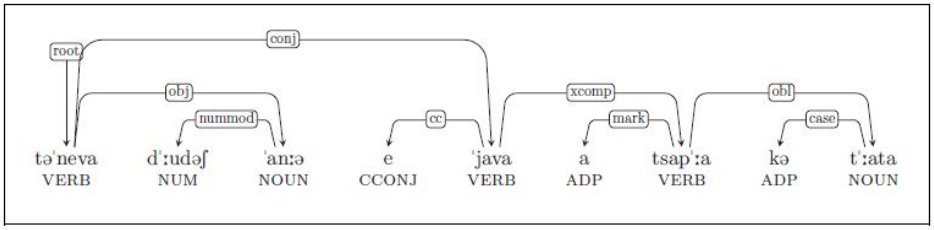

Fig. 3. – Example of dependency tree.

Importantly, the UD scheme allows scholars to define language-specific subsets of part-of-speech tags and of morphological features. Thus, for instance, in the case of Italo-Romance dialects an expansion of the tag-set is required in order to account for the different possible values of the feature GENDER. In particular, if we follow the analysis by Loporcaro and Paciaroni [19: 410 and passim] the alternating neuter and the so-called mass neuter must be considered two gender values different from each other (in addition to being different from the masculine and the feminine): *e.g.*, Amasenese alternating neuter [ʎu ˈlɛtte] / [le ˈlɛttəra] vs. mass neuter [lu ˈlatte] vs. masc. [ʎu ˈfile] / [ʎi ˈfili] vs. fem. [la ˈkasa] / [le ˈkase]. However, resorting to specific additional tags does not solve all the problems at issue. Indeed, a recurring condition in dialects such as the ones of Southern Latium is the existence of doublets consisting of a masculine form and a corresponding alternating-neuter form: *e.g.* [ʎu ˈlɛtte] / [le ˈlɛttəra] along with [ʎu ˈlɛtte] / [ʎi ˈletti] (cfr. also [27: 64]). In these cases, the singular of the alternating neuter noun can-

not be distinguished from the one of the masculine one so that the occurrence of a form like [ʎu ˈlɛtte] might be ascribed to the one as well as to the other.

On the other hand, since the UD scheme has been already applied to Italian and other Romance varieties, some possible annotation problems related to syntactic characteristics of Italo-Romance or of Romance languages in general may be dealt with by following the solutions previously adopted [2, 5]. Thus, for instance, clitics belonging to Italo-Romance pronominal verbs are to be labeled as *expl* according to de Marneffe and colleagues [5][5].

A further issue concerns whether and to what extent information related to diachrony or to cross-dialectal comparison should be included in the treebank. For instance, in the dialect of San Donato Val Comino the 3pl auxiliary in the passato prossimo is [javə], a form etymologically connected to latin HABERE, while other persons exhibit auxiliary forms related to ESSE[6]. This distribution testifies for a pattern of person-sensitive auxiliary selection which is well attested in other central-southern Italo-Romance varieties. However, from a purely synchronic point of view, there is no reason to ascribe two auxiliaries to this dialect: there seems to be no other continuation of HABERE to which [javə] could be connected synchronically (the verb for 'have' is, as expected in this area, a continuation of TENERE, while [ˈaʎʎa] 'I have to', 3pl [ˈjava] with final /a/, seems to behave synchronically as a distinct lexeme); a confirmation of this fact is provided by the circumstance that [javə] serves as a model for the creation of the 3pl forms [tjɛvə], [vjɛvə] belonging to the paradigms of [təne] 'have' and [vəni] 'come'. This change implies that [javə] has been reanalysed as [ja] + [və], where [ja] functions synchronically as an allomorph of the auxiliary stem.

Other issues are related with properties specific to spoken corpora. In the first place there is a minimal amount of pragmatic information that must be necessarily given (in particular, speaker identity and turn changes), especially

[5]UD drops the special label *clit* that is assigned to such clitics in the Italian Stanford Dependency treebank [2], the annotated corpus from which the UD Italian treebank was obtained by conversion.

[6]A survey of the periphrastic forms of this variety is in Roberta Pellegrini, *La morfologia verbale del dialetto di San Donato Val di Comino*, unpublished master's thesis, Università degli Studi di Cassino e del Lazio meridionale, aa. 2015-2016.

in conversations in which two or more informants participate. Markup guidelines specifically devised for such data have been proposed in literature, with most of them following the XML or SGML standard (see [28: 86]). For the purposes of our project, however, a comprehensive pragmatic annotation is not so urgently needed, and a simpler solution similar to the CHAT format [22] adopted in the CHILDES corpus (where, *e.g.*, participants are indicated simply by three-letter codes placed at the beginning of a text line), might be sufficient[7].

Another necessary type of metadatum concerns time alignment. At the present time most of the texts available in our corpus have been provided with marks that divide the recordings in time spans of about two or four minutes. Though being less fine-grained than other time alignment annotations (such as the one of the CHILDES corpus), these tags are useful for helping users synchronize listening and reading in the website version.

It should be noted, moreover, that not only pragmatic annotation and metadata but also morphosyntactic annotation proper can face problems specifically related to the spoken nature of the corpus. Indeed, a series of labels belonging to the Universal Dependencies basic tag-set or to extensions of it have been proposed precisely in order to deal with syntactic particularities of speech, such as repairs and restarts (see [6]). Thus, for instance, the sentence in the example above begins with a sequence ([təˈneva kwatːərdəʃ ˈ anːə təˈneva dːudəʃ ˈ anːə]) which cannot be considered as an instance of asyndetic coordination and which, according to the UD scheme, can be treated as in fig. 4.

To conclude, even if some descriptive issues will likely require some compromise, a morphosyntactic annotation based on the UD scheme promises to be a valuable tool for analysing the grammar of the linguistic varieties recorded in our corpus and for investigating diatopic (micro-)variation phenomena pertaining to morphology and syntax.

To sum up, its richness in diatopic variation and the non-negligible degree of vitality still shown by its dialects make Southern Lazio a particularly favourable subject for the creation of a multi-dialectal spoken corpus with regional

[7]In the transcriptions presently available in our website, a similar device has already been adopted.

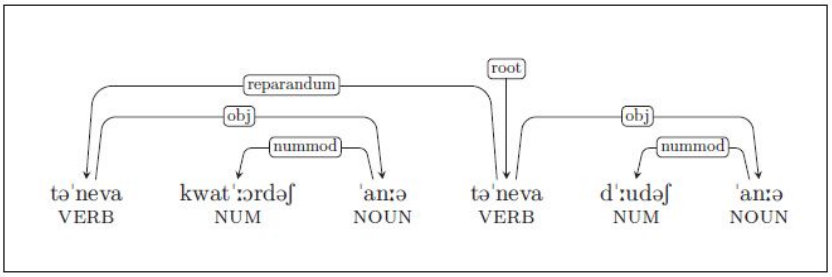

Fig. 4. – Example of reparandum relation.

scope. As we have tried to show, a series of theoretical and practical problems is encountered in designing the architecture and features of such a tool and, in making the relevant choices, one has to keep in mind the whole range of possible usages – from linguistic investigation at different levels of analysis to ethno-anthropological documentation – and users – from researchers to general public.

References

[1] R. Bauer, D. Kattenbusch, *VIVALDI, Vivaio acustico delle lingue e dei dialetti d'Italia*. Humboldt Universität Berlin, https://www2.hu-berlin.de/vivaldi/.

[2] C. Bosco, S. Montemagni, M. Simi, *Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank*. In: *The 7th Linguistic Annotation Workshop and Interoperability with Discourse. ACL workshop*. Sofia 2013, http://medialab.di.unipi.it/downloads/ISDT/MIDT-STD2013_law.pdf.

[3] A. Boyd Douglas, *Audio or Video for Recording Oral History: Questions, Decisions*. In: D. Boyd, S. Cohen, B. Rakerd, D. Rehberger (eds.), *Oral History in the Digital Age*. Washington DC 2012, http://ohda.matrix.msu.edu/2012/06/audio-or-video-for-recording-oral-history/.

[4] A. De Dominicis, *La voce come bene culturale*. Carocci editore, Roma 2002.

[5] M.C. de Marneffe, T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, Ch.D. Manning, *Universal Stanford Dependencies: A cross linguistic typology*. In: *Proceedings of LREC 2014, Ninth International Conference on Language Resources and Evaluation*. Reykjavik 2014, http://nlp.stanford.edu/pubs/USD_LREC14_paper_camera_ready.pdf.

[6] K. Dobrovoljc, J. Nivre, *The Universal Dependencies Treebank of Spoken Slovenian*. In: *Proceedings of LREC 2016, Tenth International Conference on Language Resources and Evaluation*. Portorož 2016, http://www.lrec-conf.org/proceedings/lrec2016/pdf/433_Paper.pdf.

[7] J. Gilliéron, E. Edmont, *Atlas linguistique de la France*. Champion, Paris 1902-1910.

[8] J. Gilliéron, J. Mongin, *"Scier" dans la Gaule romane du Sud et de l'Est*. Champion, Paris 1905.

[9] J. Gilliéron, M. Roques, *Études de géographie linguistique d'après l'"Atlas linguistique de la France"*. Champion, Paris 1912.

[10] J. Gilliéron, *L'aire "clavellus" d'après l'"Atlas linguistique de la France"*. Beerstecher, Paris 1912.

[11] J. Gilliéron, *Généalogie des mots qui désignent l'abeille d'après l'Atlas linguistique de la France*. Champion, Paris 1918.

[12] H. Goebl, *Dialektometrische Studien. Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. Niemeyer, Tübingen 1984.

[13] H. Goebl, *Atlante linguistico del ladino dolomitico e dei dialetti limitrofi*. Reichert, Wiesbaden - Strasbourg 1998-2012.

[14] H. Goebl, *Regards dialectométriques sur les données de l'Atlas linguistique de la France (ALF): relations quantitatives et structures de profondeur*. Estudis Romànics, XXV, 2003, 59-121.

[15] H. Goebl, *La dialettometrizzazione integrale dell'AIS. Presentazione dei primi risultati*. Revue de linguistique romane, LXXII, 2008, 25-113.

[16] K. Jaberg, J. Jud, *Sprach - und Sachatlas Italiens und der Süschweiz*. Ringier, Zofingen 1928-1940.

[17] W. Labov, S. Ash, C. Baberg, *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Walter de Gruyter, Berlin 2008.

[18] P. Ladefoged, *Preserving of sounds of disappearing languages*. UCLA Linguistics Dept., Los Angeles (CA) 2002, http://linguistics.ucla.edu/people/ladefoge/Preserving%20sounds.pdf,2002,27-38.

[19] M. Loporcaro, T. Paciaroni, *Four-gender systems in Indo-European*. Folia Linguistica XLV, 2011, 389-434.

[20] L. Lorenzetti, *Un decennio di studi linguistici sui dialetti del Lazio: bilanci e prospettive*. In: C. Giovanardi, F. Onorati (eds.), *Le lingue der monno*. Aracne, Rome 2007, 197-215.

[21] L. Lorenzetti, *Sull'innalzamento di /a/ tonico nei dialetti costieri del Lazio meridionale*. L'Italia dialettale, LXXVII, 2016, 167-82.

[22] B. MacWhinney, *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah (NJ) 2000.

[23] S. Rosen, P. Howell, *Signals and Systems for Speech and Hearing*. Brill, Bingley 2011.

[24] G. Sanga, *Sistema di trascrizione semplificato secondo la grafia italiana*. Rivista italiana di dialettologia, I, 1979, 167-176.

[25] G. Schirru, *Assimilazione permansiva e categorie nominali in un dialetto del Molise*. In: A. De Angelis (ed.), *I dialetti italiani meridionali tra arcaismo e interferenza*. Centro di Studi Filologici e Linguistici Siciliani, Palermo 2008, 291-309.

[26] G. Schirru, *La propagginazione in un dialetto della Valle di Comino*. In: F. Avolio (ed.), *Lingua e dialetto tra l'Italia centrale e l'Italia meridionale. I dialetti della Media Valle del Liri e delle zone limitrofe*. Arte Stampa Editore, Roccasecca 2013, 78-91.

[27] C. Vignoli, *Vernacolo e canti di Amaseno*. Società Filologica Romana, Roma 1920.

[28] M. Weisser, *Speech act annotation*. In: K. Aijmer, Ch. Rühlmann (eds.), *Corpus Pragmatics. A Handbook*. Cambridge University Press, Cambridge 2015, 84-116.

[29] V.R. Yow, *Recording Oral History: A Guide for Humanities and Social Sciences*. Rowman Altamira, Wanut Creek (CA) 2005.

Emanuela Cresti* - Massimo Moneglia* - Alessandro Panunzi*

THE LABLITA CORPUS & THE LANGUAGE INTO ACT THEORY: ANALYSIS OF *VITERBO* EXCERPTS

Abstract. – The LABLITA Corpus is aligned at the utterance level, characterized by high diaphasic variation, and organized with respect to a fine-grained metadata grid (CHAT and IMDI standards). The Language into Act Theory (L-AcT) provides a theoretical framework for the LABLITA collection, assuming a pragmatic basis for Information Structure and a strict correspondence between prosodic and information units. L-AcT provides criteria for Information Structure tagging in the heavily annotated, informal sub-part of the corpus for cross-linguistic comparison (available in the DB-IPIC database). The paper demonstrates how the Viterbo excerpts can be annotated in accordance with L-AcT and how the LABLITA Corpus has been exploited for comparative studies[1].

1. The LABLITA Corpus of Spoken Italian

The Language into Act Theory (L-AcT), which has been in development since the early eighties, aims at providing a framework suitable for the corpus-based study of spontaneous speech [6]. This framework constitutes the scientific background for the collection and annotation of the LABLITA corpus of Spoken Italian, the organization of which entails a set of variation parameters which are considered relevant to the purpose of representing spontaneous speech variation [2, 14]:

- *Register variation*: interactions characterized by regulated transactions (generally corresponding to a formal usage of language) vs. non-regulat-ed, informal interactions

[1]This paper is due to the joint work of the authors. In more detail, Emanuela Cresti wrote section 3, Massimo Moneglia wrote section 2, Alessandro Panunzi wrote sections 1 and 4.

*LABLITA - University of Florence - Piazza Savonarola, 1 - 50132 Florence (Italy).

- *Social context variation*: public vs. private/family interactions
- *Dialogue structure variation*: dialogues or multi-dialogues vs. monologues
- *Channel variation*: face-to-face vs. media vs. telephone
- *Written to be spoken*: transcriptions of movie dialogues

The corpus building strategy was pursued over a long term period, making use of occasional recording opportunities by students and researchers at the University of Florence. The strategy consisted of populating each of the fields with samples for different speakers, genres, topics, and tasks, allowing for the representation of any spoken event. The corpus primarily consists of face-to-face interactions and limits the sampling of formal and media contexts to their real weight in the spoken universe [17].

The recordings gathered in the corpus date from 1965 (Stammerjohann Corpus) to the present and collect samples from sessions recorded mostly in the Florence area. The corpus is anonymized and delivered in acoustic files (.wav PCM) which have been orthographically transcribed according to the CHAT/LABLITA format, and aligned via terminated units to the acoustic sources using WINPITCH. Each sample is accompanied by a rich set of metadata in the CHAT and IMDI formats. Table 1 presents an overview of the corpus.

The sampling strategy does not differentiate for speaker variation, but records sex, geographical origin, education, profession, and mutual relations for each speaker in the session metadata. For each recorded event, metadata also details situation, topic, task, and, optionally, genre. Furthermore, the acoustic quality is noted, based on the reliability of F0 computation. The sample length was determined so as to avoid over-representing individual contexts. Informal samples contain about 1,500 words, while formal samples range from 3,000 to 4,000 words. This resource, partially delivered as a file system within C-ORAL-ROM [10], will shortly be made available online as a database and released in an Open Data format (GrAF, RDF), ensuring interoperability. The corpus samples and metadata populating the database allow the generation of

Table 1 – Corpus Design Matrix of the LABLITA Corpus.

Diaphasic Variation of LABLITA CORPUS			Samples	Words	Ref.Units
Free	Family/Private	Monologues	28	55,141	5,795
turn	Context	Dialogues/Multi-dialogues	140	254,085	39,593
taking	Public Context	Monologues	1	1,723	198
		Dialogues/Multi-dialogues	39	67,434	5,842
total			**208**	**378,383**	**51,428**
Regulated	Family/Private	Monologues	1	3,023	193
turn	Context	Dialogues/Multi-dialogues	28	52,915	5,713
taking	Public Context	Monologues	38	105	4,678
		Dialogues/Multi-dialogues	40	186,103	12,944
total			**107**	**242,308**	**23,528**
Broadcasting			49	131,539	12,593
Telephone			74	32,541	4,445
Total Adult Corpus			**438**	**784,771**	**91,994**
Cinema (written to be spoken)			189	70,414	N.C.
Talking to Children			276	260,595	N.C.
Total			**903**	**1,115,780**	N.C

balanced sub-corpora suitable for comparative studies in different areas (*e.g.* diachrony of spoken variety, written vs spoken, cross-linguistic comparisons).

The L-AcT framework has been tested extensively on large spoken romance corpora (C-ORAL-ROM; C-ORAL-BRASIL, [28]) and in comparative studies. L-AcT focuses on the interface between information structure and prosody and this relation is assumed to be the essential property of speech; for this reason, it is the main information to be specified in spoken corpus annotation.

This paper shows how the L-AcT framework can coherently capture the basic properties of Information Structure and the pragmatic nature of utterances when applied to the three Viterbo excerpts. These properties are generally shared by spoken languages and for this reason we consider L-AcT to be viable for application in large, corpus-based, typological studies. For instance, C-ORAL-ROM shows that the percentage of verbless utterances vary along the corpus design in the same way in the four Romance languages and simple verbless utterances specifically characterize interactive dialogues. In Section 2 we will briefly present the L-AcT framework and the system currently em-

ployed for the annotation of the LABLITA Corpus. In Section 3 the analysis of the Viterbo excerpts at the levels of information structure, illocution, and syntax will be presented, showing the added value of this analysis. The exploitation of different LABLITA samplings for language comparison in the spoken variety is summarized in Section 4.

2. L-AcT FRAME AND THE LABLITA FORMAT

2.1. *Reference units and information units in L-AcT.*

L-AcT focuses on information structure and relies on two principles: a) the pragmatic basis of Information Structure; b) a strict correspondence between prosodic and information units. The corpus-based study of Information Structure, like the study of syntactic or semantic properties, requires the selection in the speech continuum of reference units, for which specific linguistic relations hold. Within a pragmatic tradition [3, 18, 27], we assume that the *utterance*, defined as the counterpart of the speech act [1], must be considered the primary reference unit for speech analysis. We also assume that utterance boundaries can be identified by *perceptually relevant prosodic terminal* breaks [12, 16], rather than by semantic or syntactic factors[2].

However, as large corpus-based analyses on Romance and English languages have shown, in monologues, narrative activities, and in formal speech, terminated prosodic sequences are frequently long and may contain more than one autonomous speech act. This kind of sequence, called a *stanza* in L-AcT, is not structured, but, rather, each component is added to another, like a "work in progress" during the speech process [5]. Therefore, terminal breaks signal a second type of reference unit, where the strict *terminated sequence* / *speech act* correspondence does not hold. The pragmatic counterpart of a stanza is not one act, but a sequence of linked speech acts referred to as a Bound Comment (COB).

The one-to-one correspondence of *speech act* / *terminated prosodic sequence* is also disregarded when one single Prosodic Pattern ending with a

[2]See [13, 28] and references therein for the evaluation of agreement on prosodic break detection.

terminal break corresponds to two or three speech acts (Multiple Comments, CMM) which participate in the same pragmatic activity. Each one of the chained CMMs performs its own illocutionary force, but from a pragmatic point of view they are interpreted in a unitary way as an *illocutionary pattern*. For instance, *alternative questions* are the illocutionary result derived from the combination of two questions. The possibility of patterning different speech acts into one pragmatic activity makes use of prosody, which allows the framing of illocutionary chains into a prosodic model, giving rise to effects that require more than one speech act to be achieved. *List*, *reinforcement*, *alternative question*, *comparison*, and *functional recall* are the most common illocutionary patterns. In all cases, however, the reference units are recoverable in the speech flow by the detection of terminal breaks. The three Viterbo excerpts show both utterances and stanzas, displaying these core notions of spontaneous speech analysis.

Within L-AcT, terminated units are segmented into the information units, whose types record differential pragmatic properties and correspond necessarily to prosodic unit types, which is in accordance with an overall principle of isomorphism between prosodic units and information units. The Comment is the core information unit, since its role is the expression of the illocutionary force, and thus it is necessary and sufficient to accomplish an utterance. Given that the Comment represents a pragmatic modification of the world by the speaker, it automatically conveys new information. Within reference units, other optional information units may support the Comment. L-AcT assumes that they are functionally differentiated and form an Information Pattern with the Comment in accordance with its Prosodic Pattern [6, 24]. Thus, the information relations are not limited to the main "Topic / Comment" opposition, as is generally considered. The correspondence between prosodic and information units leads to the expansion of the Information Structure theory to include a closed, but larger, list of information functions (reported in Table 2). The Viterbo excerpts contain examples of two of the main information patterns: Topic / Comment and Comment / Appendix.

Table 2 – Tagset of Information Units in the L-AcT framework.

Type of Unit	Name	Tag	Definition
Textual	Comment	COM	Accomplishes the illocutionary force of the utterance. It is necessary and sufficient for the performance of the utterance.
	Topic	TOP	Identifies the domain of application for the illocutionary act expressed by the comment, allowing a cognitive reference to the speech act. It allows the utterance to be displaced from the context (linguistic and non-linguistic).
	Appendix of Comment	APC	Integrates the text of the Comment and concludes the utterance, marking an agreement with the addressee.
	Appendix of Topic	APT	Gives a delayed integration of the information given in the Topic.
	Parenthesis	PAR	Inserts information into the utterance with a meta-linguistic value, having *backward* or *forward* scope.
	Locutive Introducer	INT	Expresses the evidence status of the subsequent locutive space (simple or patterned) marking a shift in the coordinates for its interpretation.
	Multiple Comment	CMM	Constitutes a chain of Comments which form an *illocutionary pattern i.e.* an action model which allows the linking of at least two illocutionary acts, for the performance of one conventional rhetoric effect.
	Bound Comment	COB	A sequence of Comments, which are produced by progressive adjunctions following the flow of thought. It forms a *Stanza* out of any informational model.
Dialogic	Incipit	INP	Opens the communicative channel, bearing a contrastive value and starting a dialogic turn or an utterance.
	Conative	CNT	Pushes the listener to take part in the Dialogue, or stops his uncollaborative behavior.
	Phatic	PHA	Controls the communicative channel, ensuring its maintenance. It stimulates the listener toward social cohesion.
	Allocutive	ALL	Specifies to whom the message is directed keeping his attention and forming a cohesive, empathic function.
	Expressive	EXP	Works as an emotional support, stressing the sharing of a social affiliation.
	Discourse Connector	DCT	Connects different parts of the discourse, indicating its continuation.

2.2. *The CHAT/LABLITA format.*

The LABLITA transcription format is a version of the CHAT format [20] combined with the tagging of terminal and non-terminal prosodic breaks [10]. The transcriptions are orthographic and capital letters are employed only for proper names. Speakers are identified by an asterisk and three capital letters, followed by a colon and one space; each dialogic turn is introduced by the Speaker-Id. Each utterance and each information unit are followed by a space and, respectively, a double or a single slash (//, /). The double slash denotes a terminal prosodic break, signaling the conclusion of a spoken reference unit (utterance or stanza); the single slash marks the internal segmentation of the reference unit within its enclosing information unit. Dependent layers, which are preceded by a % in accordance with the CHAT convention, may be used to add different kinds of information, concerning, for instance, the situation (%sit) or the illocutionary type (%ill). When tagging the information functions of each unit, its slash is marked with the information value using three capital letters in superscript, or with this tag in a layer of PRAAT or WINPITCH.

Fragmentation phenomena are annotated both at the word level and information unit level. The '&' character symbolizes a word fragment, while '+' signals an interrupted utterance and the '[/]' symbol a retraction, such as reformulation. In the LABLITA practice, '[/n]' also specifies the number of retracted words, which should be disregarded in text processing tasks. Overlaps are indicated with angular parentheses ('<word>'). A dollar tag ('$') at the end of each reference unit in the transcript refers to the aligned audio counterpart.

3. The L-AcT Analysis of the Viterbo excerpts

The excerpts (*Maiale*, *Oro*, *Vestiti*) seem to belong to one spontaneous dialogue between a man (UOM) and a woman (DON). The following analysis is mainly concerned with the information structure and illocutionary evaluation; however, L-AcT foresees that syntactic relations hold inside information units, therefore the analysis will also be augmented by the description of the

syntactic type of each information unit.

Maiale
*DON: è ritornato giù //^{COM}\$ ancora stavano &co +\$ colato giù a catinella /^{COB} e mh /^{TMT} così /^{DCT} nun moriva il maiale //^{COM}\$

%ill: [1] Report (Assertive); [2] Description (Assertive); [3] Description (Assertive)

The turn is composed of two reference units: an utterance and a stanza, respectively. The utterance is a single Comment unit, accomplishing an assertive illocution, probably based on the memory of a killing of a pig, producing a reporting effect[3]. A verbal predicate *(è ritornato giù*) constitutes the verbal clause and fulfills the Comment function. After the first utterance, we find an interrupted utterance concluded by a word fragment (*ancora stavano &co*), and, when the prosodic planning resumes, the second reference unit corresponds to a stanza, composed of two COBs. The syntactic filling of the first COB is a past participle (*colato giù a catinella*). The illocution is assertive, with a descriptive function. In our opinion, a hesitation begins the second COB (*e mh*), which in the default analysis is marked as coordinate conjunction. The first COB is connected to the final COM by a Discourse Connector (*così*). The COM performs a description (*nun moriva il maiale*) through a negative verbal clause.

Stanzas are not strictly governed by pragmatic principles, but rather follow strategies of textual construction [7] and typically occur when action goals are weak. Thus, the expression of semantic content plays a greater role, as happens here in a reporting / description context. For these reasons, the illocutionary force of COBs mostly belongs to the assertive and expressive illocutionary classes. IPIC shows that roughly 90% of terminated units are utterances and 10% stanzas in both Italian and Brasilian Portuguese. Stanzas mostly characterizing monologues (around 25% according to [26]). The following are Italian and English examples of stanzas, taken from the LABLITA

[3]The assignment of an illocutionary value is made following the pragmatic repertoire in [9], however, the lack of metadata and context prevents a well-based interpretation. On the contrary, the tagging of information units is fully reliable according to L-AcT principles.

and Santa Barbara Corpora, both broader than the two COB stanza of Maiale (as are most stanza):

(1) *ROS: c'è i' puro cotone /[COB] c'è lino e seta /[COB] che casca bene / [COB] e po' ci si sta freschi //[COM] [ifamdl07]

(2) *PAM: and we think of her in Jezebel /[COB] we think of her /[SCA] you know /[AUX] smoking /[COB] cigarette smoke into the faces of William Holden /[COB] and [/1][EMP] and the like //[COM] [afamdl02]

Oro

*UOM: l'oro /[TOP] niente //[COM]$
%ill: Question (Directive)
*DON: l' oro //[COM]$
%ill: Taken for granted (Assertive)
*UOM: allora se faceva //[COM]$ <coralli> //[COM]$
%ill: [1] request of confirmation (Directive); [2] suggestion (Expressive)
*DON: <xxx> //[UNC]$
%com: overlapping
*DON: pe' quello +$ ecco //[COM]$ una fila de coralli /[CMM] o una catenina /[CMM] una cosa /[SCA] così //[CMM]$
%ill: [1] none (interrupted); [2] Revelation (Expressive); [3] Description (Assertive)

The *Oro* excerpt is a brief exchange between two speakers. In the first turn, UOM asks whether golden jewelry was used (*l' oro* / *niente* //), performing a nominal utterance with a Topic / Comment structure. In DON's turn, she takes for granted that gold jewelry was indeed used (*l'oro*). In the next turn UOM requests a confirmation (*allora se faceva*) and proposes further jewelry (*coralli*). In the last turn an overlapping occurs that prevents hearing (tagged as unclassifiable). The second part of the turn is interrupted (*pe' quello* +) and finally DON performs un utterance through an interjection (*ecco*). The conclusion of the turn corresponds to an illocutionary pattern: a chain of three patterned nominal Comments (*una fila de coralli* / *o una catenina* / *una cosa così*).

Despite the presence of more than one illocutionary act, this terminated

sequence of CMMs must be considered one single *listing* activity performed by the speaker. The Prosodic Pattern type for list is performed by the speaker as the result of one single intention, thus giving the possibility of two or three illocutionary forces in one utterance. In spontaneous speech, rather frequently, utterances can be performed through patterned chains of speech acts (IPIC data shows a percentage of around 10% of the total utterances, very uniform across corpus variation). The following examples show the most common illocutionary Patterns in Italian[4].

(3) **List:** *FER: sputa la cassetta /^{CMM} la tira fuori /^{CMM} 'un la piglia //^{CMM} [ifamcv15]

(4) **Comparison:** *MAX: questa è un anno prima /^{CMM} e questa è un anno dopo //^{CMM} [ifamcv01]

(5) **Reinforcement:** *MAX: no /^{CMM} 'un ci credo //^{CMM} [ifamcv01]

(6) **Alternative question:** *MAX: li fa lei /^{CMM} o la devo fare io?^{CMM} [ifamdl19]

(7) **Recall + directive act:** *DAN: Daddo /^{CMM} che fai?^{CMM} [ifamcv01]

Paradoxically for speech act theory, illocutionary patterns show that an utterance may not necessarily be characterized by a single illocutionary force.

From a syntactic perspective, *Oro* shows the predominance of verbless clauses, which are fully autonomous since they develop illocutionary values (*l'oro / niente* //; *l'oro* //; *coralli* //; *una fila de coralli* / *o una catenina* / *una cosa così* //).

Vestiti

*UOM: quando v' avevano [/2] ve siete sposati voi /^{TOP} po' &ke [/1] che costume portavate /^{COM} k' vestito portavate //^{APC}$

%ill: Partial question (Directive)

*DON: vestiti //^{COM}$

%ill: Request of confirmation (Directive)

[4]Listening to the audio files the reader can perceive that each CMM can be interpreted in isolation as an independent act, but this results in the loss of the combinatorial meaning conveyed by their Prosodic Patterning.

*UOM: eh //^{COM}$
%ill: Assent (Dialogical Move)
*DON: embè vestiti /^{COM} 'n importa //^{APC}$
%ill: Softening (Expressive)

Vestiti is a balanced alternation of turns between two speakers, in which UOM begins asking a question to DON, who requests confirmation, expressing surprise. UOM voices assent and DON express her underestimation for the subject. The first turn by UOM corresponds to an utterance with a typical Information pattern composed of one Topic unit, starting with a partial retraction (*quando v' avevano* [/2] *ve siete sposati voi*), a Comment, preceded by a word fragment (*po' &ke che costume portavate*), and an Appendix for lexical clarification (*k' vestito portavate*). The linguistic fulfilment of the Topic corresponds to a subordinate clause, the *protasis* of a hypothetic period, while the *apodosis* is the Comment, filled by the main clause.

The second turn by DON is a simple utterance, composed of a Comment accomplishing a request of confirmation, also expressing surprise (*vestiti*). Syntactically it is a verbless clause, simply a noun. After the assent given by UOM (*eh*), DON continues with a Comment / Appendix utterance, composed of the Comment (*embè vestiti*), performing an act of softening, and the Appendix (*n'importa*). We note that the locutive content of the Comment is a verbless clause: a noun focalized by *embè*. The Comment is the core of the utterance and is sufficient to perform it, even if the linguistic content realized in the Appendix is the predicative component within the utterance. Thus, we can appreciate the mismatch between the information organization of the utterance and its syntactic fulfilment. The illocutionary information is sufficient to communicate independently of the syntactic predicate, which is added on as an optional tail. Here, as in a large majority of cases in Italian, the utterance doesn't correlate with a sentence but is a combination of syntactic islands, which are mostly phrases and clauses, pragmatically bound [8].

The LABLITA corpus can provide many examples where the Comment is fulfilled by a noun, but expresses the core of the information content (since it accomplishes the illocution), while the predicative part (performed in Appendix) can be cancelled out, being an unnecessary addition. The reader can

judge this for himself by listening to the audio of (8) and (9).

(8) *ZIA: tutto mio cognato /^{COM} assomiglia //^{APC} [ifammn01]
(9) *VER: forse /^{TOP} la Maddalena / ^{COM} fa dei biscotti // ^{APC} [ifamdl14]

L-AcT's reliability in assigning information values to chunks identified by prosody is guaranteed by standard operative conditions. On isolating the Comments in the audio tracks of the patterned Viterbo utterances, they can still be pragmatically interpreted since they express an illocutionary value. Accordingly, the annotation of the illocutionary type is aligned to the Comment unit in the PRAAT file delivered for this workshop. Conversely, the Topic and the Appendix are neither autonomous nor pragmatically interpretable. From a prosodic point of view, the Topic bears a prosodic prominence, which, in both Italian and Brazilian, can correspond to three different profiles [21, 29]. The Appendix has a falling profile, low intensity, and often an inaccurate phonetic execution.

In conclusion, despite the short size, the annotation of the Viterbo excerpts testifies to general features of spontaneous speech such as the high variability of illocutionary activities. The relevance of verbless utterances in spoken Italian goes hand in hand with the low occurrence of verbal predication, while information patterning provides the main structuring device of the utterance. Disfluencies are quite frequent, but nevertheless do not impede the flow of communication.

4. LABLITA Corpus as source for comparative studies

The LABLITA corpus structure, given its large variation and its basic prosodic annotation, easily allows the creation of balanced sub-resources suitable for comparative analyses of linguistic phenomena, with respect to different perspectives and purposes. The Italian corpus in the multilingual C-ORAL-ROM resource [10] is a subset of the LABLITA corpus. C-ORAL-ROM also contains three other corpora of French, Portuguese, and Spanish language. It has been built to ensure the representativeness of realistic spontaneous speech in each language and their mutual comparability, assuming the same design

and the same levels of basic annotation for each sub-corpus: tagging of the prosodic boundaries, utterance-based alignment to the audio source, PoS tagging. The same architecture has later been applied to the C-ORAL-BRASIL resource [28], a fifth Romance corpus of Brazilian Portuguese.

The data collected in the DB-IPIC resource (*DataBase for Information Patterning Interlinguistic Comparison*), in turn, derives from C-ORAL-ROM and C-ORAL-BRASIL. In its first release, this resource brought together texts from the Informal sections of the C-ORAL-ROM Italian corpus (74 texts for 124,735 total words and 21,007 terminated units), which were enriched with the manual annotation of information units (L-AcT tagset) and prosodic alignment. After that, a minicorpus of Brazilian Portuguese (a subset of the C-ORAL-BRASIL corpus; 20 texts, ca. 8,500 words, and ca. 5,500 terminated units) was selected and annotated using the same criteria. To ensure the highest level of cross-linguistic comparability, a second Italian minicorpus has been derived from the broader one. The choice of the recording sessions which compose the two minicorpora maximizes the similarity between these resources, both from the quantitative and the qualitative point of view [26] and allows a comparison of how information structure varies as a function of the language [23]. The same architecture has also been exploited to design two other minicorpora currently under development: one for American English, derived from the Santa Barbara Corpus [4, 15], and one for Spanish, derived from C-OR-DIAL [19].

Over and above the cross-linguistic comparison, a sample of the LABLITA corpus has been extracted with the goal of a diachronic comparison with the Stammerjohann Corpus, which was the first spoken corpus of Italian, collected in Florence by the German linguist Harro Stammerjohann in 1965. This corpus comprises spontaneous spoken interactions collected in familial and working contexts and demonstrates considerable usage of Florentine dialect of the time. A *Corpus for Diachronic Comparison*, derived from the LABLITA corpus, has been specifically designed for reproducing as close as possible the diatopic, diaphasic, and, diastratic varieties of the Stammerjohann Corpus, 30 years later. The comparison led to the evaluation of the incidence of dialect on the Florentine lexicon in this time spam [22].

The LABLITA Corpus has also been used to build the spoken part of the

CorDIC resource (*Corpora Didattici Italiani di Confronto*), a collection of two distinct corpora (one written, one spoken, about 500,000 words in 200 different texts for each corpus) which aim at representing the spoken and written varieties of Italian, specifically designed for didactic purposes [11]. In order to compare spoken and written varieties beyond the lexical level, the annotation of prosodic breaks is crucial: without this, it is hard to conceive of a reliable method for the comparison between the reference units in written (*i.e.* sentences) vs. spoken language.

References

[1] J.L. Austin, *How to Do Things with Words*. Oxford University Press, Oxford 1962.

[2] D. Biber, *Variation across Speech and Writing*. Cambridge University Press, Cambridge 1988.

[3] D. Biber, S. Johansson, G. Leech, S. Conrad, E. Finegan, *The Longman Grammar of Spoken and Written English*. Allyn & Bacon, London 1999.

[4] F. Cavalcante, A. Ramos, *The American English spontaneous speech minicorpus. Architecture and comparability*. Chimera 3(2), forthcoming.

[5] W. Chafe, *Meaning and the structure of language*. Cambridge University Press, Chicago 1970.

[6] E. Cresti, *Corpus di italiano parlato*. Accademia della Crusca, Firenze 2000.

[7] E. Cresti, *La Stanza: un'unità di costruzione testuale del parlato*. In: A. Ferrari (ed.), *Sintassi storica e sincronica dell'italiano. Subordinazione, coordinazione e giustapposizione*. Proc. of the 10th Conf. SILFI, June 30-July 3, Florence 2010.

[8] E. Cresti, *Syntactic properties of spontaneous speech in the Language into Act Theory: data on Italian complements and relative clauses*. In:

T. Raso, H. Mello (eds.), *Spoken corpora and linguistics studies*. John Benjamins Publishing Company, Amsterdam 2014.

[9] E. Cresti, *Per una classificazione empirica dell'illocuzione. Lo stato dell'arte*. In: M. Biffi, F. Cialdini, R. Setti (eds.), *Acciò che il nostro dire sia ben chiaro. Scritti in onore di Nicoletta Maraschio*. Firenze forthcoming.

[10] E. Cresti, M. Moneglia, *C-ORAL-ROM. Integrated reference corpora for spoken romance languages*. John Benjamins Publishing Company, Amsterdam 2005.

[11] E. Cresti, A. Panunzi, *Introduzione ai corpora dell'italiano*. Il Mulino, Bologna 2013.

[12] A. Cruttenden, *Intonation*. Cambridge University Press, Cambridge 1986.

[13] M. Danieli, J.M. Garrido, M. Moneglia, A. Panizza, S. Quazza, M. Swerts, *Evaluation of Consensus on the Annotation of Prosodic Breaks in the Romance Corpus of Spontaneous Speech C-ORAL-ROM*. Proc. of the 4th Conf. LREC, Paris 2004.

[14] T. De Mauro, F. Mancini, M. Vedovelli, M. Voghera, *Lessico di frequenza dell'italiano parlato*. Etas, Milano 2004.

[15] J. Du Bois, W. Chafe, Ch. Meyer, S. Thompson, *Santa Barbara Corpus of Spoken American English, Part 1*. Web Download, Philadelphia 2000.

[16] J. 't Hart, R. Collier, A. Cohen, *A Perceptual Study on Intonation. An Experimental Approach to Speech Melody*. Cambridge University Press, Cambridge 1990.

[17] H. Koiso, T. Tsuchiya, R. Watanabe, D. Yokomori, M. Aizawa, Y. Den, *Survey of Conversational Behavior: Towards the Design of a Balanced Corpus of Everyday Japanese Conversation*. Proc. of the 10th Conf. LREC, Paris 2016.

[18] G. Leech, *The Pragmatics of Politeness*. Oxford University Press, Oxford 2014.

[19] Ph. Martin, *The structure of spoken language. Intonation in romance*. Cambridge University Press, Cambridge 2015.

[20] B. McWhinney, *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associate, Mahwah 2000.

[21] M.M. Mittmann, *O C-ORAL-BRASIL e o estudo da fala informal: um novo olhar sobre o Tópico no Português Brasileiro*. PhD thesis UFMG, Belo Horizonte 2012.

[22] M. Moneglia, N. Binazzi, R. Cella, A. Scarano, A. Panunzi, M. Fabbri, *L'incidenza del lessico fiorentino nella lingua d'uso a Firenze. Un confronto tra il corpus Stammerjohann del 1965 e un corpus di parlato contemporaneo*. Proc. of the 9th Conf. SILFI, June 14-17, Florence 2006.

[23] M. Moneglia, E. Cresti, *The Cross-linguistic comparison of information patterning in spontaneous speech corpora: Data from C-ORAL-ROM ITALIAN and C-ORAL-BRASIL*. In: S. Diao-Klaeger, B. Thörle (eds.), *Linguistique interactionnelle contrastive. Grammaire et interaction dans les langues romanes*. Stauffenburg, Tübingen 2015.

[24] M. Moneglia, T. Raso, *Notes on the Language into Act Theory*. In: T. Raso, H. Mello (eds.), *Spoken corpora and linguistics studies*. John Benjamins Publishing Company, Amsterdam 2014.

[25] C. Nicolas Martinez, *Cor-DiAL*. Ediciones Liceus, Madrid 2012.

[26] A. Panunzi, M.M. Mittmann, *The IPIC resource and a cross-linguistic analysis of information structure in Italian and Brazilian Portuguese*. In: T. Raso, H. Mello (eds.), *Spoken corpora and linguistics studies*. John Benjamins Publishing Company, Amsterdam 2014.

[27] R. Quirk, S. Greenbaum, G. Leech, J. Svartvik, *A Comprehensive Grammar of the English Language*. Allyn & Bacon, London/New York 1985.

[28] T. Raso, H. Mello, *C-ORAL-BRASIL I: Corpus de referência de português brasileiro falado informal*. UFMG, Belo Horizonte 2012.

[29] S. Signorini, *Topic e soggetto nell'italiano parlato*. PhD thesis University of Florence, Florence 2005.

Philippe Martin*

PROSODIC ANNOTATION IN ADVERSE RECORDING CONDITIONS

Abstract. – Advances in acoustic speech analysis makes now possible to perform prosodic annotation on acoustically degraded recordings, such as those found in old entertainment or documentary movies. The five examples provided for the convention provide good examples of such recordings. This paper summarizes the most common problems found in prosodic annotation in these conditions, and describe the software program WinPitch, which integrates a set of functions designed to handle and correct erroneous fundamental frequency tracking, as well as a user programmable annotation system. The annotation phonetic parameters can be easily transferred to a spreadsheet package such as Excel, in order to perform a comprehensive analysis of phonetic data obtained from the annotation.

1. Introduction

Advances in acoustic speech analysis makes now possible to perform prosodic annotation on acoustically degraded recordings, such as those found in old entertainment or documentary movies. The software WinPitch [12] described here is specially designed for this purpose. It integrates various user friendly functions in order to obtain reliable pitch curves and realize prosodic annotations in adverse acoustic conditions.

Using this program, the sequence of operations for a typical annotation session are as follows:

1. Conversion of the original sound format if necessary into 22050 Hz sampling rate, mono, 16 bits, more suitable for speech recordings.

2. Alignment of fundamental frequency and narrow band spectrogram frequency scales for visual validation and correction of pitch curves.

3. Selection of adapted multi-methods fundamental frequency tracking on selected speech segments.

*LLF, UFRL - Université Paris Diderot - 5, rue Thomas Mann -75205 Paris Cedex 13.

4. Annotation of fundamental frequency variations using categories defined by prosodic models.

5. Statistical analysis (if applicable) of fundamental frequency acoustic parameters related to preselected phonological.

These operations are illustrated below on some of the examples proposed for the convention *oro*, *maiale*, *CG5*, *ayoreyo* and *vestiti*.

2. Format conversion

Although the most common commercially available sampling frequency used to convert analog speech signal into numerical data is either 44100 Hz or 48000 Hz, the resulting bandwidths of 22050 Hz or 24000 Hz appear quite excessive for speech applications. Indeed, 22050 Hz is a more appropriate sampling frequency to adequately represent the highest speech sounds frequency as found in the fricative [s], about 8000 Hz. A higher conversion rate induces unnecessary computation and memory load and should be reduced before further processing.

Likewise, stereo recordings using a single source microphone should be converted into a mono format. Furthermore, as sampling formats of 24 bits are more appropriate for large sound mixers using 64 different sound sources or more. 16-bit formats give a more than enough signal to noise ratio of about 90 dB, well below the actual noise level of microphone preamplifiers.

Compressed formats such as mp3, wma, aiff, etc., should be avoided as well, as they introduce frequency and phase distortions in their coding process. The uncompressed wav format should therefore be preferred.

To convert the original sound files, whether in wav, mp3, wma, etc. format, or sound information imbedded in video formats such as mp4, mov, flv, wmv, etc., WinPitch has an integrated format converter, acting directly during a file load operation. The user has simply to select the output format (fig. 1) while reading a sound file to operate the conversion. Praat TextGrid files can also be read and converted directly.

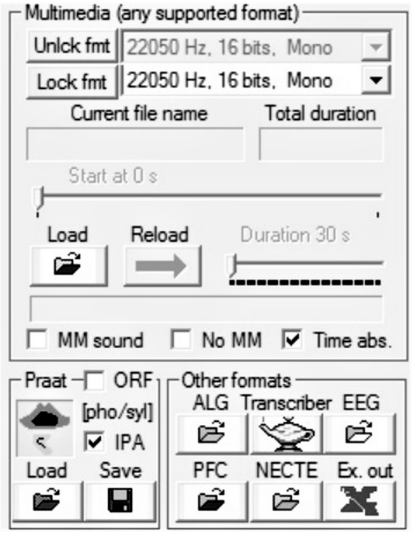

Fig. 1. – Format conversion reading TextGrid files.

3. Alignment of F0 and Spectrogram frequency scales

Although not frequently used, one of the best ways to validate the fundamental frequency curve of a given speech segment consists to compare the melodic curve with the first harmonic of the corresponding narrow-band spectrogram (or another harmonic if the first is not visible, has been filtered or is obscured by another sound source spectral component). Indeed, except in rare case of a very fast changing fundamental frequency (in the transition from a voiced stop to a high intensity vowel), the spectrum first harmonic is totally reliable, although not allowing a precise measurement of the frequency value due to the frequency resolution of the spectrogram (in the order of 20 to 30 Hz). For an easy comparison, WinPitch provides an automatic alignment function for both F0 tracking and the spectrogram frequency scale, so that both curves should be exactly superposed. Errors in F0 detection can then be easily detected visually (fig. 2).

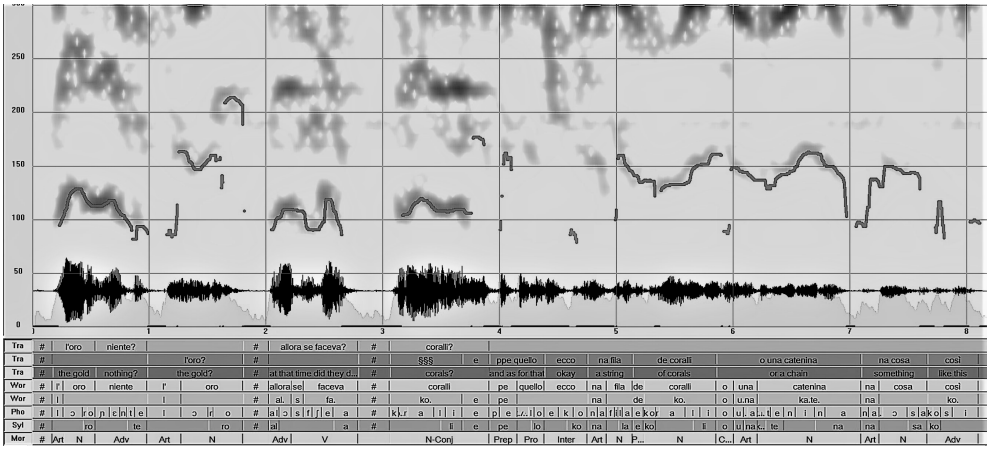

Fig. 2. – Superposition of the fundamental frequency curve and the first harmonic of the corresponding narrow-band spectrogram (File *oro.wav* and *oro.wp2*. *.wp2 files are created by WinPitch form the TextGrid transcription). This allows the user to visually check the validity of the fundamental frequency curve, and eventually select an alternate F0 tracking method, or even adopt a manual graphic selection to describe the pitch movement as observed from the spectrogram data.

4. Multi-method fundamental frequency tracking

4.1. *F0 detection problems.*

Measurement of the fundamental frequency is particularly sensitive to recorded speech signal distortions due to:

1) Poor signal to noise ratio.

2) Filtering of low frequencies (eliminating low harmonics for male voices).

3) Various distortions due to room echo during the recording.

4) Encoding in formats such as mp3 or wma with an excessive compression rate.

5) Presence of external sound sources (car engine, overlapping speech segments, etc.).

6) Presence of creaky segments where the fundamental frequency is not really defined.

Praat, the de facto standard in the speech analysis domain [10] does not really appear optimized for fundamental frequency tracking in difficult conditions. From the evaluation of the most frequent causes of F0 errors, most common problems could be listed as:

1) The use of microphones with a poor low-frequency response, implying the absence of the first harmonics in the spectrum (especially for male voices).

2) The presence of a strong echo in the signal depending on the dimensions and the material of the recording room, producing harmonic blurs. An unvoiced consonant can often appear as voiced due to the spurious continuity observed on the first harmonic (fig. 3).

3) An insufficient recording level, often due to too great a distance between the microphone and the speaker, resulting in a low signal to noise ratio.

4) The use of AVC (automatic volume control) in the recording process, which distorts the speech intensity curve and indirectly produces errors in the evaluation of vowel spectra.

5) The presence of multiple sound sources, in particular generated by low-frequency engines (ex. lighting noise in the recording room, etc.), poor microphone cable isolation, or overlapping speech.

6) Excessive compression of the speech signal (*e.g.* wma or mp3 with a high rate of compression), giving, when converted into a waveform, shifted spectral peak frequency values which are undesirable for spectrum-based algorithms such as Cepstrum, Spectral Comb... (fig. 6).

All these sources of errors are linked to the recording conditions. While some of these conditions cannot be easily modified (recording among a crowd of people for example), many problems can be avoided by taking some elementary precautions during the recording sessions:

1) Using a microphone with a satisfactory low frequency response.

2) Positioning the microphone far from reflecting surfaces (windows, bare walls, etc.).

3) Positioning the microphone close to the speaker's lips if possible (30 cm is the standard used by professional sound engineers).

4) Deselecting the AVC (Automatic Volume Control).

5) Using a wav format with a 22050 Hz sampling rate, recorded in mono and 16-bit format.

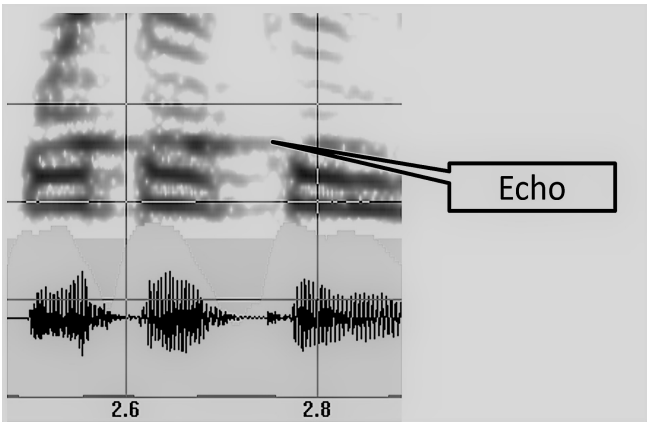

Fig. 3. – An example of an echo creating training harmonics, especially visible on the first and third harmonic on this figure.

Obviously, once the recording has been made, it is often impossible in practice to retrieve exactly the same speech production conditions, and using poor quality recording requires tedious and lengthy efforts to obtain a reasonably exploitable pitch curve, as detailed below. Although professional recording engineers are rarely part of spontaneous speech recording projects, it is relatively simple to use speech analysis software such as WinPitch, which allows visual monitoring of a narrow band spectrogram displayed in real time, so that moderately trained operators in phonetics can immediately identify visually and fix the potential recording problems described above.

4.2. *Fundamental frequency tracking.*

To cope with some if these recording conditions, a multi-method fundamental frequency tracking has been integrated in WinPitch. Since all pitch tracking algorithms, no matter their operating principle, are so far prone to errors in adverse recording conditions, and given that for a particular speech segment some algorithms are less prone to errors than others, eight different pitch tracking routines are implemented in WinPitch in order to display a more reliable fundamental frequency curve. These methods are the spectral comb [5], the spectral brush [6], autocorrelation in 3 flavors: stan-

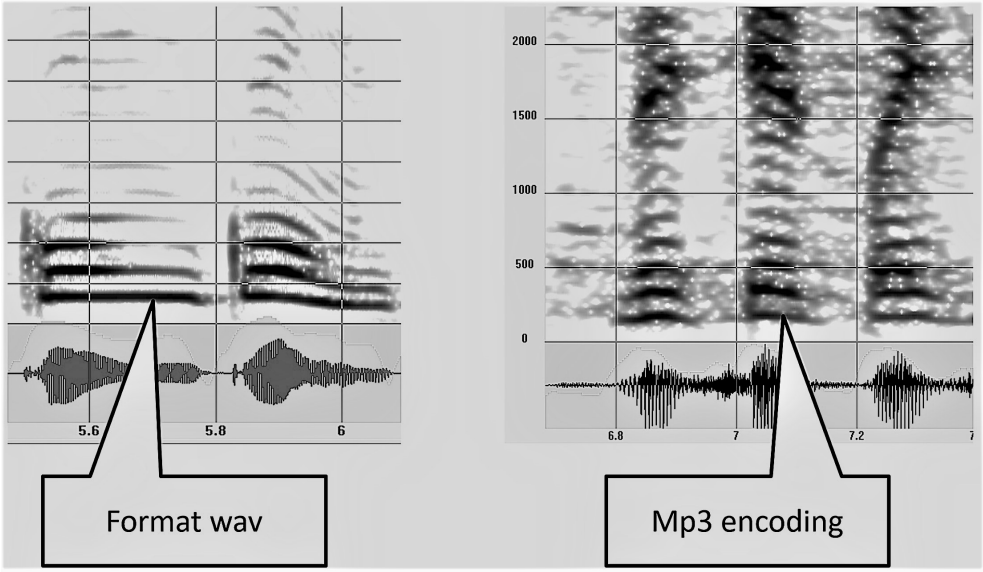

Fig. 4. – A comparison between narrow band spectra of a speech segment recorded in wav format (left) and encoded in mp3 format (right).

dard, Praat [2], Yin [3], AMDF, Cepstrum [9] and harmonic selection.

WinPitch also includes a scanning feature allowing a quality analysis of the recording in terms of fundamental frequency coherence, transition and presence of creak (diplophonia and vocal fry varieties). The implemented creak detection algorithm [7] proceeds by comparing the number of harmonics in successive speech segments of about 64 ms. A sudden doubling of this value generally indicates the presence of creak.

The software allows these algorithms and their related parameters to be independently applied on user-defined segments of the speech wave, in order to use the most satisfactory scheme for a given speech section of the recording. These various methods give globally comparable results on good quality recordings. However, for lower quality recordings, the main problems of analysis are:

1) A fast-rising melodic curve. By nature, spectral based methods (such as the Spectral Comb) evaluate the fundamental frequency of a signal from

the harmonic structure obtained from successive Fourier transforms. This requires a relatively long analysis time window (in the order of 32 ms or 64 ms for male voices with a 100 Hz F0), which in turn prevents a correct tracking for fast rising or falling F0 values. Autocorrelation-based methods such as Yin may also exhibit this limitation even if they are based on the time domain (the reason stems from the time window usually selected for autocorrelation).

2) The presence of pseudo-harmonics due to the presence of echo in the recording room can adversely affect frequency-based methods. Evaluation of the fundamental frequency of a signal by the Comb method for example is based on the detection of at least two consecutive harmonics. Echo produced by some harmonics, depending on the dimensions of the room, can generate trails of certain harmonics that have enough duration to make an unvoiced segment appear voiced and confuse the algorithm detecting these components. It is quite difficult to automatically differentiate this spectral configuration from examples where a low frequency filtering would provoke spectral patterns similar to the ones generated by an echo (fig. 3).

In practice, the spectral comb is the default method of choice for most recordings. Autocorrelation can then be used on segments exhibiting (on the narrow band spectrogram) a noise source with their harmonics, and is more resistant to echo effects and spectral peak shifts and smearing. However, when both the spectral comb and autocorrelation fail, the AMDF method can provide a satisfactory local solution. Extreme cases require pre-selection of the harmonics by the user followed by a spectral comb analysis. This pre-selection, again operated on the narrow band spectrogram, should of course discard unwanted noise components.

A more tedious process consists in manually correcting pitch markers directly on the speech wave (dedicated signal editing functions are provided for this operation), after automatically placing default pitch markers.

4.3. *F0 curve cleaning.*

To apply one of these methods, the user first selects a F0 tracking method in the command window. Then a time segment is selected on screen with the mouse guided by visual inspection of a narrow band spectrogram

displayed underneath. By releasing the left button of the mouse, the corresponding segment of the signal is automatically reanalyzed with the selected method, replacing F0 data with the new values obtained.

The new F0 curve segment is displayed in a color specific to the tracking method chosen, so that the user can visually identify on the overall F0 curve the tracking method pertaining to a specific time segment. Furthermore, by moving the cursor on screen, the corresponding command box corresponding to the F0 tracking method used for the wave segment defined by the cursor is displayed dynamically in the command box, together with all the parameter values used for the chosen tracking method (fig. 5 and fig. 6).

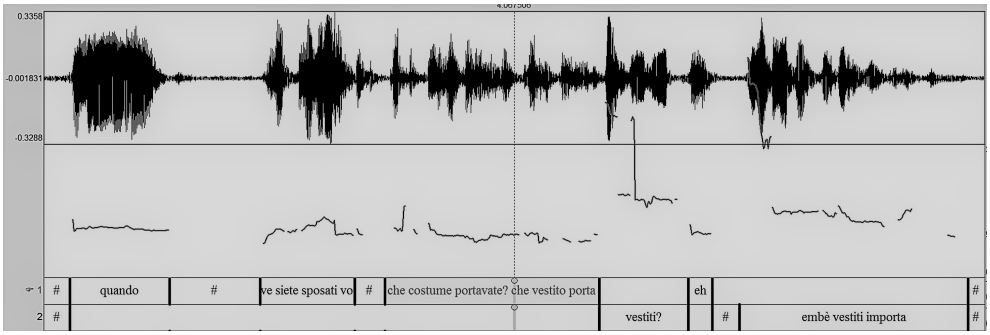

Fig. 5. – The melodic curve of the example *vestiti* obtained with Praat (File *vestiti.TextGrid*).

In fig. 6, the speech segment was analyzed with the autocorrelation method, at the exception of short segments on *voi # che costume* (spectral comb), *vestiti* (AMDF) and *importa* (comb and AMDF).

5. Prosodic annotation

WinPitch integrates prosodic annotation functions which allow the user to graphically represent segments of melodic variations with linear a piecewise functions (*i.e.* one or more straight lines). These linear segments approximate phonetic variations of phonological categories, which are programmable and can be adapted to virtually any phonological model of sentence intonation.

In the incremental prosodic structure model for instance [8], sentence intonation uses melodic contours categories C0 (terminal falling and low for

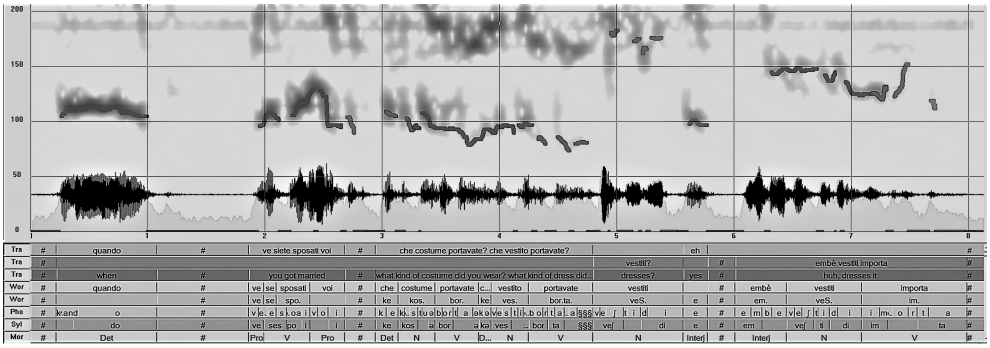

Fig. 6. – The melodic curve of the example *vestiti* after applying the F0 multi-method pitch tracking. F0 is now correctly displayed for the female voice and wrong transitions are removed on *vestiti (*Files *vestiti.wav* and *vestiti.wp2)*.

declarative sentences), C1 (rising above the glissando threshold), C2 (falling above the glissando threshold) and Cn (rising or falling below the glissando threshold). Once the glissando threshold has been defined [11], the labelling of each segment placed graphically by the user to fit the actual melodic variation, the corresponding phonological category is automatically assigned by the software to the segment, together with the color coding eventually assigned beforehand (fig. 7).

Using ToBI notation [1, 4] such as H*H%, L*L-, L+H* for F_ToBI is equally easy to predefine and actual acoustic melodic movements are then approximated in terms of targets.

Once the operations of prosodic annotation have been achieved, acoustic parameters of each annotated melodic segment are transferred to the program Excel (or any other similar spreadsheet program) in a single mouse click for further analysis (fig. 8).

The table shown on fig. 8 is generated with one single mouse click, and displays for each contour: color, graphic width, start and end time in seconds, start and end fundamental frequency values in Hz, contour duration in seconds, frequency range in Hz, start and end frequency values in semitones, start and end intensity values, glissando value reported to the corresponding glissando threshold, glissando ration (*i.e.* glissando value / glissando threshold).

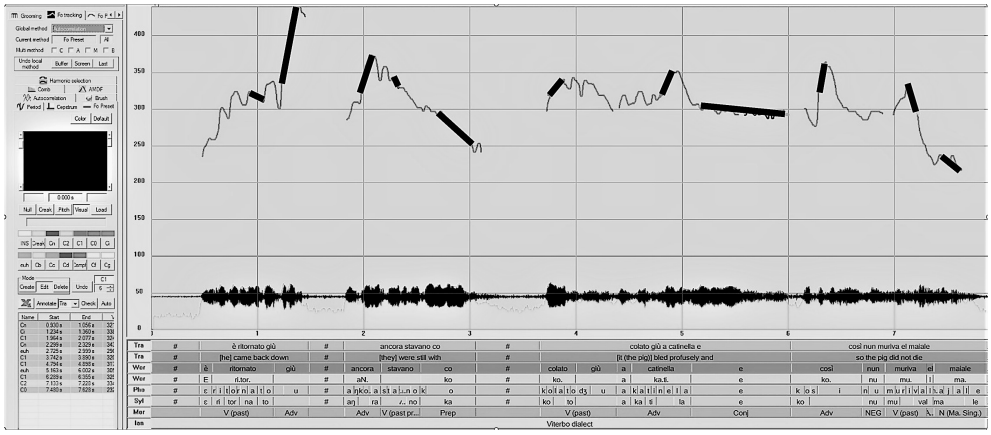

Fig. 7. – Prosodic annotation with linear variations C1, Ci, hesitation, C1, C, C1, hesitation, C0 (according to the incrementation prosodic structure categories) on the example *maiale* (File *maiale.wav* and *maiale.wp2*) using predefined categories (left window). Classes of contours are color coded: red for final conclusive C0, pink for rising above the glissando threshold C1, brown falling above the glissando threshold C2. These categories are user programmable.

6. Statistical analysis

After conversion into an Excel file, many statistical tests can be executed from the set of acoustical data, pertaining to the starting and ending points of each linear segment fit on the actual fundamental frequency values: fundamental frequency, intensity and time, as well as segment duration and glissando values (fig. 8). The large set of statistical function available on Excel (or any suitable statistical package easily interfaced with table files in csv or tab formats) allows to investigate the validity of some phonological model implicitly defined in the annotation.

7. Conclusions

For a large variety of recording conditions found today in speech corpora, and even more for old documents such as movie soundtracks or cassette recordings, the generation of a reliable fundamental frequency curve is not a

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Name	Width	T 1 [s]	T 2 [s]	F0 1 [s]	F0 2 [Hz]	Duration [s]	Range [Hz]	F0 1 [ST]	F0 2 [ST]	Int 1 [dB]	Int 2 [dB]	Glissando [st/s]	Glissando ratio
2														
3	Cn	6	0.93	1.056	321	311	0.126	-10	20	19	34	28	9//10	0.9
4	Ci	6	1.234	1.36	338	438	0.126	100	21	25	35	29	40//10	4
5	C1	6	1.964	2.077	324	373	0.113	49	20	22	24	35	32//12	2.666
6	Cn	6	2.299	2.329	343	329	0.03	-14	21	20	33	32	24//172	0.139
7	euh	6	2.725	2.999	296	254	0.273	-42	18	16	36	27	18//2	9
8	C1	6	3.742	3.89	320	340	0.147	20	20	21	38	33	11//7	1.571
9	C1	6	4.794	4.898	317	352	0.104	35	19	21	30	36	23//14	1.642
10	euh	6	5.163	6.002	305	293	0.838	-12	19	18	36	21	15//0	10
11	C1	6	6.289	6.355	325	364	0.065	39	20	22	34	26	37//37	1
12	C2	6	7.133	7.228	334	294	0.095	-40	20	18	30	35	28//17	1.647
13	C0	6	7.48	7.628	232	218	0.147	-14	14	13	31	25	12//7	1.714
14														

Fig. 8. – Table of fundamental frequency, duration, intensity and glissando values derived automatically from the prosodic annotation mode on the *maiale* example fig. 5 (file *maiale.wav* and *maiale.wp2*).

trivial operation. In recent corpora elaboration projects, especially involving spontaneous speech, suboptimal acoustic quality of recordings lead to pitch curves carrying a large number of errors to the point they cannot be exploited.

This paper summarizes the most common problems found in prosodic annotation in these conditions, and describe the software program WinPitch, which integrates a set of functions designed to handle and correct erroneous fundamental frequency tracking, as well as a user programmable annotation system. Furthermore, the program allows for the annotation phonetic parameters to be transferred in one mouse click into a spreadsheet package such as Excel, in order to perform a comprehensive analysis of phonetic data obtained from the annotation.

WinPitch can be downloaded free from `www.winpitch.com`.

References

[1] M.E. Beckman, J. Hirschberg, S. Shattuck-Hufnagel, *The original ToBI system and the evolution of the ToBI framework*. In: S.A. Jun (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing*, 9-54. Oxford University Press, Oxford 2005.

[2] P. Boersma, *Accurate short time analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound*. Proceedings of

the Institute of Phonetic Sciences (University of Amsterdam), 17, 1993, 97-110.

[3] A. de Cheveigné, H. Kawahara, *YIN, a fundamental frequency estimator for speech and music*. Journal of the Acoustical Society of America, 111(4), 2002, 1917-1930.

[4] E. Delais-Roussarie, B. Post, M. Avanzi C. Buthke, A. Di Cristo, I. Feldhausen, S.A. Jun, P. Martin, T. Meisenburg, A. Rialland, R. Sichel-Bazin, H. Yoo, *Developing a ToBi system for French*. In: S. Frota, P. Prieto (eds.), *Intonational Variation in Romance*, chapter 3. Oxford University Press, Oxford 2015.

[5] Ph. Martin, *Extraction de la fréquence fondamentale par intercorrélation avec une fonction peigne*. Proc. 12e Journées d'Etude sur la Parole, Montréal, 1981, 221-232.

[6] Ph. Martin, *Cross-correlation of adjacent spectra enhances fundamental frequency estimation*. Proc. Interspeech 2008, Brisbane, 2008, 26-28.

[7] Ph. Martin, *Automatic detection of voice creak*. Proc. Speech Prosody 2012, Shanghai, 2012, 43-46.

[8] Ph. Martin, *The Structure of Spoken Language. Intonation in Romance*. Cambridge University Press, Cambridge 2015.

[9] M.A. Noll, *Cepstrum Pitch Determination*. Journal of the Acoustical Society of America, 41(2), 1967, 293-309.

[10] Praat, computer program, `www.praat.org` (accessed July 26, 2016).

[11] M. Rossi, *Le seuil de glissando ou seuil de perception des variations tonales pour la parole*. Phonetica, 23, 1971, 1-33.

[12] WinPitch, computer program, `www.winpitch.com` (accessed July 26, 2016).

Amedeo De Dominicis*

ON REPETITIONS

Abstract. – What should we annotate on a speech recording and particularly of a conversation? Traditionally, the phonetician tries to stay as close as possible to what he "hears" and sees on a spectrogram, *i.e.* close to the "real" phonetic details. But, in order to transcribe in IPA the repetitions of a given speech chunk by the same speaker and/or by a different speaker, it will be worthwhile to take into account the fact that a speaker often "hears" the previous turn differently from what a standard phonetic annotation would transcribe. The reasons may be various: for instance, because the speaker wants to "correct" the first pronunciation, or because (s)he is unable to understand it. Thus, the ideal phonetician should transcribe what the speaker shows he has perceived. So, what is the right way to annotate a repetition? Annotating from a "neutral" phonetic standpoint, or from the point of view of the speakers, of their different ways to "perceive" their speech and the speech of their interlocutors?

1. Introduction

Although two exemplars of a linguistic expression cannot ever be identical, members of a linguistic community treat some features as criteria and thereby some sequences as if they were the same. Traditionally, repetition is considered a homogeneous category; that is, all second mentions of words, regardless of the function of repetition within the talk, were considered to be 'the same'.

On the contrary, as Brown [3: 224] claims, "repeating something calls attention to the prior thing, brings it into the now, claims its relevance; repetition is therefore crucial in establishing discourse coherence. Of course, not all repetitions are alike: we can distinguish self-repetition from repetition of a prior turn at talk, and exact repetition from repetition with expansion (going beyond the initial version) or ellipsis (leaving something out in the repeat). Self-repetition facilitates language production, enabling rapid fluent talk, by

*University of Tuscia and Lincei (Italy).

setting up a syntactic frame and slotting new information into it ('He did A, and he did B, and he did C . . . '). Self-repetition also occurs predictably in self-repair, and may be used to make a bid to retain the floor or tie a referent to the prior discourse."

Clearly, repetitions can serve different functions: they can be employed as repairs; they can be used to demonstrate affiliation and display agreement [25]; they can show appreciation for the humor of the repeated utterance. By demonstrating a phonetic difference between self-repetitions produced in response to the other-initiation of repair, Curl [5] shows that not all repetitions can be considered 'the same' without careful analysis of the sequential structure they are produced in. Curl [5] argues that rather than considering phenomena as organized by reference to outside, analyst-imposed categories (*e.g.*, 'repetition'), we must consider the social actions that utterances perform (*e.g.*, repair), because the phonetics almost certainly matches the social actions rather than the lexical or syntactic items.

1.1. *Perturbations.*

In repetitions and overlapping talk, Schegloff [27: 12] found some hitches and perturbations: "the talk can get suddenly (i) louder in volume, (ii) higher in pitch, or (iii) faster or slower in pace, depending on where in the overlapping talk the change in pace occurs (more on this below). The talk-in-progress may be (iv) suddenly cut off, most commonly with what linguists call a glottal, labial, dental, or some other oral stop; or (v) some next sound may be markedly prolonged or stretched out; or (vi) a just prior element may be repeated".

1.2. *Transcriptions.*

Moreover, Munson *et al.* [23], and Kerswill and Wright [17] carry out a critique of phonetic transcription as a tool in sociolinguistic research. They show that listeners perceive consonants gradiently, contrary to the discrete nature of IPA transcription. As discussed in depth by Ladd [19], the IPA was designed in the late 19th century, long before the variation in phonetic detail presented in this paper had been studied, or even could have been studied.

Of course, we are not suggesting that categorical perception does not exist. Rather, we argue that categorical perception is a consequence of the task used to measure perception: whether a listener perceives a sound as categorical or not depends on the extent to which the task requires strict categorization. When the task promotes the perception of categories (either because of the difficulty of the task itself, or because of the use of categorical labels), people behave as if they can only hear categories and not the phonetic detail that these categories subsume. When different methods are used, individuals show exquisite sensitivity to the phonetic variation within categories (*e.g.* [28]). Generally, many listeners appear to have a rich enough knowledge of how sounds vary across social groups that they are able to parse out this variability when perceiving speech.

An example comes from Dell Hymes. Hymes [13, 14] sets out to retranscribe – or rather, reformat – the transcriptions of other scholars; he adds only those prosodic details that he considers important to his interpretation of the narrative, but in omitting other details he makes it difficult for readers to challenge or revise his account.

Yet Hymes himself often relies on details in others' transcripts to explore aspects of the data that the original researcher did not pursue.

It is for this reason – he says – that analysts are committed to transcribing in greater detail than is usually required by their analyses; the details that are not considered relevant should also be transcribed: the hope is that later readers and researchers will thereby discover new things in the data (*e.g.* [11]).

For instance, John W. Du Bois [7] presents his data, which are drawn from naturally occurring conversations and other speech events in the Santa Barbara Corpus of Spoken American English [8]. Du Bois [7: 54] specifies that in some cases the transcription has been slightly simplified for the sake of conciseness and accessibility: in particular this simplification pertains to cited utterances in which the speaker's words overlap with another utterance, not included in the citation, by a second speaker. What Du Bois describes is a very common practice among discourse analysts of all kinds in order to make their texts more reader-friendly by eliminating unnecessary clutter. Yet as researchers have repeatedly shown, what counts as an unnecessary clutter in discourse cannot be determined in advance. Once again variability in

transcription, whether deliberate or unintended, alters and often limits the ways in which published data may be read and interpreted.

1.3. *Transcription of talker-percept calibration.*

Moreover, there are some variations in transcription concerning the details of the speaker's pronunciation in a conversation. These details reveal that different auditors may hear different things in the same recording, not only because of differences in the quality of the equipment used and the acoustics of the setting, but also because of differences in what counts as salient, depending on the listeners' own dialectal backgrounds and phonological expectations as well as the practices of 'professional hearing' into which they have been socialized [1, 9, 10]. Hearing speakers in the same way the interlocutors do is the analytic challenge to take on. For instance, in the transcription of the famous conversation between Ava and Bee by Schegloff [26] the crucial point is the vocalization or the post-vocalic cancellation of [r] (often called *r-lessness*), which is characteristic of many USA dialects in informal speech (New York English, Boston English, Southern English and African American English) (cfr. [18]). In the following example [26: 66], the *parking place* by Ava is *r-less* token; but the repetition by Bee is an *r-ful* token.

Ava: [<I wan]'dih know if yih got a-uh:m
wutchimicawllit. A:: **pah(hh)khing** place °th's mornin'. ·hh
Bee: A **pa:rking** place,
Ava: Mm hm,
(0.4)

This means that Bee *hears* that *parking place* by Ava is not only *r-less*, but she considers this fact as very important, so that she highlights it by an *r-ful* token.

In fact, this kind of talker-percept calibration is very common, and many other examples can be found in the literature, including effects of presumed speaker gender on the perception of fricatives [21, 22, 29, 30], the influence of presumed speaker gender on the perception of vowels [15], and the influence of presumed dialect [24], presumed age [6] and presumed social class [12] on the perception of vowels.

Curl [5] highlights the interrelationship of phonetic structure and sequence organization. In particular, she presents a phonetic analysis of repetitions occurring in other-initiated repair sequences in American English. Despite their lexical similarities, the repairs are shown to have two distinct phonetic patterns. These patterns correspond systematically with a sequential and interactional difference between fitted and disjunct trouble source turns. Trouble source turns that are fitted to the prior turn are sequentially relevant actions, appropriate and based on the structure of the sequence-so-far, and in terms of turn-taking organization, do not intrude into the turn space occupied by another speaker (they are either overlapped by incoming talk, or produced in the clear). Trouble source turns that are disjunct at the place where they occur are either violative from a turn-taking standpoint – they overlap ongoing talk from the other participant – or, if they begin in the clear, they are ill-fitting or problematic actions in the sequence-so-far. An example of a repetition repair after a fitted trouble source turn is the self-repetition "she's coming alo:ng (heh)" following the turn "yeah she's comin' along". An example of a disjunct trouble source turn is the "d'you know her?" overlapping the turn "everybody's looking at me". Trouble source turns that are fitted are repeated more loudly and have expanded pitch ranges, longer durations, and changes in the articulatory settings. Trouble source turns that are disjunct are repeated more quietly and have nonexpanded pitch ranges, shorter durations, and no major differences in articulation.

One hypothesis about the value of such displays by speakers may be, as suggested by Couper-Kuhlen [4: 361], the acceptance or rejection of the social roles of offender and offended. The assignment of these roles is implicit in the production of the request for repair [27: 36 ff], because this activity calls attention to the problematicity of a particular turn in the sequence. The accused producer of the trouble source turn can display, by his/her choice of phonetic pattern, to accept or to reject the proposed role of 'troublemaker'.

2. Data

In our experimental work we analyzed two examples of repetition which came from the Conference Corpus: the file called "oro", and the file called

"vestiti". The methodology of the analysis will be explained below. The phonetic investigation of each repetition in our corpus will instantiate the main issues, and bring us to a more cogent discussion concerning how to transcribe a conversation, taking into account the perceptive approach, and particularly the standpoint of the interlocutors.

3. Methodology

We compared the formant chart of the vowels and the vocal quality of the linguistic expressions in repetition. In particular, we compared the degree of F0 (*jitter*) and amplitude (*shimmer*) perturbation.

3.1. *Formant Chart.*

A formant chart is a schematic plotting of the vowel formant frequencies. Vertical position on the diagram denotes the vowel aperture, with close (or high) vowels at the top of the diagram; and the horizontal position denotes the vowel backness, with front vowels on the left of the diagram. The vertical axe represents the values of the first formant F1 (in reverse order); the horizontal axe represents the values of the second formant F2 (also in reverse order).

The first two formants are important in determining the quality of vowels, and are frequently said to correspond to the open/close and front/back dimensions. Thus the first formant F1 has a higher frequency for an open vowel (such as [a]) and a lower frequency for a close vowel (such as [i] or [u]); and the second formant F2 has a higher frequency for a front vowel (such as [i]) and a lower frequency for a back vowel (such as [u]).

3.2. *Jitter.*

Vocal jitter is a case of F0 perturbation. The frequency of a speaker's voice varies from one cycle to the next. The random period variability is called vocal jitter. Vocal jitter is a measurement of vocal stability, and it increases in voice disorder and is responsible for hoarse, harsh or rough voice quality. Normal voices are usually less than 1% frequency variability. Mean percent jitter is measured taking the mean absolute jitter and dividing it by

the mean vocal period used during the phonation, the proportion is then multiplied by 100 to get a %age. As for local Jitter, MDVP (*Multi Dimensional Voice Programme*) [16] calls this parameter *Jitt*, and gives 1.040% as a threshold for pathology; as for rap Jitter, it is the Relative Average Perturbation, the average absolute difference between a period and the average of it and its two neighbors, divided by the average period: MDVP gives 0.680% as a threshold for pathology; as for ppq5 Jitter, it is the five-point Period Perturbation Quotient, the average absolute difference between a period and the average of it and its four closest neighbors, divided by the average period: MDVP calls this parameter *PPQ*, and gives 0.840% as a threshold for pathology.

3.3. *Shimmer.*

Vocal shimmer is a case of amplitude perturbation. It serves as an index of vocal stability: an excessive value of shimmer gives the perception of hoarseness. A mean cycle-to-cycle amplitude difference of 0.7 dB or less variation or less than 7% of mean amplitude is normal.

As for local Shimmer, it is the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude. MDVP (*Multi Dimensional Voice Programme*) calls this parameter *Shim*, and gives 3.810% as a threshold for pathology. As for Shimmer (local, dB), it is the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20. MDVP calls this parameter *ShdB*, and gives 0.350 dB as a threshold for pathology.

3.4. *Harmonics-to-Noise Ratio.*

We also measured the mean Harmonics-to-Noise Ratio (HNR) on the vowels in repetitions. HNR represents the degree of acoustic periodicity. HNR is expressed in dB: if 99% of the energy of the signal is in the periodic part, and 1% is noise, the HNR is 10*log10(99/1) = 20 dB. An HNR of 0 dB means that there is equal energy in the harmonics and in the noise. According to *Praat* guide, a healthy voice phonating /a/ or /i/ should have an HNR of 20, and 40 for the phonation of the vowel /u/. Consequently, an HNR below 20 is considered to be a measure of noticeable hoarseness.

Table 1 – Formant frequencies and voice qualities in the repetition “oro” (gold).

	oro1		oro2	
	o1	o2	O1	O2
Mean F1 (Hz)	541	501	508	520
Mean F2 (Hz)	1471	1194	1000	1225
Mean F3 (Hz)	2404	2100	2597	2443
Mean F4 (Hz)	3733	3572	4037	3801
Duration (ms)	165	72	231	257
Jitter (local)	0.808%	2.950%	0.988%	2.497%
Jitter (rap)	0.290%	1.893%	0.419%	1.348%
Jitter (ppq5)	0.270%	1.749%	0.580%	1.719%
Shimmer (local)	4.085%	6.054%	13.208%	23.858%
Shimmer (local, dB)	0.349 dB	0.647 dB	1.159 dB	2.072 dB
HNR (dB)	14.224	6.054	10.340	8.468

4. Analysis

We analyzed two cases of repetition. The corresponding audio files are called “oro.wav” and “vestiti.wav”.

4.1. *Oro.*

The interviewer speaks a semi-standard dialect (the regional variety of Viterbo), whereas the woman speaks the old dialect from Bomarzo (a little village in the North of Viterbo). This asymmetry is the main source of the problem.

The word *oro* ‘gold’ is initially introduced by the interviewer, then the woman repeats the same word. We analyzed the vowels of both words of both interlocutors. Conventionally *oro1* is the word uttered by the interviewer, and *oro2* is the repetition by the woman.

The vowels of *oro1* and *oro2* are labeled [o1ro2]1, and [O1rO2]2.

4.1.1. *oro1* [o1ro2]1.

As it can be seen in table 1, from [o1] to [o2] the values of jitter (local, rap and ppq5) increase over the thresholds for normal voice quality (thresholds = 1.040%; 0.680%; 0.840%); also the values of local and local-dB shimmer increase over the thresholds for normal voice quality (thresholds =3.810%; 0.350 dB).

4.1.2. *oro2* [O1rO2]2.

As it can be seen in table 1, from [o1] to [O1] all jitter and shimmer values increase, whereas the value of HNR decreases (and HNR below 20 is considered to be a measure of noticeable hoarseness). From [o2] to [O2] the local jitter is stable, whereas jitter (rap and ppq5) decreases (but still stays over the threshold for pathological voices), and local shimmer increases over the threshold for pathological voices.

Once again from [O1] to [O2] the values of jitter (local, rap and ppq5) increase over the thresholds for normal voice quality (thresholds = 1.040%; 0.680%; 0.840%); also the values of local-dB shimmer increase over the thresholds for normal voice quality (thresholds =3.810%; 0.350 dB); and the value of HNR decreases (and stays always below 20, which is considered to be a measure of noticeable hoarseness).

Thus the woman seems to have perceived that the interviewer's [o2] is harsher or more pressed, or hoarser than [o1]: as a consequence, in her repetition she realizes both her [O1] and [O2] as similar as possible to the interviewer's [o2], that is [O1] and [O2] are even harsher or more pressed, or hoarser than respectively [o1] and [o2].

As for the formant chart, fig. 1 and fig. 2 show that the woman's [O1] is further back and (slightly) higher than the interviewer's [o1]; the woman's [O2] is (slightly) lower and more front than the interviewer's [o2]: in other words, the woman overturns (in a sort of chiasm) the timbre (*i.e.* the formant space) of the vowels uttered by the interviewer, she manages to realize [O1] similar to [o2], and [O2] similar to [o1].

In short, because of the different dialect of the interlocutors, the woman's repetition shows that she "hears" the interviewer's pronunciation of the Ita-

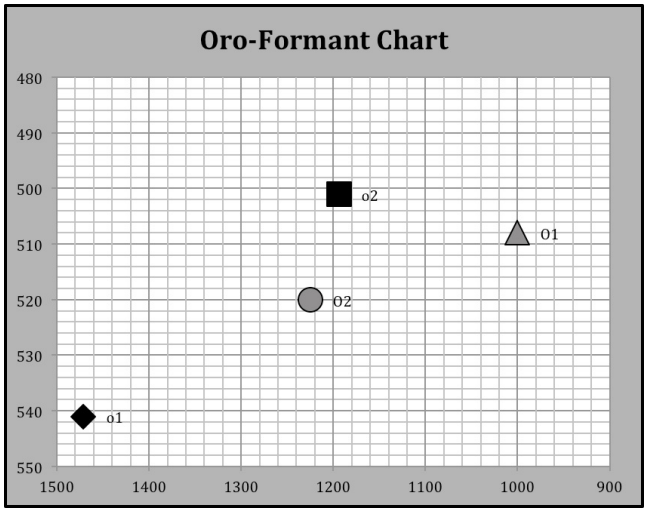

Fig. 1. – Vowels Formant Chart in the repetition “oro” (gold).

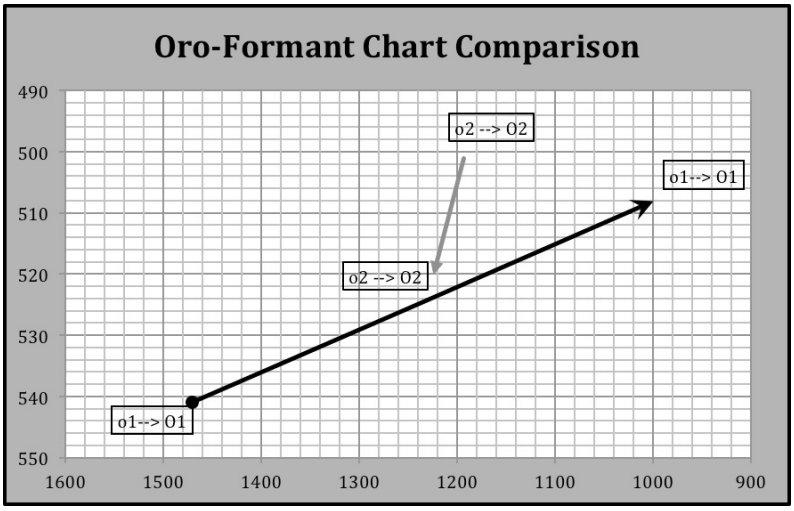

Fig. 2. – Comparison of the Vowels Formant Charts in the repetition “oro” (gold).

lian word *oro* differently from what a standard phonetic annotation would transcribe (in IPA): the woman perceives the interviewer's vowels as harsh or pressed, or hoarse, whereas phonologically they are not; she perceives his vowels as phonologically off-target, "wrong", that is with a reversed timbre, and in her repetition she "corrects" the interviewer's pronunciation.

Summing up, the woman perceives *oro1* as [ˈɔro], and realizes *oro2* as [ˈo̞ro̞] (where 'V̞' means an harsh vowel, 'V̞̞' means an harsher vowel, 'ɔ' stays for a [V̞] a sort of [o̞], and 'o' stays for [V̞] a sort of [o̞])[1]. But from a phonological standpoint both words would be transcribed /ˈɔro/, because none of the observed formant frequency differences between the speakers is phonological. So, what is the right way to annotate a repetition? Annotating from a "neutral" phonetic standpoint, or from the point of view of the speakers, of their different ways to "perceive" their speech and the speech of their interlocutors?

4.2. *Vestiti.*

In this audio recording we find again the voices of the same speakers as in the previous case (*oro*). Thus the interviewer speaks a semi-standard dialect (the regional variety of Viterbo), whereas the woman speaks the old dialect from Bomarzo. This asymmetry is the main source of the problem.
The word *vestito* 'dress' is initially introduced by the interviewer, then the woman repeats two times the same word but expressed in the plural form *vestiti* 'dresses'. We analyzed the vowels in the first occurrence *vestito* and in both repetitions of *vestiti*. Conventionally *vestito* is the word uttered by the interviewer, *vestiti1* and *vestiti2* are the repetitions by the woman.

[1]The IPA has no symbol for harsh voice. For articulations that do not have existing symbols, IPA may use an asterisk placed underneath. In the literature one can find different diacritics: the under-tilde used for a creaky voice; the double under-tilde used as the *ad hoc* diacritic for strident vowels; and an *ad hoc* underline. In the Voice Quality Symbols or VoQS [2], possibly relevant voice-quality symbols are {V!} ("harsh voice"), {V!!} ("ventricular phonation"), and {V̲} ("pressed phonation/tight voice").

In VoQS usage, "harsh voice" does not involve vibration of the ventricular folds, while in "pressed" or "tight" voice the arytenoid cartilages adduct so that only the anterior ligamental vocal folds vibrate.

Table 2 – Formant frequencies and voice qualities in the repetition “vestiti” (dresses). “Undef.” means that the value has not been detected.

	vestito			vestiti1			vestiti2		
	eØ	iØ	oØ	e1	i1	i2	E1	I1	I2
Mean F1 (Hz)	359	356	441	388	356	341	372	314	378
Mean F2 (Hz)	1989	1836	1366	2129	2466	2000	2264	2622	2503
Mean F3 (Hz)	2502	2618	2328	2540	2829	2639	2735	3078	2823
Mean F4 (Hz)	3991	4019	3788	3497	4086	3923	4015	4292	4071
Duration (ms)	38	43	35	44	117	306	61	97	204
Jitter (local, %)	1.426	0.830	2.386	1.407	1.301	4.213	1.237	1.862	2.310
Jitter (rap, %)	undef.	0.553	undef.	0.432	0.461	2.032	0.333	0.674	1.182
Jitter (ppq5, %)	undef.	undef.	undef.	0.290	0.777	2.405	0.655	0.995	1.427
Shimmer (local, %)	undef.	32.019	undef.	10.682	13.067	9.570	11.678	17.635	24.710
Shimmer (local, dB)	undef.	2.413	undef.	1.026	1.146	1.177	0.985	1.549	2.081
HNR (dB)	6.861	6.364	9.232	14.503	15.285	9.804	13.176	10.427	7.546

We analyzed the vowels of *vestito*, *vestiti1* and *vestiti2*. These vowels are labeled [v eØ sˈt iØ t oØ], [v e1 sˈt i1 t i2]1, and [v E1 sˈt I1 t I2]2.

4.2.1. *vestito* [v eØ sˈt iØ t oØ].

As it can be seen in table 2, from [eØ] to [iØ] the value of local jitter decreases; from [eØ] to [oØ] the value of local jitter increases. The values of shimmer are undefined for all vowels ([eØ] and [oØ], but [iØ]): thus they are not considered. The value of HNR increases from [eØ] and [iØ] to [oØ] (but stays always below 20, a value of HNR that is considered to be a measure of noticeable hoarseness).

4.2.2. *vestiti1* [v e1 sˈt i1 t i2]1.

As it can be seen in table 2, the values of jitter (local, rap and ppq5) increase from [e1] and [i1] to [i2]. The values of shimmer (local and dB) increase from [e1] and [i2] to [i1]. The value of HNR increases from [i2] to [e1] to [i1] (but stays below 20, a value that is considered to be a measure of noticeable hoarseness).

4.2.3. *vestiti2* [v E1 sˈt I1 t I2]2.

As it can be seen in table 2, in [I1] the values of jitter (local, rap and ppq5) increase compared to [i1]; and always in [I1] the values of shimmer (local and dB) increase compared to [i1], whereas the value of HNR in [I1] decreases compared to [i1] (but all HNR values stay below 20, a value that is considered to be a measure of noticeable hoarseness). Table 2 also shows that in [I2] the values of jitter (local, rap and ppq5) decrease compared to [i2]; and always in [I2] the values of shimmer (local and dB) increase compared to [i2], whereas the value of HNR in [I2] decreases compared to [i2] (and HNR below 20 is considered to be a measure of noticeable hoarseness).

As for the comparison among *vestito vestiti1* and *vestiti2*, table 2 illustrates that the value of jitter increases from *vestito* to *vestiti1* and then decreases from *vestiti1* to *vestiti2*. The values of shimmer decrease from *vestito* to *vestiti1* and *vestiti2*. The value of HNR increases from *vestito* to *vestiti1*, and then decreases from *vestiti1* to *vestiti2*.

Thus the woman seems to have perceived the interviewer's vowels as too harsh, or pressed, or hoarse, whereas phonologically they are not; and she tries to "amend" that in her repetitions: that is, she realizes her vowels (in *vestiti1* and in *vestiti2*) sometimes as hoarser, sometimes as less hoarse than the interviewer's vowels.

As for the formant chart, fig. 3 and fig. 4 show that the woman's [e1] is more front and lower than [eØ]; also her [i1] is more front than [iØ]. Comparing her second and first repetitions, the woman's [E1] is more front and higher than [e1]; [I1] is more front and higher than [i1]; [I2] is more front and lower than [i2].

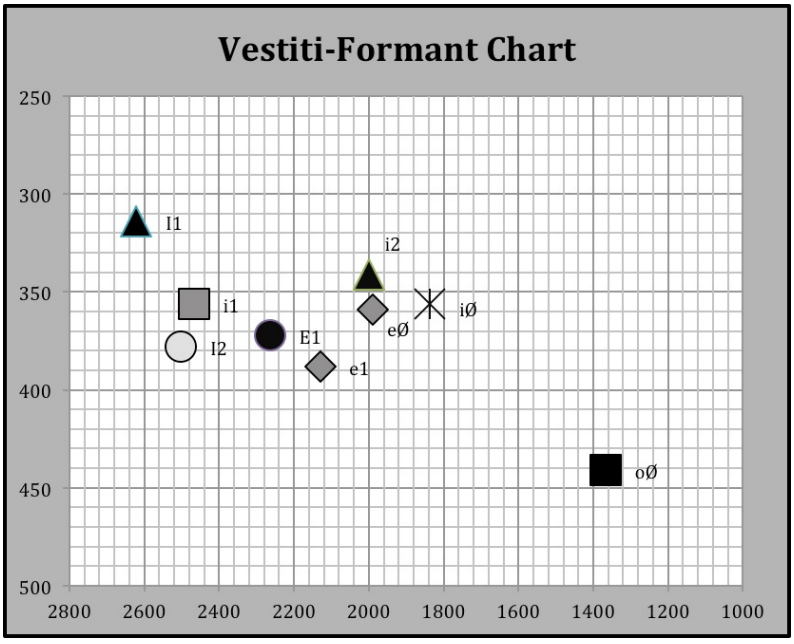

Fig. 3. – Vowels Formant Chart in the repetition “vestiti” (dresses).

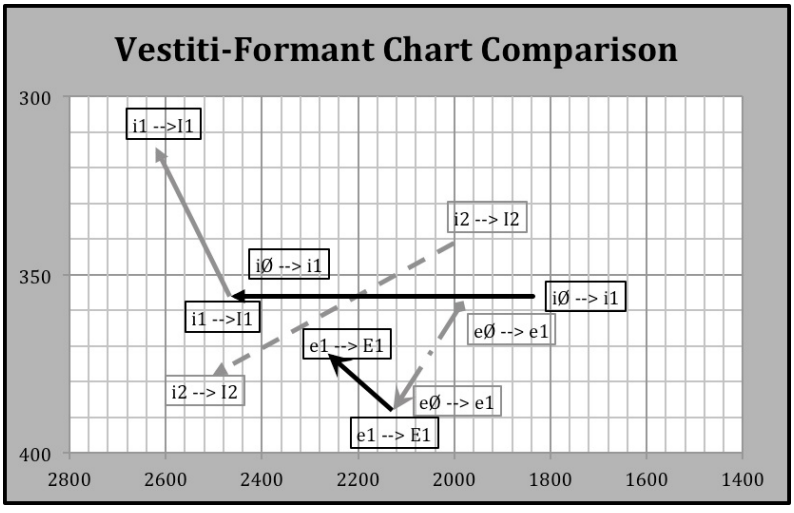

Fig. 4. – Comparison of the Vowel Formant Charts in the repetition “vestiti” (dresses).

Thus in her first repetition, the woman advances or fronts the interviewer's pronunciation. Then, in her second repetition, she increases her fronting pronunciation. So she shows that she has perceived the interviewer's vowels as too much back. As a consequence, the phonetic transcription of the interviewer's vowels should better represent the point of view of the woman, than – as traditionally – the point of view of the analyst. In short, the interviewer's vowels have to be annotated in IPA as backed or retracted.

In short, because of the different dialect of the interlocutors, the woman's repetitions show that she "hears" the interviewer's pronunciation of the Italian word *vestiti* (dresses) differently from what a standard phonetic annotation would transcribe (in IPA): the woman perceives the interviewer's vowels as harsh or pressed, or hoarse, whereas phonologically they are not; she also perceives his vowels as too back, *i.e.* phonologically off-target, "wrong"; and in her repetitions she tries to "amend", by fronting her vowels; in her repetitions she also tries to "amend" what she perceives as an excess of hoarseness by the interviewer, but she does not succeed, and then her vowels are sometimes hoarser, sometimes less hoarse than the interviewer's ones.

Summing up, the woman seems to perceive *vestito* as [vɔsˈt̪it̪o], and she realizes *vestiti1* as [vɛ̝̈sˈt̪it̪ï̝], and *vestiti2* as [vë̝sˈt̪it̪ï̝] (where 'V̈' means a harsh vowel, 'V̝̈' means a harsher vowel, 'ɛ' is for a [V̞] a sort of [e̞], and 'ɔ' is for [V̟] a sort of [e̟]). But from a phonological standpoint both repetitions would be transcribed /vesˈtiti/, because none of the observed formant frequency differences between the speakers is phonological. So, once again, what is the right way to annotate a repetition? Annotating from a "neutral" phonetic standpoint, or from the point of view of the speakers, of their different ways to "perceive" their speech and the speech of their interlocutors?

5. Discussion and conclusions

In 4.1 we have observed that the standard annotation of the interviewer's *oro* (gold) would be /ˈɔro/, but this would deplete the data, because we would miss an important fact: the woman perceives this word as [ˈɔro], and as a consequence she realizes her repetition as [ˈö̝rö̝].

In 4.2 we observed that the interviewer's *vestito* apparently should be transcribed in IPA as /vesˈtito/, but this transcription would miss an important fact: in the conversation the woman perceives this word as [vəsˈt̤i̤t̤o]. If we don't take into account this fact, then we will not understand why the woman's repetitions are [vɛ̱sˈti̱ti̱], and [ve̱sˈti̱ti̱].

So the study of repetitions helps us to reconsider the question of how to transcribe a conversation, taking into account the perceptive approach, and particularly the standpoint of the interlocutors. The traditional, let's say, "neutral" phonetic standpoint misses out on a number of facts and reduces the richness of the data; a more perceptive and interpersonal approach will cast a new light on the practice of annotating speech data.

This paper gives an overview of the linguistic annotation system and a range of pronunciation variations occurring in spoken conversation repetition data. Not only we do want to provide well-annotated spoken data for linguistic and engineering studies. We also hope to be able to more systematically and consistently identify spontaneous speech phenomena than has been done till now. Prosodic annotation has been neglected at this stage of research for reasons of time. In near future, this area will certainly form an independent project.

References

[1] M. Ashmore, K. MacMillan, S.D. Brown, *It's a Scream: Professional Hearing and Tape Fetishism*. Journal of Pragmatics, 36(2), 2004, 349-374.

[2] M.J. Ball, J. Esling, C. Dickson, *The VoQS System for the Transcription of Voice Quality*. Journal of the International Phonetic Association, 25(2), 1995, 71-80.

[3] P. Brown, *Repetition*. Journal of Linguistic Anthropology, 9(1/2), 1999, 223-226.

[4] E. Couper-Kuhlen, *Contextualizing discourse: The prosody of interactive repair*. In: P. Auer, A. di Luzio (eds.), *The Contextualization of Language*. John Benjamins, Amsterdam 1992, 337-364.

[5] T.S. Curl, *Practices in Other-Initiated Repair Resolution: The Phonetic Differentiation of "Repetitions"*. Discourse Processes, 39(1), 2005, 1-43.

[6] K. Drager, *Speaker age and vowel perception*. Language and Speech, 54(1), 2011, 99-121.

[7] J.W. Du Bois, *Discourse and Grammar*. In: M. Tomasello (ed.), *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure, Vol. 2*. Lawrence Erlbaum, Mahwah NJ 2003, 47-87.

[8] J.W. Du Bois, W.L. Chafe, C. Meyer, S.A. Thompson, *Santa Barbara Corpus of Spoken American English, Part 1*. Linguistic Data Consortium, Philadelphia PA 2000.

[9] M. Goldrick, *Linking Speech Errors and Generative Phonological Theory*. Language and Linguistics Compass, (5/6), 2011, 397-412.

[10] C. Goodwin, *Professional Vision*. American Anthropologist, 96(3), 1994, 606-633.

[11] P. ten Have, *Reflections on Transcription*. Cahiers de praxématique, 39, 2002, 21-43.

[12] J. Hay, P. Warren, K. Drager, *Factors influencing speech perception in the context of a merger-in-progress*. Journal of Phonetics, 34(4), 2006, 458-484.

[13] D. Hymes, *'In Vain I Tried to Tell You': Essays in Native American Ethnopoetics*. University of Pennsylvania Press, Philadelphia 1981.

[14] D. Hymes, *Ethnopoetics and Sociolinguistics: Three Stories by African-American Children*. In: D. Hymes (ed.), *Ethnography, Linguistics, Narrative Inequality: Toward an Understanding of Voice*. Taylor and Francis, Bristol PA 1996, 165-183.

[15] K. Johnson, E.A. Strand, M. D'Imperio, *Auditory-visual integration of talker gender in vowel perception*. Journal of Phonetics, 27, 1999, 359-384.

[16] Kay Elemetrics, *Multi-Dimensional Voice Program, Model 5105*. Kay Elemetrics Corporation, Lincoln Park NJ 2008.

[17] P. Kerswill, S. Wright, *The validity of phonetic transcription: Limitations of a sociolinguistic research tool*. Language Variation and Change, 2, 1990, 255-275.

[18] W. Labov, *The Social Stratification of (r) in New York City Department Stores*. In: W. Labov, *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia 1972, 43-69.

[19] D.R. Ladd, *Phonetics in phonology*. In: J. Goldsmith, J. Riggle, C. Alan, L. Yu (eds.), *The Handbook of Phonological Theory: Second Edition*. Blackwell Publishing, Oxford 2011, 348-373.

[20] C. Langstrof, *Acoustic evidence for a push-chain shift in the Intermediate Period of New Zealand English*. Language Variation and Change, 18, 2006, 141-164.

[21] B. Munson, *Gender biases in fricative identification revisited. Oral presentation at the annual meeting of the Linguistic Society of America*. San Francisco 2009; http://www.tc.umn.edu/~munso005/LSA2009_Munson.pdf.

[22] B. Munson, *On Voiceless Fricative Perception: Vocal-Tract Normalization, and Socioindexicality. Oral presentation at the International Phonetics and Phonology Forum*. Kobe University, Japan 2009; http://www.tc.umn.edu/~munso005/Munson_Fricatives_August25_Final.pdf.

[23] B. Munson, J. Edwards, S. Schellinger, M.E. Beckman, M.K. Meyer, *Deconstructing Phonetic Transcription: Language-Specificity, Covert Contrast, Perceptual Bias, and an Extraterrestrial View of Vox Humana*. Clinical Linguistics & Phonetics, 24(4/5), 2010, 245-260.

[24] N. Niedzielski, *The Effect of Social Information on the Perception of Sociolinguistic Variables*. Journal of Language and Social Psychology 18, 1999, 62-85.

[25] E.A. Schegloff, *Confirming Allusions: Toward an Empirical Account of Action*. American Journal of Sociology, 102(1), 1996, 161-216.

[26] E.A. Schegloff, *Turn Organization: One Intersection of Grammar and Interaction*. In: E. Ochs, E.A. Schegloff, S.A. Thompson (eds.), *Interaction and Grammar*. Cambridge University Press, Cambridge, 1996, 52-133.

[27] E.A. Schegloff, *Overlapping Talk and the Organization of Turn-Taking for Conversation*. Language in Society, 29(1), 2000, 1-63.

[28] J. Scobbie, *Flexibility in the face of incompatible English VOT systems*. In: L.M. Goldstein, C. Best, D. Whalen (eds.), *Papers in Laboratory Phonology 8*. Cambridge University Press, Cambridge 2004.

[29] E. Strand, K. Johnson, *Gradient and visual speaker normalization in the perception of fricatives*. In: D. Gibbon (ed.), *Natural language processing and speech technology: Results of the 3rd KONVENS Conference*. Beilefeld, 1996. Mouton de Gruyter, Berlin 1996, 318-336.

[30] J. Stuart-Smith, *Empirical evidence for gendered speech production: /s/ in Glaswegian*. In: J. Cole, J.I. Hualde (eds.), *Laboratory Phonology 9*. Mouton de Gruyter, New York 2007, 65-86.

Brigitte Bigi*

ANNOTATION REPRESENTATION AND FILE CONVERSION TOOL

Abstract. – Annotating corpora is of crucial importance in Corpus Linguistics. Linguistics annotation, especially when dealing with multiple domains, makes use of different tools within a given project. More and more annotated corpora are now available, and so are tools to annotate automatically and/or manually. Due to the diversity of linguistic phenomena, annotation tools lead to a variety of models, theories and formalisms. This diversity results in heterogeneous description formats, each tool developing its own framework. Then, none of the annotation tools are directly interoperable, each one using a native format, some of them on top of XML, some others developing an ad hoc markup language. The mapping between user formats is then an important issue. This paper presents an efficient annotation representation framework and the related tool to convert from/to some of the existing annotation file formats of various annotation software tools for audio recordings.

1. Introduction

Corpus Linguistics is a Computational Linguistics sub-field which aims to study the language as expressed in corpora. Nowadays, *Annotation, Abstraction and Analysis* – the 3A from [22], is a common perspective in this field. *Annotation* consists of the application of a scheme to recordings (text, audio, video...). *Abstraction* consists of the mapping of terms in the scheme to terms in a theoretically motivated model or dataset. *Analysis* consists of statistically probing, manipulating and generalizing from the dataset. Within these definitions, this paper is part of *Annotation* field, focusing on linguistic annotations of audio recordings.

Annotating speech recordings is relevant for many sub-fields of linguistics such as phonetics, prosody or discourse studies. Corpora are annotated thanks to the use of specialized software. Generally, "different annotation tools are designed and used to annotate the audio and video contents of a

*Laboratoire Parole et Langage - CNRS, Aix Marseille Univ. - 5 avenue Pasteur - 13100 Aix-en-Provence (France).

corpus that can later be merged in query systems or databases" [1]. To be useful for purposes such as qualitative or quantitative analyses, the annotations of recordings must be time-synchronized (*i.e.* time-aligned). Temporal information makes it possible to describe behavior or actions of different subjects that happen at the same time. Indeed, "when multiple annotations are integrated into a single data set, inter-relationships between the annotations can be explored both qualitatively (by using database queries that combine levels) and quantitatively (by running statistical analyses or machine learning algorithms)" [10].

For a researcher looking for an annotation software tool, it might be difficult to select the most appropriate one. See [18] for a comparison of multimodal annotation tools. To decide about usefulness and usability of a software, it is advisable to consider the software license, the portability, the ease of use, the strengths/weaknesses for specific annotation purposes, the type of data or analysis the tool/software is specifically designed to complete and the software compatibility with other annotated data, *i.e.* the availability of files to be imported from and exported to several other data formats.

The choice of a software is then of crucial importance: it determines the framework that will be utilized during the annotation process and the suitability of the annotated files. Due to the diversity of linguistic phenomena, annotation tools lead to a variety of models, theories and formalisms. This diversity results in heterogeneous description formats, each tool developing its own framework. As a consequence, none of the annotation tools are directly interoperable, each one using a native format: some of them on top of XML, some others developing an *ad hoc* markup language.

SPPAS – the automatic annotation and analysis of speech, is a recent computer software tool designed and developed by the author to handle multiple language corpora and/or tasks in the same software environment [7]. SPPAS allows to automatically create and to analyze annotations; it offers a *fully-automatic or semi-automatic annotation process*, including for speech segmentation [3, 4]. SPPAS includes a generic enough annotation representation framework that allows it to import from and export to a variety of file formats, which is the purpose of this paper. It also proposes its own native XML-based format.

After a background on data representation, this paper briefly presents the annotation representation framework of SPPAS and the related tool to convert from/to some of the existing file formats of various software. Such tool is illustrated via several conversions of one file of the set of data distributed for the International Workshop "Speech audio archives: preservation, restoration, annotation, aimed at supporting the linguistic analysis", organized by Accademia Nazionale dei Lincei, in Roma Italy in May, 2017.

2. Background on data representation

2.1. *Standardization of annotation representation frameworks.*

Nowadays, most of linguistic annotation projects require a rich framework in order to deal with all the information levels involved. For example, dialog act annotation involves multi-dimensional communicative functions that are organized in a complex taxonomy. Manipulating linguistic annotated resources requires to be independent from the coding format which implies two kinds of pre-requisites:

1. to specify and organize the information to be encoded independently from the constraints or restriction of the format (or the annotation tool);

2. to encode the information into an exchange format, readable whatever the edition or annotation system.

One way toward interoperability is the standardization of annotation frameworks. "The ISO/TEI specifications implement the full power of feature structures and define inheritance, unification, and subsumption mechanisms over the structures, thus enabling the representation of linguistic information at any level of complexity." [12]. Recently, Liégeois *et al.* [15] proposed to extend the TEI and to use it as a pivot format for oral and multimodal language corpora. However, this initiative is in its early stage and despite standardization efforts it is unlikely that a unique generic schema shared by all annotation tools emerge.

2.2. *Annotation tool frameworks.*

Trees or graphs are a very common formalism used for various linguistic representation levels in annotation tools. Syntactic trees are the clearest example of such linguistic representations. Unlike other linguistic annotations data structures of speech/video annotated data are “tiers”, “layers” or “tracks” of annotations. In that context, a *Tier* consists of series of *Annotation* instances, each one defined by a temporal localization (mostly an interval) and a label. It results in the following advantages:

- inclusion of various annotation types in the same document;
- flexible merging of documents;
- easy comparison of annotations (mainly on the basis of temporal information);
- representation of alternative/concurrent annotations, by means of separated tiers;
- representation of partial annotations (tiers with “holes”);
- such annotation representation framework is not theory-dependent.

Fortunately, as shown in the left-vs-right illustrations of fig. 1, some mapping is possible between representations. In this example, we limited the annotations to: syntactic groups (g), syntax categories (c), words (w), lemmas (l), phonemes (p), syllables (s), left hand movements (lh), right hand movements (lr). Tiers can be mapped into graphs and vice-versa if graphs have "categorized-nodes" like it is represented by various intensities of gray in fig. 1.

As soon as the annotations are associated to a time representation, trees can easily and systematically be mapped into tiers. However, it is not always possible to map a hierarchy of tiers into a tree – it depends on the structure of the hierarchy, as shown below:

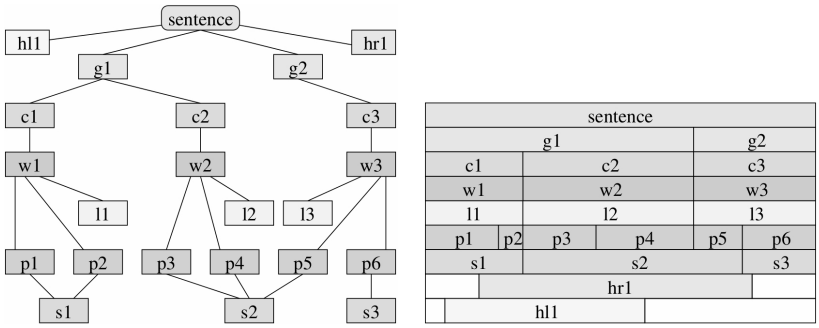

Fig. 1. – Example of annotations representation using graphs at left, and tiers at right.

```
Parent tier:  Phonemes | l | a | n | e | l | a |
Child tier:  Tokens | l' | âne | est | là |
Parent tier:  Phonemes | l | a | n | e | l | a |
Child tier:  Syllables | l.a | n.e | l.a |
```

Notice that there is no possibility of hierarchy between Tokens and Syllables. Consequently, the map from a representation to another one is something difficult but doable in many situations.

3. Annotation representation framework in SPPAS

Representing annotated data in SPPAS is of crucial importance for its automatic annotations, its analysis of annotations and for the software to be able to automatically annotate and analyze any kind of data files. Data representation of SPPAS is then commonly based on tiers, made of annotations. However, contrariwise to most of the existing annotation software tools, the time localization and label of each annotation are represented as complex objects to increase the availability of both: representing complex annotations and being compatible with other existing annotation representations. SPPAS also allows to represent uncertainty of annotations: it can concern the time precision [5], the localization and/or the label (*i.e.* it can include alternative localizations and/or alternative labels). An *Annotation* in SPPAS is then made of:

- a *Location*, which is a list of *Localization* and its *score*. A *Localization* is either a *Point,* an *Interval* or a *Disjoint*. A *Point* includes a midpoint value and a radius to represent the vagueness of the localization [5]. *Disjoint* intervals are mostly used for dialog annotations;

- eventually a *Label* which is a list of *Tag* and its *score*. A *Tag* is either a string or a numerical value.

A *Tier* is then used to distribute the annotations of a recording. It can also contain a controlled vocabulary, a link to a primary media, etc. The definition of a tier proposed in [9] is then largely recovered: "a tier is a group of annotations that all describe the same type of phenomenon, that all share the same metadata attribute values and that are all subject to the same constraints on annotation structures, on annotation content and on time alignment characteristics". Tiers are stored into a *Transcription* object that may indicate a *Hierarchy* between tiers in the form of a list of *TimeAssociation* or *TimeAlignment* links. All the 3 objects *Transcription*, *Tier*, and *Annotation* can also contain metadata in the form *key:value* like for example: date:2016-12-01; annotator: John; etc.

A native format named XRA was developed to fit in such data representation. The physical level of representation of XRA obviously makes use of XML, XML-related standards and stand-off annotation. Due to an intuitive naming convention, XRA documents are human readable as far as possible within the limits of XML. The format specification as well as the texts encoded in XRA are "reusable, *i.e.*, potentially usable in more than one research project and by more than one research team, and extensible, *i.e.*, capable of further enhancement" [11].

The proposed framework is implemented using the programming language Python. This Application Programming Interface is part of SPPAS and also distributed under the terms of the GNU Public License. The developers implement behaviors in abstract classes and each class is well-documented. Actually, it is quite easy to read some existing annotation file formats and to instantiate them into the proposed framework to use it (like exemplified in section 4.3).

4. Converting files with SPPAS

4.1. *Principle.*

SPPAS makes use of its internal data representation to convert files. A conversion then consists of two steps: First, the incoming file is loaded and mapped to the SPPAS data framework described in the previous section and second, such data will be saved to the expected format. These two steps are applied whatever the organization structure of annotations in the original or in the destination file format. This is illustrated by fig. 2 which includes the list of file formats and the corresponding software that are supported by version > 1.8 of SPPAS. Arrows are representing the following actions:

- *import from*, represented by a continued line with an arrow from the file format to the SPPAS framework;
- *partially import from*, represented by a dash line with an arrow from the file format to the SPPAS framework;
- *export to*, represented by an arrow from the SPPAS framework to the file format.

SPPAS version 1.8.0 supports Phonedit & Signaix [21], Praat [8], HTK [24], Sclite [17], Xtrans [16], Transcriber [2], Anvil [13], Elan [23], Annotation Pro [14], subtitles and tab-separated files. The support of external formats is regularly extended to new formats by the author on-demand from the users and distributed to the community in the SPPAS regular updates (about ten a year).

4.2. *The given example.*

In the context of the International Workshop "Speech audio archives: preservation, restoration, annotation, aimed at supporting the linguistic analysis", a set of audio and TextGrid files were distributed. The file `Ayoreyo.TextGrid` will be used as example in the following.

TextGrid is the native format of *Praat,* a tool for manually annotating sound files. It provides different visualizations of audio data - waveform or

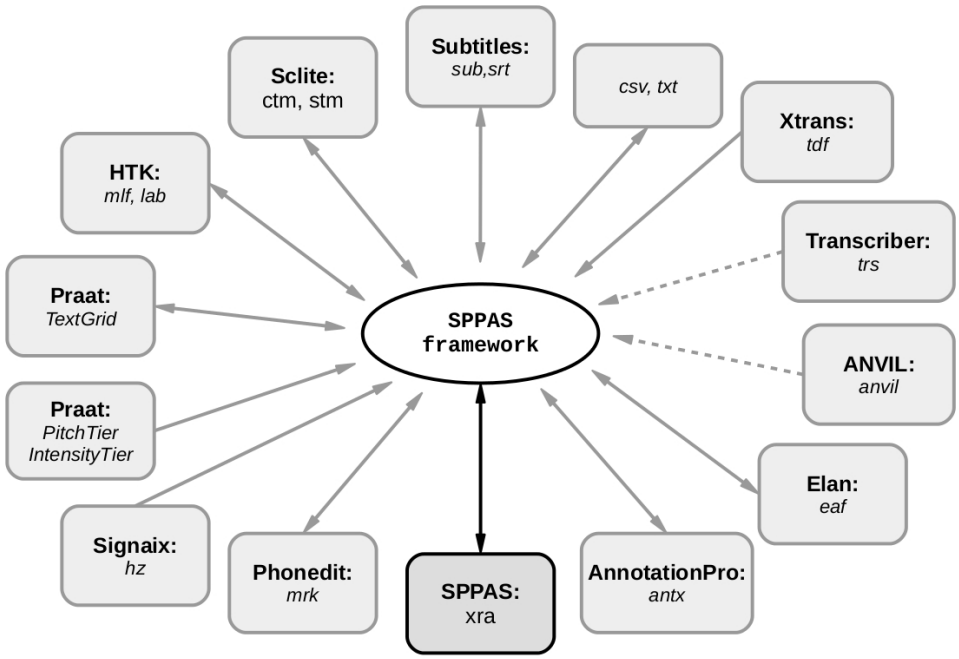

Fig. 2. – SPPAS conversion method and formats (version 1.8.0).

spectrogram display - and, among other things, enables pitch contour as well as formant calculation and visualization. The Praat native files are in several native text formats; “TextGrid” is the one used to save annotated data.

To be used with SPPAS, it was first necessary to save the given file with UTF-8 encoding. There are 9 tiers with name TranscripAYO, TranslationENG, Words, WordsPhones, Phones, Syllables, Morph-Syntax, language and Notes. We observe that tiers are used for both annotated data and metadata. Annotated data of tiers are represented as below:

```
intervals [1]:
   xmin = 0
   xmax = 0.1377517184116469
   text = "#"
intervals [2]:
```

```
        xmin = 0.1377517184116469
        xmax = 0.4928648872410315
        text = "chi"
    intervals [3]:
        xmin = 0.4928648872410315
        xmax = 1.5253691515655043
        text = "#"
    intervals [4]:
        xmin = 1.5253691515655043
        xmax = 3.747445456657043
        text = "siquere chise yocayode ore nani"
    intervals [5]:
        xmin = 3.747445456657043
        xmax = 4.186875287771049
        text = "#"
    intervals [6]:
        xmin = 4.186875287771049
        xmax = 8.13501133786848
        text = ""
```

Each annotation is represented by an interval, with 2 real numbers to represent respectively the beginning and the end of the interval, and a text. We can observe that the number of digits can increase up to 15. In our example, the audio file is sampled at 44,100 Hz, a frame is then during 0.000022676 seconds. This means that time value of an annotation in this TextGrid is about 100,000 times more precise that the framerate of the annotated recording.

4.3. *Converting using SPPAS*

.

There are several ways to use SPPAS: The Graphical User Interface (GUI), the Command-Line User Interface (CLI) and the Application Programming Interface (API).

There are two solutions to convert files from the GUI (fig. 3).). The

first one is directly from the File Explorer by selecting the file(s) to convert, clicking on the "Export" button and then choosing the file format from the given list. The second solution is the DataRoamer analysis component that is useful if the file as to be modified: renaming tiers, deleting tiers, changing their order, etc. In both cases, an export fails if:

- the given file is not in UTF-8 encoding;
- the given file contains several tiers and the output format supports only one;
- the given file contains corrupted annotations (as for example: two annotations of the same tier are starting and ending at the same time, the end of an interval is lower or equal to the beginning, etc).

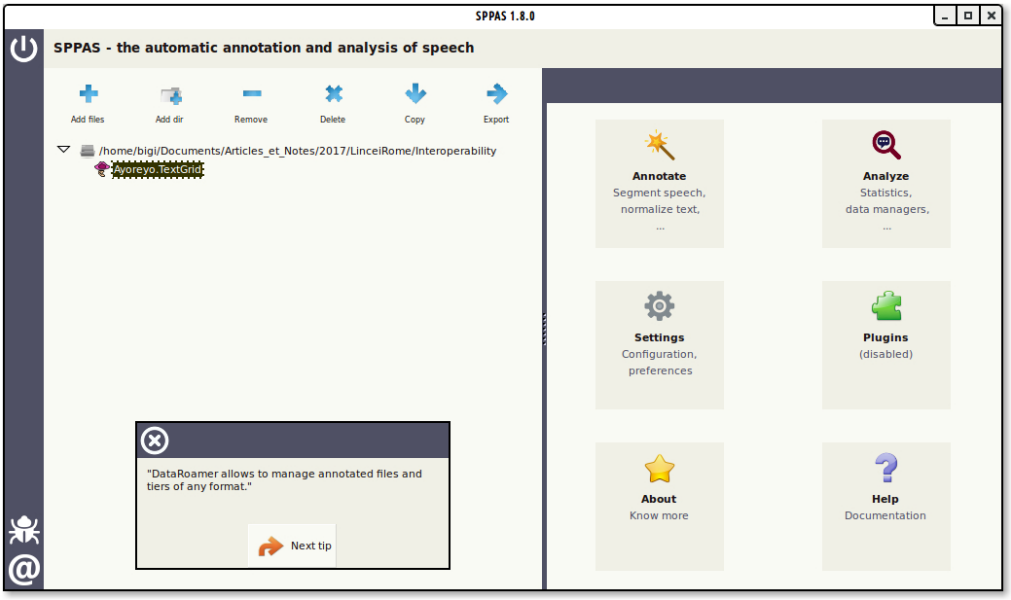

Fig. 3. – SPPAS main frame.

Converting with the CLI is performed thanks to a Python script: trsconvert.py. It allows to select the tier(s) to export. Finally, the more powerful

solution to convert files is the API but it requires some basic skills about Python programming language. Fig. 4) is an example of the script that was specifically created for the purpose of this study. The first step is to load the file and to map it into the SPPAS framework. Then, the following optional modifications are performed: fixing hierarchy links between tiers, assigning a recording, replacing the last two tiers by metadata and fixing the language as metadata. The modified data representation is saved in `Ayoreyo.xra` file.

```
import sys
import os
SPPAS = os.getenv('SPPAS')
sys.path.append(os.path.join(SPPAS, 'sppas', 'src'))
from annotationdata.transcription import Transcription
from annotationdata.media import Media
import annotationdata.io

# Read input file
trs = annotationdata.io.read( "Ayoreyo.TextGrid" )

# Add "flexibility" in boundary time values
for tier in trs:
    tier.SetRadius(0.005)

# Optionnal:  Set Hierarchy links
trs.GetHierarchy().addLink("TimeAlignment",trs.Find("Words"),trs.Find("TranscripAYO"))
trs.GetHierarchy().addLink("TimeAssociation",trs.Find("Words"),trs.Find("WordsPhones"))
trs.GetHierarchy().addLink("TimeAssociation",trs.Find("Words"),trs.Find("Morph-Syntax"))
trs.GetHierarchy().addLink("TimeAlignment",trs.Find("Phones"),trs.Find("Words"))
trs.GetHierarchy().addLink("TimeAlignment",trs.Find("Phones"),trs.Find("Syllables"))

# Optionnal:  Set Media
m = Media('AyoreyoId', 'Ayoreyo.wav', 'audio/wav')
trs.AddMedia(m)
for tier in trs:
    tier.SetMedia(m)

# Optionnal:  Set Metadata
trs.SetMetadata('iso639-3', u"ayo")
trs.SetMetadata('notes', u"# = pause; ÂĝÂĝÂĝ = unintelligible")
trs.Find("TranslationENG").SetMetadata("iso639-3", "eng")

# Optionnal:  Remove/Modify the tiers that was used to declare metadata
trs.Pop()
trs.Pop()
trs.Find("TranscripAYO").SetName("Transcription")

# No set of a controlled vocabulary

# Write

annotationdata.io.write("Ayoreyo.xra", trs)
```

Fig. 4. – A Python script to convert the example file, with SPPAS version 1.8.0.

An annotation of the resulting XRA file looks like the following one:

```
<Annotation>
  <Location>
```

```
        <Localization score="1.0">
         <Timeinterval>
          <Begin midpoint="1.5253691516" radius="0.005"/>
          <End midpoint="3.7474454567" radius="0.005"/>
         </Timeinterval>
        </Localization>
       </Location>
       <Label scoremode="max">
        <Text score="1.0">siquere chise yocayode ore nani
</Text>
       </Label>
      </Annotation>
```

As it was previously mentioned, the framework for annotations is very rich and contain several information like alternative labels or alternative localizations. The midpoint value was stored exactly as it is in the original file, which is useful to save back in a TextGrid file without any loss of information.

Either the original TextGrid file or the newly created XRA file can be converted in any of the supported formats; next section reports on some of them.

4.4. *Example of results.*

Two exports into XML formats are supported: Elan and Annotation Pro. Elan export (.eaf) of the above mentioned annotation is:

```
     <ANNOTATION>
      <ALIGNABLE_ANNOTATION ANNOTATION_ID="a3"
TIME_SLOT_REF1="t3" TIME_SLOT_REF2="t4">
       <ANNOTATION_VALUE>siquere chise yocayode ore nani
</ANNOTATION_VALUE>
      </ALIGNABLE_ANNOTATION>
     </ANNOTATION>
```

Time values, in milliseconds, are stored in a list of time slots and annota-

tions refer to such time slots. Only intervals are supported and overlaps are forbidden. Elan allows to set a controlled vocabulary, a hierarchy between tiers and a media. The other XML export supported by SPPAS is the AnnotationPro native format (.antx). Each annotation is considered as a *Segment* of a previously defined *Layer*. Like for most of annotation tools, only intervals are supported and overlaps are forbidden. Each segment also contains graphical information used by the software to display it, like its colors. Finally, a segment can be enriched by 1 to 3 parameters.

Table 1 presents the other formats supported by SPPAS. CSV is not related to an annotation tool however it is supported by spreadsheet software. Phonedit (.mrk) is using the informal standard of INI configuration files: a simple text file with a basic structure composed of sections, properties and values. In that case, sections are used to represent tiers; then properties are the annotations identifiers for which a value is assigned: the label of the annotation. Only intervals are supported, however the degenerated interval is allowed and overlaps are also allowed. The time marked conversation scoring input (.ctm) is a native format of Sclite – a tool for scoring and evaluating the output of speech recognition systems. This file format is a concatenation of time mark records for each word in each channel of a waveform. The records are separated with a newline. Each word token must have a waveform id, channel identifier [A | B], start time, duration, and word text. Optionally a confidence score can be appended for each word. The segment time mark input file (.stm) is also a Sclite native format. The file consists of a concatenation of text segment records from a waveform file. Each record is separated by a newline and contains: the waveform's filename and channel identifier [A | B], the talker ids, begin and end times (in seconds), optional subset label and the text for the segment. SubRip files (.srt) contain formatted lines of plain text in groups separated by a blank line. The plain text is made of a numeric counter identifying each sequential subtitle, the time that the subtitle should appear on the screen, followed by --> and the time it should disappear and the subtitle text itself on one or more lines. Finally, the SubViewer files is a INI style file with a header section that contains metadata and rendering instructions. Immediately following is a [SUBTITLE] section, consisting of

Table 1 – Example of one annotation as it is represented in several file formats.

`Transcription,1.5253691516,3.7474454567,siquere chise yocayode ore nani`	.csv
`LBL_LEVEL_AA_000003= "siquere chise yocayode ore nani" 1525.369152 3747.445457`	.mrk
`Ayoreyo.wav A 1.5253691516 2.2220763051 siquere chise yocayode ore nani 1.0`	.ctm
`Ayoreyo.wav Transcription none 1.5253691516 3.7474454567 siquere chise yocayode ore nani`	.stm

`4`
`00:00:01,525 --> 00:00:03,747`
`siquere chise yocayode ore nani`
	.srt

`00:00:01.52,00:00:03.74` `siquere chise yocayode ore nani`	.sub

comma-delimited time ranges (accurate to one hundredth of a second) and a caption to be displayed during each range.

5. Conclusion

Computational Linguistics make it necessary to produce and share reference annotations to which linguistic and computational models can be compared. The development of an expressive and generic framework for linguistic annotations is then very beneficial and many efforts are currently being made in this direction. Considerable attention has been devoted to the development of means to represent annotations so that phenomena at different levels can be merged and/or analyzed in combination. A particular attention has been on the development of standards for representing linguistic annotations.

This paper particularly focused on the problem of speech annotation representation framework and conversion. An XML file format was proposed, and a conversion tool was described and illustrated with an example. With such simple example, we observed that SPPAS framework is interoperable with a large number of file formats: like Elan, it supports hierarchy of annotations, controlled vocabularies and media assignment; like Annotation Pro,

it supports an advanced media definition; like Phonedit, it allows overlaps of annotations; like Sclite, it supports alternative labels; like Praat, it allows to store all the 15 digits for time values, and it supports interval tiers and point tiers. Moreover, several other features already implemented in the framework will allow to support more formats in the future. Actually, the development of new file import/export is done by the author on-demand.

Some other components are also published within SPPAS software tool for the analysis of any kind of annotated files supported for importing by SPPAS. The DataStats analysis component is used to count occurrences of annotations and to estimate descriptive statistics as standard deviations, etc. It is also intended to be used to estimate the Kappa value in a near future. The DataFilter analysis component is used to extract annotations from tiers [6]. With respect to searching, linguists are typically interested in locating patterns on specific tiers, with the possibility to relate different patterns of annotations on some tiers. DataFilter offers a powerful way to request/extract data, with the help of Allen's relations, without requiring any XML-related knowledge or skill. It has been already used in several studies as to find correlations between speech and gestures [19], or to find which gestures are produced while pausing [20], just to cite some of them. Finally, the use of the SPPAS framework in several studies or projects, and the user feedbacks allowed us to verify that the proposal of this paper is robust, useful and reliable.

SPPAS can be freely downloaded at: `http://www.sppas.org`.

References

[1] Á. Abuczki, E. Baiat Ghazaleh, *An overview of multimodal corpora, annotation tools and schemes*. Argumentum Hungria, 1(9), 2013, 86-98.

[2] C. Barras, E. Geoffrois, Z. Wu, M. Liberman, *Transcriber: development and use of a tool for assisting speech corpora production*. Speech Communication, 33(1), 2001, 5-22.

[3] B. Bigi, *A Multilingual Text Normalization Approach*. Less-Resourced Languages workshop, 5th Language & Technology Conference, Poznań Poland, 2011.

[4] B. Bigi, *The SPPAS participation to Evalita 2011*. In: B. Magnini *et al.* (eds.), *LNAI 7689*. Springer, Berlin Heidelberg, 2012, 312-321.

[5] B. Bigi, T. Watanabe, L. Prévot, *Representing Multimodal Linguistic Annotated Data*. Proceedings of the Ninth International Conference on Language Resources and Evaluation, European Language Resources Association, Reykjavik Iceland, 2014, 3386-3392.

[6] B. Bigi, J. Saubesty, *Searching and retrieving multi-levels annotated data*. Gesture and Speech in Interaction - 4th edition, Nantes France, 2015, 31-36.

[7] B. Bigi, *SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech*. The Phonetician, International Society of Phonetic Sciences, 111-112, 2015, 54-59.

[8] P. Boersma, D. Weenink, *Praat, a system for doing phonetics by computer*. Glot International, 5(9/10), 2001, 341-345.

[9] H. Brugman, *Annotated Recordings and Texts in the DOBES project*. Proceedings of the EMELD Language Digitization Project Conference: Workshop on Digitizing and Annotating Texts and Field Recordings. LSA Institute, University of Michigan, Lansing MI USA, 2003.

[10] C. Chiarcos, S. Dipper, M. Götze, U. Leser, A. Lüdeling, J. Ritz, M. Stede, *A flexible framework for integrating annotations from different tools and tagsets*. Traitement Automatique des Langues, 49(2), 2008, 271-293.

[11] N. Ide, C. Brew, *Requirements, tools, and architectures for annotated corpora*. Proceedings of data architectures and software support for large corpora, Citeseer, 2000, 1-5.

[12] N. Ide, *Annotation science: From theory to practice and use: Data structures for linguistics resources and applications*. Proceedings of the Bienniel GLDV Conference, Tübingen Germany, 2007.

[13] M. Kipp, *Anvil, DFKI, German Research Center for Artificial Intelligence*. http://www.anvil-software.de/ (accessed January 26, 2018).

[14] K. Klessa, *Annotation Pro. Enhancing analyses of linguistic and paralinguistic features in speech*. Wydział Neofilologii UAM, Poznań Poland, 2016.

[15] L. Liégeois, C. Etienne, C. Parisse, C. Benzitoun, C. Chanard, *Using the TEI as a pivot format for oral and multimodal language corpora.* Text Encoding Initiative Conference and Member's meeting, Lyon France, 2015.

[16] LDC - Linguistic Data Consortium, *Xtrans: A next generation transcription tool developed by LDC*. Trustees of the University of Pennsylvania, https://www.ldc.upenn.edu/language-resources/tools/xtrans (accessed January 26, 2018).

[17] NIST – National Institute of Science and Technologies. Speech Recognition Scoring Toolkit, version 2.4.0. http://www.itl.nist.gov/iad/mig/ (accessed January 26, 2018).

[18] K. Rohlfing, D. Loehr, S. Duncan, A. Brown, A. Franklin, I. Kimbarra, *Comparison of Multimodal Annotation Tools - Workshop Report*. Online-Zeitschrift zur Verbalen Interaktion, Ausgabe, 7, 2006, 99-123.

[19] M. Tellier, G. Stam, B. Bigi, *Same speech, different gestures?* In 5th International Society for Gesture Studies, Lund Sweden, 2012.

[20] M. Tellier, G. Stam, B. Bigi, *Gesturing While Pausing In Conversation: Self-oriented Or Partner-oriented?* The combined meeting of the 10th International Gesture Workshop and the 3rd Gesture and Speech in Interaction conference, Tillburg The Netherlands, 2013.

[21] B. Teston, A. Ghio, B. Galindo, *A multisensor data acquisition and processing system for speech production investigation*. International Congress of Phonetic Sciences (ICPhS), University of California, 1999, 2251-2254.

[22] S. Wallis, G. Nelson, *Knowledge discovery in grammatically analysed corpora*. Data Mining and Knowledge Discovery, 5, 2001, 307-340.

[23] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, H. Sloetjes, *ELAN: a professional framework for multimodality research*. Proceedings of the fifth international conference on Language Resources and Evaluation, Genoa (Italy) 2006.

[24] S.J. Young, S. Young, *The HTK hidden Markov model toolkit: Design and philosophy*. Department of Engineering, University of Cambridge, 1993.

Antonio Rodà* - Alessandro Russo* - Sergio Canazza* - Piero Cosi**

AUDIO DOCUMENTS RESTORATION AS A DOCUMENTARY SOURCE IN THE LINGUISTIC RESEARCH: COMPARISON OF INSTRUMENTS

Abstract. – In the last years, always more speech audio archives digitized their historic corpora with the aim of preserving them and exploiting the advantages offered by digital signal processing techniques, such as formant analysis, automatic transcription, content-based information retrieval. Many of these applications are less effective when the signal to noise ratio (SNR) decreases, as it often happens in field-recorded linguistic corpora, due to the use of non-professional portable analogic devices and environmental noises. Audio restoration tools can therefore be very useful to increase SNR and improve subsequent analyses. At the same time, restoration algorithms should be carefully chosen and tuned to avoid the distortion of essential characteristics, such as formant position and energy. The paper describes, with examples taken from on field-recorded sound excerpts, several algorithms, able to cover different audio restoration categories. The algorithms were then evaluated by means of an automatic speech recognition tool. Results showed that the noise reduction algorithms can improve the phoneme recognition task carried out with the Kaldi toolkit.

1. Introduction

In the last years, archives and cultural institutions have started to be fully aware of the risks arising from unsuitable handling of sound documents in the remediation process, *e.g.*, during the transfer from analogue to new media for preservation and/or distribution and/or restoration purposes. This is also the case of archives which didn't receive adequate attention so far, in particular in the fields of ethno-musicology and speech (dialect, in particular), often intersected between them. These archives are often collected by the owners, in poor condition, using non-professional carriers. Nevertheless,

*CSC-Sound and Music Computing Group - Dept. of Information Engineering - University of Padova.

**National Research Council - ISTC - Via Martiri della Libertà, 2 - 35117 Padova.

these archives preserve recordings almost always in a unique copy, such as important documents for different research fields, including ethno-musicology, linguistics, and sociology. The value of these archives is evident by the gradual appreciation of the audio recordings by the international digital libraries. Several research projects have been funded by the European Community in the framework of preservation and digital curation[1] (*e.g.*, IST, Culture2000, Mediaplus, Interreg, eContent, Creative Culture, FP5-6-7), but very few of these have considered speech or ethno-music archives (*e.g.*, European project in Culture 2000 framework POFADEAM: Preservation and On-line Fruition of the Audio Documents from the European Archives of Ethnic Music). In addition, the multidisciplinary research that concerns the preservation of cultural musical heritage – which includes preservation and restoration of audio documents – fits perfectly in the Sound and Music Computing roadmap [2], as defined by the scientific community.

The growing interest in audio documents is motivated by the fact that the sound recordings are characterized by a shorter life expectancy than other cultural heritage goods (measured in years or decades, not centuries such as for sculpture and mosaics) and without concrete actions they are doomed to disappear. The urgency is heightened by the fact that the problems introduced by this type of sound files requires the definition of innovative approaches, specifically designed for the needs of these archives. In ethnic music and linguistic documents, a written notation generally doesn't exist (as in modern music Western tradition): the recordings are the only memory of those repertoires.

For the reason stated above, between the wide variety of audio recordings, speech documents represent a very interesting case study also from the preservative point of view. In addition, the preservation of this material is not limited to long-term preservation strategies, but rather includes all actions aimed to allow and improve the access to sound documents by experts and general users for study, research, reinterpretation, entertainment and more. Of course, there is a different situation in musical repertoires, such as clas-

[1]Digital curation [18] is the process of establishing and developing long term repositories of digital assets for current and future reference by researchers, scientists, and historians, and scholars generally.

sical, rock/pop, or jazz, where the record companies have already taken the necessary measures to protect the audio documents, particularly those of high commercial value.

This paper presents a preservation methodology (Section 2) designed for speech archives, some restoration tools developed by the authors (Section 3). They are used to process some audio files shared by the conference scientific committee (Section 4). Section 5 compares the results of automatic speech recognition carried out before and after the restoration, in order to observe if (and how much) the speech enhancement process has actually improved the phonemes recognition.

2. Preservation Methodology

Different approaches to the preservation of sound documents came to light during the debate carried out by the archival community over the last forty years about remediation process [5]. These approaches distinguish themselves for their different degree of accuracy to the integrity of the original document.

Depending on the approach adopted, the signal processing stages (after the A/D process) will be different. The authors are strongly convinced that a preservation intervention is preliminary to any other action. Regardless of the final use for which the recording is made, the quality of the starting digital material is critical, because is the only thing that can maintain a strict continuity between the original document and all copies made (low-quality extracts for online streaming, medium quality versions for local access, high-quality versions for scholars request, ...).

This section presents the preservation methodology defined by the authors on the basis of their experience in national and international research projects. In [5] is described in detail a general procedure including all the steps from A/D process to the restoration and the storage of the audio documents. In the following, the authors will discuss this issue, focusing on speech recordings.

When an authoritative version of an original documents disappears – because lost, stolen, accidentally destroyed, or corrupted for physical degradation – it's necessary to find a new document replacing the first as authoritative reference. The master copy of an audio document is exactly designed for this

purpose. As defined by International Association of Sound and Audiovisual Archives (IASA), the master (or archival) copy is "the artefact designated to be stored and maintained as the preservation master". Such a designation means that the item is used only under Exceptional Circumstances" [13]. Its purpose is to preserve the documentary unit and its bibliographical equivalents are the facsimile and the diplomatic copy. The restoration is permitted only for the repair/optimization of the physical carrier. The documentation accompanying the preservation master is defined by Schüller [17], who specifies that all the compensation and the processing applied during the remediation process must be "based on the capacity for precise counteraction" (which implies the reversibility of each operation and, consequently, the ability to go back to the original characteristics). For the master copy (logical) structure, see [5].

The actions starting from the assessment of the document condition to the time when the document is ready to be re-archived are included on the remediation process (see [5]). In general, a restoration laboratory can process multiple documents at once, although some activities require careful and constant control by the operator: in particular, the monitoring of the signal transfer aims to document any corruptions and other relevant events (related to the audio signal and not to the content). In the case in which the operator knows the (presumed) content of the recording, s/he can discern more clearly the noise produced by the recording process from non-intentional alterations resulting from the equipment of reproduction. The determination of the causes of a local corruption (clicks, pops, crackles, bumps) is not a trivial task. And the fact that the artifact is to be attributed to the playing equipment implies that the signal transfer must be repeated from the beginning. This kind of recognition can be performed only if the operator monitors the signal transfer during the whole process (including the silent parts). A lack of documentation of disturbances in the digital signal invalidate the reliability of the preservation master: actually, when a sound file will be analysed in future, and the original media will be lost or otherwise inaccessible, it would be impossible to determine the source of the noise. Monitoring, with the purposes described above, is effective only if the monitored audio is not taken directly from the playback device, but is first processed through the A/D conversion (*i.e.*, the

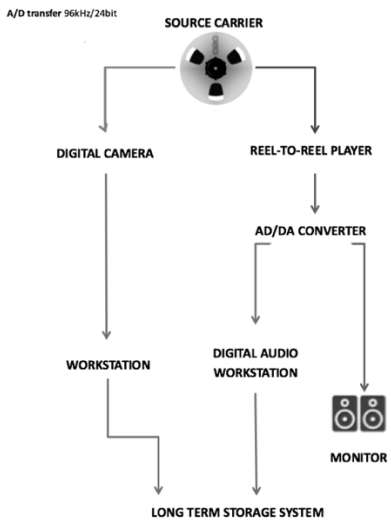

Fig. 1. – A generic signal flow scheme proposed by the authors.

signal should be redirected from digital workstation, via D/A converter, to the speakers or headphones input). In fig. 1 is reported a resuming signal flow scheme proposed by the authors.

2.1. *First and Second Level Access Copies.*

The aims of the digitization projects by archival institutions are: (i) the possibility to grant the access to a larger user base; (ii) the desire to promote new kind of access, and (iii) the documents preservation. If the ultimate goal is the documents preservation, the work process ends when the preservation copies are created. If the goal is the access, other kinds of documents must be produced. A crucial step missing between the conservative and, for example, audio files downloaded from an internet user, is the description of documents content. At this level, a change from a documentary approach to a content oriented one is required. The cataloguer sees in the magnetic tape (wax cylinder, phonographic disc, cassette, ...), not a sound recording featuring technical audio parameters, but the music, the interview, the conference (with the same gap that separates a manuscripts restorer from the one who studies the

text, essay, paper contained in the manuscript). The *first level access copies* are versions generally of lower quality than the preservation copies, and they are used by cataloguers to accomplish the analysis interpretation step. The main function of the first level access copies is to allow the access to the preservation master content, which must be "used only under exceptional circumstances" [14], and which are usually stored in different locations with respect to the archive with slow access (for example, tape drives for long-term storage). In the first level access copies, the length, the tracks number (and their order) should remain unchanged from the preservative master. At this level, some processing is permitted on the audio signal: for example, de-hiss and de-noise filtering for improving the Signal-to-Noise Ratio (SNR) could be useful in order to increase the intelligibility (speech enhancement) to help cataloguers.

The *second level access copies* are digital audio files generated during the cataloguing process. Their length depends on the content. The second level access copies are the containers of sound events (abstract entities) identified by cataloguers during her/his analysis and interpretation of content, organizing the sound material that was provided them in unmediated manner with respect of the content (first level access copies). For examples: (i) a work in mp3 format, from which the silent parts of the tape have been eliminated, and which has been divided into tracks corresponding to the parts of an interview; (ii) a collection of songs in a ethno-musical research. The result of the cataloguer work is a collection of digital resources of variable duration, which can vary in size and quality depending on the flexibility of the protocol adopted by the preservation project. In any case, the flexibility is permitted by the fact that a preservation master exists, and that it is possible to go back to it by means of the information provided by cataloguers during the step of creation of access copies.

The dichotomy between carrier and content (*i.e.*, between artefact and information) distinguishes audio recordings by other cultural heritage, such as sculptures and paintings: in these cases, preservation and restoration actions are directed to the object representing the heritage, that can't be separated from its physical expression [4]. Conversely, this separation can be applied to the audio recordings (the re-mediation process, described in this paper, per-

forms an abstraction of the content, in this way the content can survive in new containers), with the result that, in presence of a preservation master, (potentially) infinite number of different restorations (*i.e.*, interpretations) can be performed without affecting directly the source document. The second level access copies are a logic re-organization of sound material, but they are also the type of document that incorporates the results of the various restoration proposals.

3. Audio Restoration: Techniques and Tools

The speech audio documents are usually recorded in non-professional carriers by means of amateur recording system. Thus, for their appropriate fruition and/or for a suitable use of automatic speech recognition techniques, it is often useful to process the audio signals by means of audio restoration algorithms, during the creation of second level access copies.

The audio restoration algorithms can be divided into, at least, three categories [8]:

1. frequency-domain methods, such as various forms of non-casual Wiener filtering or spectral subtraction schemes and recent algorithms that attempt to incorporate knowledge of the human auditory system; these methods use little *a priori* information;

2. time-domain restoration by signal models such as Extended Kalman Filtering (EKF): in these methods, a lot of *a priori* information is required in order to estimate the statistical description of the audio events;

3. restoration by source models: only *a priori* information is used.

The advantage of frequency-domain methods is that they are straightforward and easy to implement. However, the limitations are as follows: musical noise (short sinusoids randomly distributed over time and frequency) is unavoidable; the results depend on a good noise estimation. Restoration by source model is limited to very few cases (*e.g.* only monophonic recordings) and it is not generalizable. The EKF is able, in principle, to simultaneously

solve the problems of filtering, parameter tracking and elimination of the outliers, but it is very sensitive to parameter setting (*e.g.*, the order model; the length of the signal vector, the length of the initial training segment in the bootstrap procedure, the adaptation speed, the threshold for detection of impulsive noise).

The authors developed innovative algorithms, using the Virtual Studio Technology (VST) plug-in architecture developed by Steinberg. The algorithms are (for a detailed description, from a computer science point of view, see [6]):

- CREAK (Canazza REstoration Audio - extended Kalman filter): a de-noise and de-click system based on Extended Kalman Filter [7], dedicated to the restoration of audio signal re-recorded from discs coated with shellac, the most common material used in the early production of the 78 RPM recordings, typically affected by low SNR, clicks, pops, crackle. At medium/high SNR, the performance of such filter is (at least) comparable to that of other standard methods like STSA (see below), nevertheless its plain use does not guarantee the best results. One reason for this is that the non-stationarity of audio signals leads to errors in parameter and tracking noise filtering, especially during fast transients. In order to achieve maximum performance from the EKF, it is essential to optimize its implementation.

- CMSR (Canazza-Mian Suppression Rule): a de-noise algorithm based on STSA (Short Time Spectral Attenuation), dedicated to the restoration of audio signal re-recorded from wax and amberol cylinders and shellac discs: low SNR. Adaptations of the spectrally-based techniques applied in speech processing have received most attention. They are based upon the well-known spectral weighting [3], in which individual spectral components are weighted according to expected noise and signal components. Such techniques can be viewed as finite block-size approximations to frequency domain Wiener filtering. As a result of these approximations (necessary to follow the time-varying nature of the useful signal) undesirable distortions can occur, the most notable being known as “musical noise” in which statistical fluctuations in the

frequency components of noise lead to random tonal artefacts in the processed signal. Various techniques have been applied to mask or eliminate these distortions. CMSR adopts the Ephraim and Malah suppression rule [12].

- PAR (Perceptual Audio Restoration): a de-hisser based on perceptual algorithm for reel-to-reel tapes and cassettes: high SNR. The "perceptual" processing task requires to transform the audio signal from an "outer" to an "inner" representation, that is to resort to a representation that takes into account how the human ear perceives the sound. The combination of the psychoacoustic model and frequency-domain algorithms, permits to define a promising restoration methodology [9].

Some files, chosen among the set shared by the conference scientific committee, are restored, using CREAK, PAR (CMSR war discarded because it is specifically designed for wax cylinders) and a commercial software tool (iZotope RX, based on a STSA algorithm able avoid the "musical noise" artefact, designed to repair, enhance, and restore audio signals, improving sound quality and clarity), with different setups. In particular, files named "oro" and "maiale" (8 s long, wave format) were chosen. No further information about the way the files were recorded are known by the authors.

4. Audio Restoration: Results

4.1. *iZotope RX Audio Restoration Tool.*

Audio files were processed with a -20 dB de-noise filter (iZotope RX, Algorithm Type D) and with a 50 Hz de-hum filter. The intention was to apply a massive removal of the noise floor, extracting the vocal part as much as possible. Although this process might increase the risk of creation of audio artifacts, it also enhances the speech components and allows improved automatic phoneme recognition by the software. The noise profile was extracted from the first second of the record, in a fragment in which vocal parts were not present (fig. 2). In fig. 3 is showed an example of noise profile, extracted from the first second of a not processed audio file. After the processing,

the spectrograms show how the spoken part appears more defined and emphasized in comparison to the noise background (fig. 4). Spectrograms were obtained with the free software Sonic Visualizer developed by London University [10]. In the first seconds of the fragment is particularly clear the action of the de-hum and the -20 dB de-noise filter, especially in the frequency range between 100 and 1000 Hz, which appears now less intense. In some cases, the effect of the de-noise algorithm can be also observed along the whole spectrogram as a lighter color band in the frequency range between 1000 and 1700 Hz (fig. 5). In fact, observing the noise profiles it is possible to notice how in this particular frequency range the noise components are more intense. In order to further emphasize the extraction of the spoken parts, a second -20 dB de-noise filter has been applied. The second noise profile was extracted from the residual audio information still present in the first second of the record after the DD (Denoise-Dehum) processing (fig. 6). As shown in fig 7, the second de-noise algorithm caused a significant further enlargement of the lighter color band around 1000 and 1700 Hz, due to the attenuation of middle-high range frequencies and the high amount of information removed. In fact, although the noise background was almost removed in its totality, the filter influence area included also fundamental components of the speech part.

4.2. *Perceptual Audio Restoration*

In comparison to the results obtained after the application of the algorithms developed by Izotope, the application of the de-hisser based process PAR has led to a minor alteration of the original audio files. PAR was applied with three different growing intensities: “Soft”, “Medium” and “High”. As shown in fig. 8, as opposed to the former results, the background floor reduction is now inferior and it slowly increases with an heavier application of the algorithm. In fact, vocal component frequencies are emphasized and they appear more intense in the high-intensity spectrogram, in comparison to the low-intensity one. It is possible to notice that, in this case, there is no massive intensity reduction in the frequency range between 1000 and 1700 Hz, as confirmed by the absence of the lighter colour band like the one that was created after the application of the former commercial algorithm. With

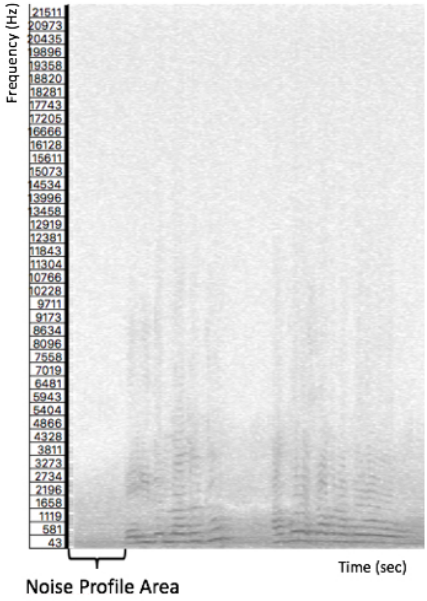

Fig. 2. – A generic noise area chosen as target for the extraction of the noise profile.

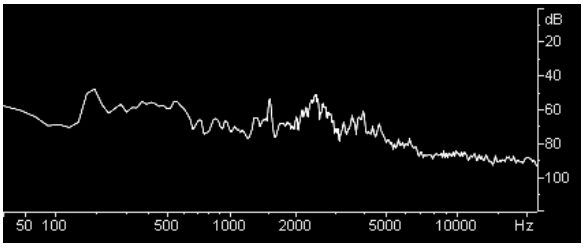

Fig. 3. – Noise profile extracted from “oro” audio file.

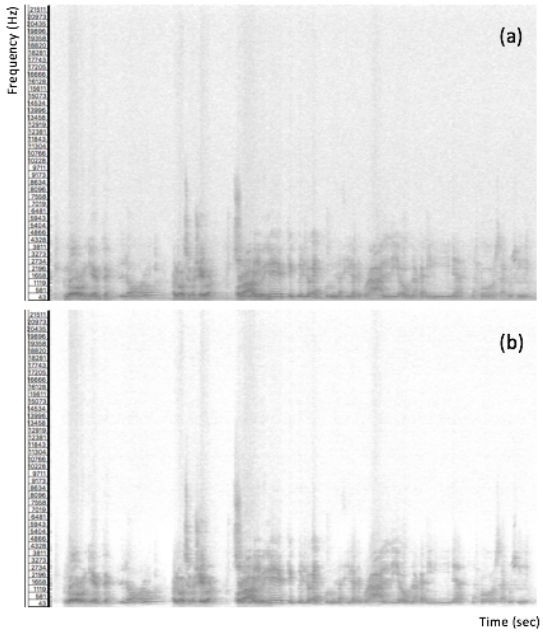

Fig. 4. – Comparison between “oro” audio file not processed (a) and after the application of de-noise and de-hum algorithms (Izotope RX).

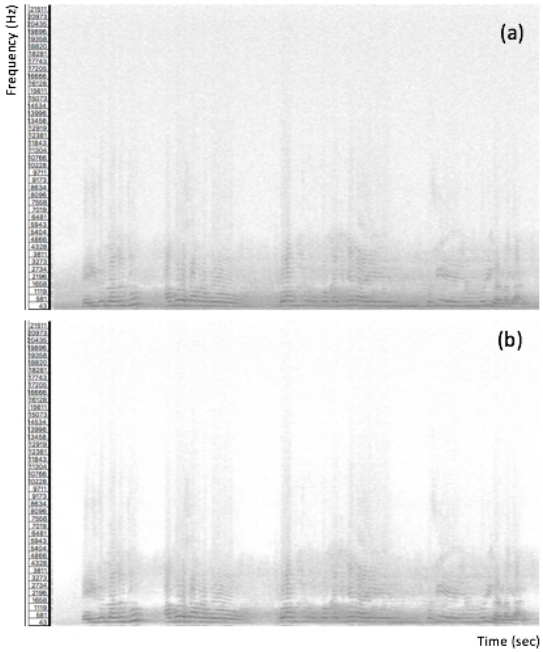

Fig. 5. – Comparison between "maiale" audio file not processed (a) and after the application of de-noise and de-hum algorithms (Izotope RX) (b). It is possible to observe the lighter colour band in the frequency range around 1000-1700 Hz, due to the removal of some fundamental audio components from the speech part.

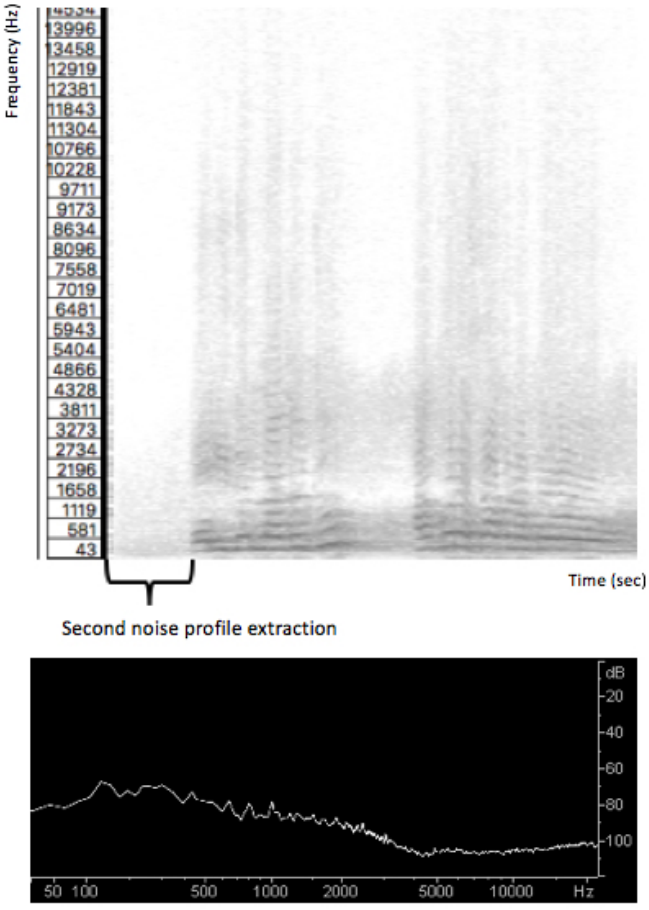

Fig. 6. – Extraction of a second noise profile, used as target to process the file with a second de-noise algorithm (Izotope RX).

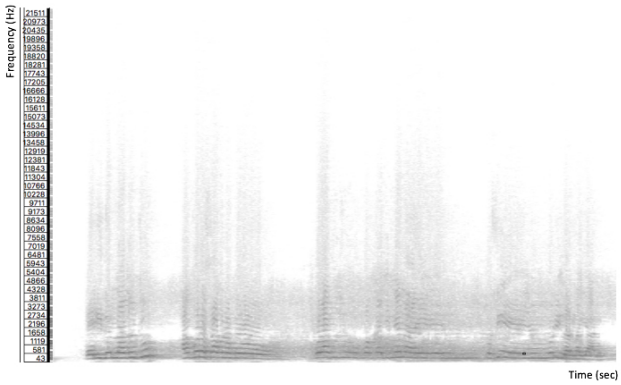

Fig. 7. – Spectrogram of "maiale" audio file after the processing with the first de-noise filter (-20 dB, Izotope RX), de-hum and the second de-noise filter (-20 dB, Izotope RX). It is possible to observe the spread of the lighter colour band due to the massive removal of fundamental audio components.

the application of PAR process, high frequencies (>1.5 kHz) components of the audio signal are also emphasized.

In general, this process permitted to optimize the frequency attenuation for the speechless parts, in comparison to the vocal ones, less deprived of frequency components.

4.3. *Extended Kalman Filter.*

The application of this particular restoration chain led to less evident results and the processed files result equal or slightly different from the original ones.

Neither the analysis of the spectra profiles showed any relevant differences in comparison to the not processed ones (fig. 9). In fact, the noise components appear slightly less intense in the files processed with the EKF, but almost in a imperceptible way. Generally, EKF best results are obtained with low SNR and for impulsive noises like pop and crackle and not for continuous noise disturbs like the ones that the processed files presented; that's probably the reason why, in comparison with the iZotope commercial algorithms and the

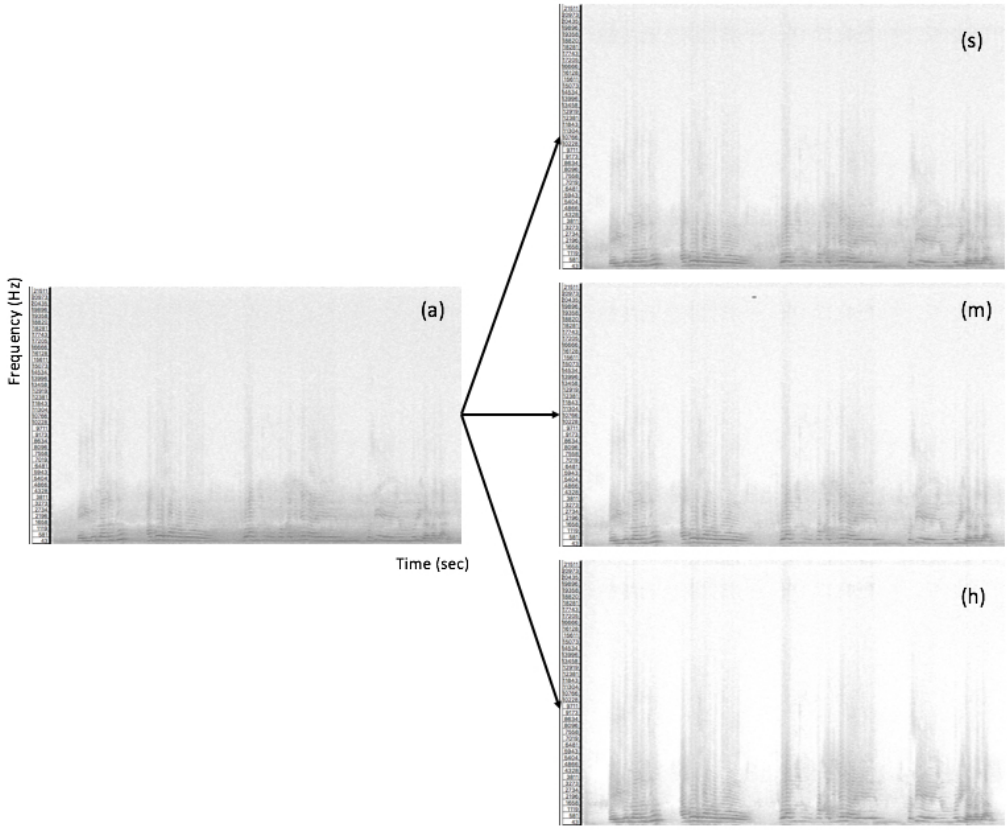

Fig. 8. – Comparison between unprocessed “maiale” audio file (a) and the same file processed with PAR algorithm with three different intensities (s), (m) and (h).

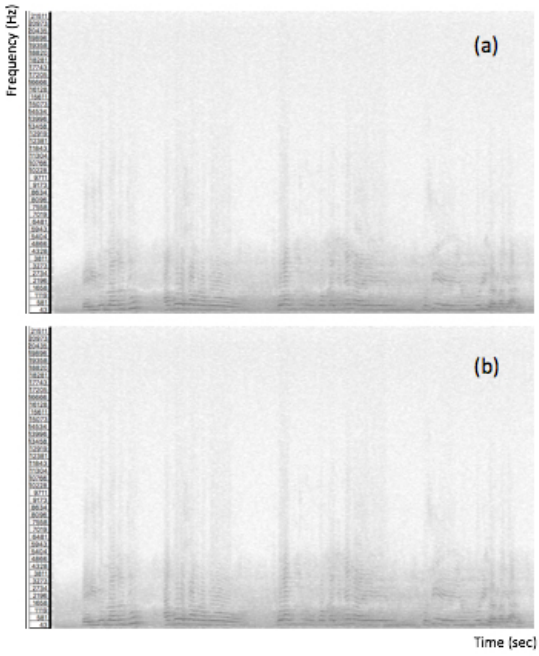

Fig. 9. – Comparison between unprocessed "maiale" audio file (a) and after the application of the EKF (b).

PAR, EKF leaded to a less efficient removal of the undesired audio information.

5. Evaluation by means of ASR

Automatic speech recognition is always more used in many applications, from multimodal human-computer interfaces and virtual assistants to automatic transcription and annotation. With the aim of evaluating how the noise reduction process influences the automatic recognition of speech, the original recording and the de-noised versions described in the previous section were analyzed by an acoustic model implemented by means of Kaldi [16], a free and open-source toolkit for speech recognition research. Kaldi includes tools for calculating standard features, such as MFCC and PLP, and supports

model-space transformations and projections both speaker-independent, *e.g.* LDA and MLLT/STC, and speaker-specific, such as fMLLR. Moreover, Kaldi supports conventional acoustic models (*i.e.* diagonal GMMs) and Subspace Gaussian Mixture Models (SGMMs). As the excerpts to be evaluated are both in Italian language, we employed the adaptation of Kaldi to Italian detailed in Cosi *et al.* [11]. The Italian FBK APASCI [1] was employed to train the acoustic model. APASCI is an Italian speech database recorded in a silent room with a Sennheiser MKH 416 T microphone; it includes 5290 phonetically rich sentences and 10800 isolated digits, for a total of 58924 word occurrences (2191 different words) and 641 minutes of speech material which was read by 100 Italian speakers (50 male and 50 female).

Table 1 shows the performance of the speech recognition algorithm applied to the excerpts "maiale" (A) and "oro" (B), both the original versions and the processed ones. The results are reported in terms of Phoneme Error Rate (PER), calculated on the basis of the Levenshtein distance [15], and defined as

$$PER = \frac{S + D + I}{N}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the number of phonemes in the sentence. Consequently, a lower value means a better recognition.

Overall, the PER is included between 61.84% and 80.88%. As one might expect, the phonemes of the excerpt "maiale", that is characterized by a worse sound quality, are less recognizable in comparison to excerpt "oro". In all the cases but one, the PER is less for the processed versions in comparison to the original recordings, meaning that the noise reduction algorithms improved the automatic phoneme recognition. Only the version processed with the Extended Kalman Filter of the excerpt "oro" obtained the same PER of the original recording: this result could be explained by the fact that the excerpt "oro" was recorded with a quite good sound quality and the application of the EKF, that is known to work better with low SNR [6], resulted in a signal that is very similar to the original one.

Apart EKF, which results are equal or slightly lower than the original recording, the noise reduction algorithms improved the phoneme recognition by

Table 1 – Phoneme Error Rate calculated for both the excerpts "maiale" (A) and "oro" (B) is reported. The original recordings have been processed by different de-noise algorithms (DDD = Denoise+Dehum+Denoise, DD = Denoise+Dehum, EKF = Extended Kalman Filter, P = Perceptual, H = High, M = Medium, S = Small).

	PER [%]
A	80.88
ADDD	76.47
ADD	75.00
AEKF	77.94
APH	72.06
APM	75.00
APS	75.00

	PER [%]
B	68.42
BDDD	64.47
BDD	61.84
BEKF	68.42
BPH	64.47
BPM	65.79
BPS	65.79

an amount that depends on the type of algorithm and the applied settings. As concern the excerpt "maiale", the best performance (PER = 72.06%) is obtained by the PAR (Perceptual Audio Restoration) algorithm with the setting High, that implies a higher noise attenuation. Regarding the excerpt "oro", instead, the commercial tool iZotope obtained the best performance (PER = 61.84%) by applying in sequence a de-noise and a de-hum algorithm (DD). In general, PAR worked better with the setting High that removes more noise, whereas iZotope obtained a better performance with the sequence DD, that removes less noise in comparison to the sequence Denoise-Dehum-Denoise (DDD).

6. Conclusions

The state-of-the-art techniques in audio noise reduction offer interesting tools for improving the quality of audio recordings, as a preliminary task for subsequent linguistic analyses. In this paper, a set of commercial and research tools for audio noise reduction have been described and applied on a couple of on-field recorded excerpt. For each original record, a set of different post-processed versions was obtained. The utilization of the commercial

de-noise and de-hum algorithms developed by iZotope was intentionally heavy and the amount of audio information removed was finalized to enhance and emphasize the speech part as more as possible, permitting a best intelligibility of phonemes by the software. In addition to the files processed with the commercial algorithms, other versions were obtained processing the audio material with the Perceptual Audio Restoration (soft, medium and high intensity) and with an Extended Kalman Filter-based restoration algorithm (CREAK).

While EKF did not lead to significant results and the files appeared to be almost identical to the original ones, in PAR processed files the speech components were emphasized, without an excessive alteration of the original record.

The noise reduction algorithms were then evaluated by means of Kaldi, a toolkit for automatic speech recognition. Results showed that, at least for the analyzed cases, the noise reduction algorithms reduce the phoneme error rate, improving the automatic recognition task.

Some limitations of the present paper should be acknowledged. Firstly, the results were obtained with only two short (8 s long) audio excerpts and a more extensive set of audio recordings would be necessary to have more significant results. Secondly, the values of PER obtained with the ASR tool are quite high: for such values, the improvement due to the noise reduction algorithms, with a phoneme error rate however greater than 60%, could not imply a real usefulness for automatic annotation/transcription tasks. The low performances in terms of PER could be explained with the differences between the dataset used to train the acoustic model and the two analyzed excerpts. In particular, the training dataset is composed by Italian utterances recorded with professional equipment in a controlled environment (anechoic room), whereas the excerpts were recorded on-field, with dialect inflections. Specific adaptation of the acoustic model is therefore necessary to improve the recognition task.

References

[1] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, M. Omologo, *A baseline of a speaker independent continuous speech recognizer of Italian*. Proceedings of the 3rd European Conference on Speech Communication and Technology, 1993, 847-850.

[2] N. Bernardini, X. Serra, M. Leman, G. Widmer, *A roadmap for sound and music computing*. The S2S2 Consortium, 2007.

[3] S.F. Boll, A.V. Oppenheim, *Suppression of acoustic noise in speech using spectral subtraction*. IEEE Trans. Acoustics, Speech and Signal Processing, 2(27), 1979, 113-120.

[4] C. Brandi, *Teoria del restauro*. Edizioni di Storia e Letteratura, Roma 1963 (*Theory of Restoration*, trad. Cynthia Rockwell, rev. Dorothy Bell, Firenze, Nardini 2005).

[5] F. Bressan, S. Canazza, *A systemic approach to the preservation of audio documents: Methodology and software tools*. Journal of Electrical and Computer Engineering, 2013.

[6] S. Canazza, *The digital curation of ethnic music audio archives: from preservation to restoration*. International Journal of Digital Libraries, 2/3(12), 2012, 121-135.

[7] S. Canazza, G. De Poli, G.A. Mian, *Restoration of audio documents by means of Extended Kalman Filter*. IEEE Trans on Audio Speech and Language Processing, 6(18), 2010, 1107-1115.

[8] S. Canazza, A. Vidolin, *Preserving electroacoustic music*. Journal of New Music Research, 4(30), 2001, 289-293.

[9] S. Canazza, G. De Poli, S. Maesano, G.A. Mian, *On the performance of a noise reduction technique based on a psychoacoustic model for the restoration of old audio recordings*. Proceedings of Diderot Forum, 1999, 29-35.

[10] C. Cannam, C. Landone, M. Sandler, *Sonic Visualiser: An Open Source Application for Viewing, Analysing, and Annotating Music Audio Files*. Proceedings of the ACM Multimedia International Conference, 2010.

[11] P. Cosi, G. Paci, G. Sommavilla, F. Tesser, *Kaldi: yet another ASR toolkit? Experiments on adult and children Italian speech*. Proceedings of 11th Convegno Nazionale Associazione Italiana di Scienze della Voce, 2015, 429-438.

[12] Y. Ephraim, D. Malah, *Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator*. IEEE Trans. Acoustics, Speech and Signal Processing, 21(6), 1984, 1109-1121.

[13] IASA-TC 03, *The Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy*. IASA Technical Committee, 2005.

[14] IASA-TC 04, *Guidelines on the Production and Preservation of Digital Audio Objects*. IASA Technical Committee, 2004.

[15] V.I. Levenshtein *Binary codes capable of correcting deletions, insertions, and reversals*. Soviet Physics Doklady, 10(8), 1966, 707-710.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely *The Kaldi speech recognition toolkit*. IEEE workshop on automatic speech recognition and understanding, 2011, 404-439.

[17] D. Schüller, *Preserving the facts for the future: Principles and practices for the transfer of analog audio documents into the digital domain*. Journal of Audio Engineering Society, 7/8(49), 2001, 618-621.

[18] Yakel, *Digital curation*. OCLC Systems & Services. International digital library perspectives, 23(4), 2007, 335-340.

George Brock-Nannestad*

THE UNREHEARSED RECORDING: SOURCE-CRITICAL OPTIMISATION OF THE SOUND THAT WAS CAPTURED

Abstract. – The corpus prepared for the workshop is pedagogical and represents a low quality, sometimes found in unrehearsed tape recordings of the 1950s-70s. Each recording needs special technical treatment, and one approach to analysis is duly presented in the paper. Most recordings are now historical, and the originator cannot help us. Focusing on corpora recorded via the mechano-acoustic process a source-critical approach is introduced in order to correctly handle distortions of speech recordings perhaps more than 100 years old. Most known digitalisations have not handled these distortions, and although they were performed 1980-2005, the equipment used was not calibrated. Some utilisations require re-digitalisation.

1. Introduction and fundamental considerations

The alluring promise of recording technology has at all times been "transparency": the trust that what is reproduced is a true representation of the phenomenon it was wished to record. A similar mirage of "true representation" is known from photography, another technical aid to memory that historically appeared only a few decades earlier. Transparency is only possible if one of two conditions is fulfilled. One is a completely standardised and fully documented recording-reproduction chain, from the placement of the microphones to the placement of the loudspeakers and with full control of the sound pressure levels involved.

The other is that a thorough knowledge is available as to the recording procedure, the conservation of the carrier that represents the recorded phenomenon, and the conditions of reproduction. In the beginning of audio technology this was a painstaking work, but entirely possible and it was done in practice for both scientific and commercial recordings, albeit for different reasons.

*Patent Tactics - Resedavej, 40 - 2820 Gentofte (Denmark).

Alas, as technology progressed the physical manipulation involved diminished, and with it, the respect for the problems of the recording-reproduction chain. Reliance on the recording grew without foundation, and this may lead to undetected erroneous conclusions [9].

1.1. *Scientific recording*

Scientific recording, such as that for ethnomusicology and the phonetic sciences was mostly annotated, but certain recordings had to occur when the informant was ready and not necessarily in a laboratory environment. In the present text such recordings are termed "unrehearsed recordings", and they do pose special problems.

In the absence of a standardised and fully documented recording-reproducing chain, the ideal situation is obtained when everything concerning a recording and its subsequent reproduction for analysis is meticulously annotated, because that permits the determination of the influence of each contributing influence and suitable compensation. That is not the case with the "unrehearsed recording", because then reliance has to be put on more general knowledge when dealing with the influences. The commercial record companies made full notes of the actual parameters for any given recording, because they might be required to have a repeat of the "rehearsed recording" at a later time in order to create the proper illusion [4] of a continuous, live performance when the record was reproduced – alone or in a particular sequence. These notes were commercial secrets, however, and obviously still only made sense in connection with the apparatus used at the time. But they were essential in order to recreate the musical ambience and balance at the repetition. Due to the confidentiality of the information it was not spread and it is in most cases not available for study.

The elements that unite the study of ethnomusicology, performance practice, and the phonetic sciences are 1) it is the study of human sonic behaviour as it is determined by physical abilities and culture, and 2) the apparatus used to enhance the study process [5]. The principle of sound recording and later reproduction of that sound recording has permitted first of all sufficient repetition of the recorded phenomenon to permit focusing on details in an aural

annotation process and secondly machine-assisted analysis [10]. It is only in later years that it has been realised from within these disciplines that the similarities between the fields are greater than the differences, and that e.g. tools from one field may be applied to another [11]. It should be noted that a recent re-assessment of the method of Wilhelm Wundt [19] for documenting prosody discusses his use of the musical scale, although he was not the first [7][1].

We shall introduce the field of *source criticism* as a general approach to sound recordings as sources, irrespective of their provenance. The sound emanating from a loudspeaker or through headphones comes from an electronic input to the amplifier, and that electronic input is derived from the stored version of the sound originally uttered by the informant. Moving in a field in which recordings were and are intended to preserve utterances, we do not need to concern ourselves with the processed sound that goes into modern artistic (and, obviously, commercial) recordings, in which an actual utterance that has been witnessed live may constitute a minute percentage of the actual sound presented via a loudspeaker. There are hence many phenomena that have no need for identification and evaluation in the present context.

We shall also introduce the concept of the *value* of a recording for a specific scientific or academic purpose. This is done by means of a *thought experiment*: we want to compare the value of two magnetic tape recordings (a) and (b) that are considered to be of the same sonic event – this is what is available. The reason it is useful to introduce the concept via two recordings is that the value of one recording taken alone is infinite: the alternative would be no recording at all.

How would we approach this problem? Superficially these recordings appear to be very different. (a) has a bright tone and appears to be spoken rather quickly by young female voices. (b) is both at a lower level and with a smaller treble content, but with a more "normal" cadence to the speech.

[1]It is here not the purpose to perform an extensive analysis of the historical approaches to responsible handling of recorded and annotated sound, but it must be noted that it would cover a time span from ca. 1860-1960.

Nothing more than the date of the event and the names of the informants are available as metadata.

One may approach the problem of possible identity of the recorded sonic event from several sides. One tape is shorter than the other, but this is not due to editing: there are no splices in the tapes as presented. Proper comparison requires several replays and timings. It is in fact an investigation that uses techniques from the forensic field, and this is the response to the *thought experiment*: the absolute requirement to use forensic techniques.

At the time these tapes were recorded the tapes were fresh, and nobody would have considered multiple reproductions and wind-rewind operations to be problematic. This is not the case today, 30-60 years later: tapes are in danger of destruction if they are mechanically stressed. This means that analytical listening (repetition) can no longer be safely performed. Instead, a transfer of the recordings by merely one replay to a different medium is necessary and all analytical work will have to be done on the thereby created copy. Before the replay is performed the replay parameters have to be adjusted based on non-invasive analysis of the magnetic tape. A magnetical track developer permits the determination of the track configuration and thereby the choice of magnetic replay head. The azimuth of the historical recording head cannot be determined with precision by means of the track developer, and the first transfer must be monitored and the signal must present a maximum of high frequency. This is obtained by varying the azimuth of the replay head. This, hence, may necessitate several reproductions after all. There are several other mechanical features of the tape-machine interface that have to be taken into account, including degradation of the tape itself. A reasonably rich literature exists on this aspect of the transfer process.

Modern copying involves a digital target file. In view of the fact that much fundamental phonetic research was done with a maximum frequency of 8kHz one might consider that bandwidth is not a problem. However, in case advanced noise analysis and reduction is contemplated, it is recommended to use a file based on minimum 96 kHz sampling frequency and 24 bit resolution, which is the international archival recommendation [18].

Once the best possible pickup of the signal has been obtained from the tape and transferred at a known speed of reproduction to the digital domain it is

feasible to perform reproductions, loopings, and signal processing as required by aural analysis and measurement.

2. The Study Corpus of the International Workshop – an appreciation

The purpose of a corpus of speech recordings may be defined beforehand, or it may be used to extract information that was not considered or suspected at the time of recording. Frequently a corpus is established subsequently to stocktaking according to some organising principle.

Speech may be analysed with respect to phonetics, phonemics, linguistics, mood, narrative. This is predominantly related to the communicative aspect of speech as a source. But speech may equally be merely taken as an input signal and may be used to determine the physical input-output characteristics of the recording chain.

In the present corpus of digital files based on tape recordings we are presented with five time slices and the annotation that has been possible. They have been presented with nothing more than file names, and hence no prior knowledge is available concerning any methodological grouping of the items. Indeed, the sounds appear as if they were found on the shelf of an archive in individual small tape boxes without any markings. Are they worthless in this condition? No, not entirely. There is no information whether the annotation is made on the basis of the original tapes or from the digital files. The information made available is essentially the syllables identified, and notes about pauses and places where the syllables have been unintelligible to the annotator. There are no field notes, nor any metadata that might assist the interpretation. In a way, these files may be considered a worst-case situation.

However, some of these digital files contain much more that must be considered relevant secondary information: the noises and disturbances. Some of these are masking and contribute to the unintelligibility. This places these files in the field of audio forensics. Some of the results of audio forensic analysis are primarily directed towards intelligibility and would not provide true phonetic information on the voice as a source.

On the other hand, however, with the lack of analytical metadata it cannot be excluded that the original tapes, properly treated, may yield more secondary information that will enhance the primary information.

In terms of quality of sound, the following ranking of the Study Corpus may be made by means of observation and FFT-imaging:
oro.wav: good level and treble, some hum-type background noise
vestiti.wav: very similar to oro.wav, but with more pronounced hum-type background noise
CG5.wav: low level, very little treble (upper frequency limit 4 kHz), possible azimuth mismatch, cyclic ventilator or compressor background noise
maiale.wav: reasonable level, very little treble (in essence, upper frequency limit 6.5 kHz), road-type background noise
Ayoreyo.wav: loud, very wide frequency range, overload and clipping (on positive peaks), bass emphasis, reverberation, some hum-type background noise

3. Let us make a short historical overview of the consciousness of source critical issues

In 1909 Abraham and von Hornbostel wrote "Vorschläge für die Transkription exotischer Melodien" [2], in which chapter II deals with the technique for transcribing. This came after an earlier publication, in which a responsible technique for *recording* was defined [1]. It may be derived that a recording would have to be replayed at the very least 20 times. It must be remembered that many of the recordings were made, not only by ethnographers, but also by other travelling scientists, missionaries, and colonial officials. Hence it was important to devise uniform methods for transcribing, and although the words "source criticism" were not used, the consciousness was quite obviously present. Perhaps to our modern astonishment it is worthwhile to examine the experiences expressed in 1909 in the form of proposals.

3.1. *The steps in Abraham and von Hornbostel (1909).*

According to the custom in "comparative musicology" (the previous name for "ethnomusicology") one goal of collection was to transcribe the words and

music into the Western notation system. Special signs were used to indicate cases of deviation from any of the 12 semitones to one octave. The recordings had in most cases not been done by specialists, but musically endowed persons would receive the recordings and perform repeated listening in order to identify the features to be written down on a time axis. Many of the phenomena had to be revisited many times, and Abraham and von Hornbostel attempted to inject confidence in the endeavour (pp. 16-17) by identifying problems and their resolution, e.g. by repetition – or repetition at slower speed, by not being caught in an erroneous conclusion too early, etc. In cases of the musicologist possessing absolute pitch, a suitable adjustment of the speed is to occur in order to obtain a minimum of deviation signs while satisfying the needs of the endowed ear. If the words had been written down at the time of collection, that text could be used as a control by comparison with the words and syllables understood from the recording. Some errors attributable to the equipment are also openly discussed.

On pp. 16-18 there is a discussion of the measurements that may be undertaken on the sound recordings. The current frequency measuring instruments are discussed as is the need for a constant rate of revolution of the cylinder. All frequency measurements are to be connected to the time axis established in the form of the transcription obtained by listening.

Already now, the modern reader will realise that very many reproductions might be needed, and that it would be surprising if anything of the recording was left; it would simply be worn down. This fact was quickly understood in early scientific wax cylinder recording, and the practice was developed (p. 19) that the field-recorded cylinder would be given a coat of copper by electrotypical means; the original cylinder would be removed by melting and rinsing with a solvent, and the remaining copper tube that had the modulation of the cylinder as a negative on the inside would be used as a master for casting any number of copy wax cylinders needed for the task of repetition. The wax would be an alloy that was more durable than the original cylinder. The copper negative became the archive master and has to a large degree survived to this day.

Even though it is not readily apparent that Agostino Gemelli in the 1930s reproduced the sounds that he recorded on film, there is no doubt that he wor-

ked to the best technical standards in his work [15] in analysing the tracings. In this respect he must also be considered to be so much aware of the technical limitations that he was able to include that awareness in his analyses, and thereby work source-critically. In [16] the value of reproducible sound recordings is discussed, and although the work is concentrated on photographical oscillographic tracings, there is ample evidence of a source critical approach. Like so many in both phonology and linguistics Gemelli used standard texts to be read aloud by the test persons.

3.2. *An ethnomusicologist with a philosophical approach.*

In 1983 Bruno Nettl's fundamental and philosophical work "The Study of Ethnomusicology. Twenty-nine Isssues and Concepts" appeared [21]. In Chapter 6 the automatic recording or analysis is treated. One will also find a reference to the Norwegian Karl Dahlback's report "New Methods in Vocal Folk Music Research" [12], which is discussed as an example of the "new development" (post-Metfessel [20]) [10] in the 1950s, and Nettl describes Dahlback as a pioneer. However, in the whole book by Nettl one finds no doubt as to the sources, and Nettl does not appreciate that Dahlback's most important contribution are not the actual measurements or technical methods. Rather, it is his source-critical approach to obtaining true recordings by means of a tape recorder. On 20 pages with full documentation Dahlback treats what the precise requirements are to both recording and reproduction. It is important that this concerns the modern medium of the time – the tape – and not early recordings. A recent literature search has revealed that no ethnomusicological work, nor any work relating to use of tape recordings makes reference to the fundamental problems of the tape recording method. We shall return to Dahlback's work.

The next important connection between early sound recordings and reproduction technology was the Japanese symposium on B. Piłsudski's Phonographic Records [13]. Some of the fundamental discussions were repeated in the introduction to the catalogue "Early Field Recordings" [23] from the Archives of Traditional Music. However, the discussions were not systematic and only gave an impression of responsible handling of the original recordings

in the Archives proper and were not really of great help to outsiders.

We get the impression that the modern user of a historical recording would rather not deal with these technical matters; it is much easier just to trust the technology implicitly. Not even a conference paper from 1990, "Die Schallaufzeichnung als historisches Dokument" (The sound recording as a historical document) [22] indicates that a source critical method exists, and yet that method was originally developed for the study of history! That, which was found as essential to the founders of this field, probably because the phonograph was unreliable, yet extremely useful, was forgotten and completely neglected until 1993. This was most likely because the simpler-to-use (and prima facie better sounding) modern recording techniques seduced everybody involved. For about 70 years there was no development and conscious adoption of source criticism in the field of sound recording.

4. The pioneering work of Karl Dahlback

Dahlback worked with folk song recorded in many places in Norway, and he used the modern tool, the tape recorder. For his folk song study purposes he had a melody-writer constructed, and as it was a very precise and stable instrument it was possible to measure and display very small frequency deviations as a function of time. He realised that as the tape recorder was the source of the signals that it was intended to measure, he needed to establish the quality obtained from tape recording and reproduction. Dahlback set out to determine the usefulness of a number of tape recorders available in Norway for this work, and among other parameters he determined their sensitivity to frequency and voltage variation.

4.1. *Background*

Norway had an electricity supply that was dependent on local, essentially hydroelectric generation and no national grid. In contrast to for instance the United Kingdom (where local, historical conditions determined both frequency and voltage well into the 1960s), Norway had a defined goal of 50 Hz average frequency and 220 V household voltage. The hydroelectric ge-

neration provided power that was available without marginal costs, and the subscription was for installed wattage, not consumption. This was an advantage for the stability of frequency and voltage, because load-dependent regulation of a hydroelectric plant is very slow [14]. Such was the power supply available to the travelling researchers who recorded Norwegian folk song in the 1950s.

Dahlback's methodological work is to be found pp. 18-38. He introduces the subject tersely by stating "Scientific folk music research based on such recordings presupposes, as a matter of course, *that the sound tape will provide an exact reproduction of the subject matte*r" (Dahlback's emphasis). He then discusses the various influences observed. Using crystal-controlled oscillators and a high grade professional tape recorder he recorded test tapes that were subsequently reproduced on a number of consumer tape recorders typically used in ethnomusicological field research, while varying the supply voltage and/or frequency. He also traced start-up, and found that for some tape recorders, 15 seconds might elapse before the speed was stable and constant. A series of experiments were made by recording controlled frequencies on consumer tape recorders that were later reproduced on the same machine, at times generating e.g. twice the flutter of a single pass through the tape recorder in question. After having determined the extent of the problem Dahlback associated with the laboratories of the Norwegian Broadcasting Corporation to run a series of tests of commercial and professional tape recorders. The risks involved in corrupting the signals by using undocumented tape recorders are very clearly expressed in his report.

After a footnote that expresses his incredulity that folk music collectors have in practice also used *wire* recorders (ca. 1946-1953) for fieldwork (because these display deviations up to two semitones when reproducing a test tone that has been recorded), Dahlback concludes with the following conditions that have to be fulfilled (p. 39):

Tapes

- must be run through (without recording) and checked before use
- must be relatively free from electrostatic charge

Tape recorders

- must be tested in the laboratory with measuring tones and melody-writer (or with frequency meter if desired) in order to determine the starting-graphs, warming-up time following variations of temperature, sensitivity to external disturbances (in transit), and diagrams of mains frequency/voltage so that sources of error may be duly taken into consideration

Recording and playback

- to be controlled by measuring the mains frequency and voltage

The recording

- not to be used if conditions vary to any appreciable extent; allowance for variations in current or constant current values depends upon the diagrams of the apparatus and upon the degree of accuracy of reproduction deemed to be necessary for the research in hand

- to be reproduced in the laboratory under similar conditions of current (corresponding values) as existed when the recording was being made.

Indirectly he also stresses the need for a precise logbook in which the conditions of recording have been noted. In essence, although details obviously differ, his recommendations correspond very well with those of Abraham and von Hornbostel in [1].

Nobody can doubt the importance of Dahlback's folk music work, which forms the main body of his report, in particular as it came at a time when several competing systems for automated melodic analysis were under development. However his respect for the subject and his request for the best and calibrated recording and reproducing equipment for his sources has much more far-reaching consequences, an important one being that his recordings may be revisited with perhaps even better or more time-saving equipment, provided the modern transfers of the original tape recordings have been done to the same or better standards.

4.2. *The problem of copies and 'generations'.*

We have seen above that it was unrealistic to consider an original cylinder recording as durable enough for the thorough analysis requiring more than 20 reproductions, and for this reason copies were cast, thereby giving access to any required number of 3rd generation copies. Why 3rd? Because the negative obtained by electrodeposition and used for casting and archiving was the 2nd. With careful work, very little high frequency was lost and only a little noise was added.

Let us inspect fig. 1 (from [6]), which in a very compact form identifies the various versions that are available to researchers. We see that we need to distinguish between analyses and versions made available by the original collector and by researcher who merely receives "a sound recording". There is also a need to be aware of the difference between a modern transfer of the earliest generations (1st or 3rd) made by non-informed technicians, by informed technicians, or supervised by a researcher versed in the subject (phonetics, ethnomusicology).

There is no doubt that the experienced listener (here first and foremost the original collector) will be able to hear and identify phenomena that are not readily available to somebody without experience in the subject matter. Hence the experienced listener will be able to work with worse material, in the sense that he or she is more tolerant towards disturbing noises. And similarly, such researchers would be able to work with a low-quality transfer. However, this should not be an excuse to aim for less than the optimal in transfer work. Nobody can predict the future user.

5. Conclusion

Transparency has always been the desire of scientists who wish to make use of recordings of phenomena. It has been defined in general terms as an input-output process that does not contribute distorting factors to the output. Transparency has also been behind all commercial sound recordings, but mainly as a lofty illusion for most of our recording history, because of the use of montage.

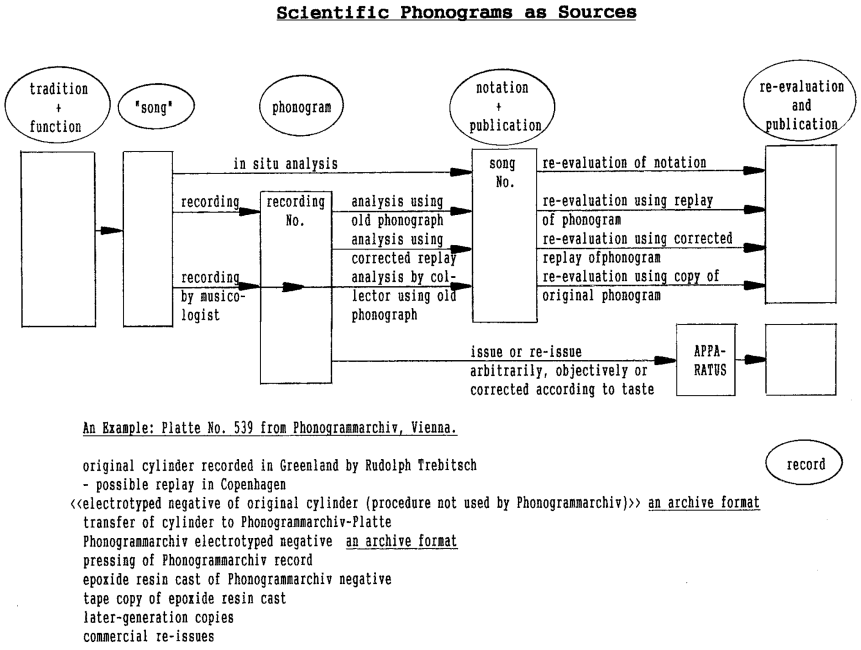

Fig. 1. – The source value of various access copies of scientific recordings.

The mechanical complexities of early sound recording were so great that it was impossible to be unaware of the distortions introduced in the recording-reproduction process, and it is conjectured that this caused an awareness of a necessary protocol for dealing with recordings. As recording as a process more and more approached push-button service, this awareness of protocol was diminished, and even when dealing with recordings that are only 50 years old, transparency is expected and not worked consciously towards. It has been the purpose of the present paper to provide access to the field of *source criticism* and the manner in which such early sources are documented, even *post facto*.

Appendix: A modern corpus of sound recording transfers from originals dated 1905-84

Modern generalised access to early recordings occurs via transfers into the digital domain, such as editions of compact discs or as files to be downloaded from the internet. This in essence means that there has been a monopolistic access to the originals, and the version provided contains a technical and sometimes an aesthetical interpretation by those responsible for the edition. Any researchers desirous of using such a material are dependent on the knowledge and skill of those who have crystallised this particular version. For this reason, the so-called *archival transfer* is universally accepted as a version that is as unmodified as possible (IASA TC03), in other words that it contains a maximum of the signal and noise present on the original recording. Many archives world-wide have such a store of transfers, which are then further processed in order to make them appeal more to the general public in the form of the published corpus. It is a difficult balance between making an unmodified version available and the possibility of recovering some of the cost of issuing a corpus by sales. A fee is furthermore usually required for accessing the "original" archival transfers.

This is a very general problem, and we see more and more academic work based on such secondary sources. If this fact were recognised and perhaps reacted on by compensatory measures to improve the transfers, or at least by discussing the difference between an original and a copy, such academic work could still have permanent value. In the present discussion I shall limit myself to one particular corpus published in a narrow ethnomusicological field.

The study of Greenland Inuit music had always been performed by a very small group of researchers, but in 1992 a monograph was published [17] accompanied by a Compact Disc with the recordings that were discussed in the monograph. The corpus as presented and the book that was expected to contain the necessary metadata were reviewed by the present author [6], and attentive listening and some measurements determined that 21 recordings displayed technically problematic transfers as shown in fig. 2. A number of recordings had not even been identified. A source-critical approach was derived from a previous approach [3] that had been developed for the field

of commercial sound recording (the documentation problems as discussed above), and the various versions possible were identified (fig. 1). This brings the compilation into the field of "unrehearsed recording", and many of the recordings represented in the corpus have to be subjected to analyses such as those presented in the present text. If we look at fig. 2, which is a part of a table analysing the complete published corpus we may note that here are many omissions, both of technical standards and of metadata. Relevant metadata would have comprised information on whether the transcript and transfer were made on the basis of the original cylinder, or via a tape recording. In the latter case, the original speed of the cylinder as well as the speed of the tape recorder would have been relevant. We are given the duration of each song but not how this was determined, e.g. was it from the original cylinder or measured via tape.

The aural analysis/appreciation of a corpus is based on the experience of the analysing listener. In mechanical records (disc or cylinder) there is frequently the advantage that a surface scratch will give cause to a regular clicking, which may be timed, whereby the rpm of the record may be determined. An experienced listener to a mechanical record will recognise different types of distortion and their probable causes, which permits to distinguish between e.g. local wear of the groove and a stylus that does not fit the groove. Such approaches have created the "Observations" in the present table. A special mention should be made of track no. 27, because it refers to a tape recording, the main subject of the present text. The lack of treble may be caused, either by an extreme deviation between the azimuths at recording and at reproduction, or, more likely, that the tape was played with the non-magnetic surface of the tape towards the magnetic head.

References

[1] O. Abraham, E.M. von Hornbostel, *Über die Bedeutung des Phonographen für vergleichende Musikwissenschaft*. Zeitschrift für Ethnologie, 36, 1904, 222-236.

[2] O. Abraham, E.M. von Hornbostel, *Vorschläge für die Transkription*

Analysis of 55 recordings published on CD ULO75 – "Traditional Greenlandic Music"
with reference to the book "Traditional Greenlandic Music" by Michael Hauser (no date, but 1992)

Source identifier	Page	Track on CD	Recorder/Year	Observations
21 transfers which are technically problematic:				
T&T 88	44	4	Thalbitzer 1906	rumble, ca. 180 rpm, bad tracking
T&T 121	48	5	Thalbitzer 1906	low pass, 1'44" phonograph struck on replay
T&T 70	50	7	Thalbitzer 1906	180 rpm (meas. from 30"), coughs, flutter=bad tracking
T&T 52	53	8	Thalbitzer 1906	low pass
T&T 8	57	10	Thalbitzer 1906	very resonant, distant, shallow cut/oversize stylus
T1923 p.543	79	13	Thalbitzer 1914	resonant, DRUMS
T1923 No. 143	99	21	Thalbitzer 1914	220 rpm?? (p. 100: "fast"), filtered, noise gating
T&T 130	118	22	Thalbitzer 1905	noise, filtered
?	144	23	Leden 1912	no bass, echo – wrong stylus
?	145	24	Treb./Stias. 1906	5 kHz low pass = PhA 593 or 594
?	149	25	Treb./Stias. 1906	5 kHz low pass = PhA 586
?	150	26	Leden 1912	distortion
?	151	27	Jensen 1959	TAPE: no treble at all!! hum
?	152	28	Leden 1912	filtered, no bass
?	199	30	Holtved 1937	LAQUER: no treble
R&J 35	206	36	Jenness 1916	echo, noise – wrong stylus? low pass
?	217	39	Holtved 1937	LACQUER: 77 rpm
?	218	40	Leden 1909	low pass, rumble
?	220	41	Lumh./Haus. 1984	TAPE: clear drums
?	223	43	Holtved 1937	LACQUER: low pass 2kHz?, needle skip 4 grooves at 38"
R&J 42	230	44	Jenness 1916	weak, echo, low pass 2 kHz?, gain riding!

furthermore among the above, 11 give no precise reference to source for a given song.

Fig. 2. – Extract of sound quality observations of the corpus of Greenlandic music from CD ULO75.

exotischer Melodien. Sammelbände der Internationalen Musikgesellschaft, 11(1), 1909, 1-25.

[3] G. BROCK-NANNESTAD, *Zur Entwicklung einer Quellenkritik bei Schallplattenaufnahmen*. Musica, 35, 1, 1981, 76-81.

[4] G. BROCK-NANNESTAD, *Tracking the Sound an Exhibition on Recording and Reproduction of Music*. The Museum of Music History, Copenhagen 1987.

[5] G. BROCK-NANNESTAD, *An Evaluation of Early Use of Sound Recordings In the Analysis of Performance Practice and in Phonetics*. In: I. FUCHS (ed.), *Internationaler Musikwissenschaftlicher Kongress zum*

Mozartjahr 1991 Baden-Wien. Bericht II. Schneider, Tutzing 1993, 505-514.

[6] G. Brock-Nannestad, *Review of Michael Hauser: Traditional Greenlandic Music, Copenhagen 1992*. IASA Journal, 1, 1993, 89-92.

[7] G. Brock-Nannestad, B. Elverdam, *Pierre Schaeffer's Objets Musicaux and Their Relationship to Phonetic Research*. In: E. Schubert, K. Buckley (eds.), *Proceedings of ICoMCS*. Australian Research Council HCSNet, Sydney 2007, 23-27.

[8] G. Brock-Nannestad, *The Sound Recording As a Source To Performance*. In: M.E. Marinelli, A.G. Petaccia (eds.), *Musicus Discologus 2, Musiche e Scritti per l'80º anno di Carlo Marinelli*. Edizioni ETS, Firenza 2007, 203-221.

[9] G. Brock-Nannestad, *The Use of Somebody Else's Sound Recordings. Source-Critical Complexes when Working with Historical Sound Recordings*. In: R.E. Mohrman (ed.), *Audioarchive. Tondokumente digitalisieren, erschliessen und auswerten*. Waxmann, Münster 2013, 25-32.

[10] G. Brock-Nannestad, *The mechanization of performance studies*. Early Music, LXXII(4), 2014, 623-630.

[11] D. Cooper, I. Sapiro, *Ethnomusicology in the Laboratory: From the Tonometer to the Digital Melograph*. Ethnomusicology Forum, 15(2), 2006, 301-313.

[12] K. Dahlback, *New Methods in Vocal Folk Music Research*. Oslo University Press, Oslo 1958.

[13] Executive Committee of the International Symposium, *Proceedings of the International symposium on B. Piłsudski's phonographic records and the Ainu culture*. Hokkaido University, Sapporo 1985.

[14] F. Fischer, H. Lichte, *Tonfilm. Aufnahme und Wiedergabe nach dem Klangfilm-Verfahren (System Klangfilm-Tobis)*. Verlag, Leipzig 1931, 250.

[15] E. Galazzi, *Agostino Gemelli et l'analyse electro-acoustique du langage*. In: L.J. Boë, C.E. Vilain (eds.), *Un siècle de phonétique expérimentale. Fondation et éléments de développement. Hommage à Théodore Rosset et John Ohala*. ENS Éditions, Lyon 2010, 179-190.

[16] A. Gemelli, G. Pastori, *L'analisi elettroacustica del linguaggio*, Vol. I. Società Editrice, Milano 1934, 1-54.

[17] M. Hauser, *Traditional Greenlandic Music*. Copenhagen 1992.

[18] IASA Technical Committee, *The Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy*. IASA Technical Commitee Standards, Recommended Practices and Strategies, IASA-TC 03, 1997.

[19] G. Konopczynski, C. Dodane, *Wilhelm Wundt: une proposition originale de transcription musicale de la prosodie*. In: L.J. Boë, C.E. Vilain (eds.), *Un siècle de phonétique expérimentale. Fondation et éléments de développement. Hommage à Théodore Rosset et John Ohala*. ENS Éditions, Lyon 2010, 259-276.

[20] M. Metfessel, *Phonophotography in Folk Music: American Negro Songs in New Notation*. Univ. of North Carolina, Chapel Hill 1928.

[21] B. Nettl, *The Study of Ethnomusicology: Twenty-nine Issues and Concepts*. University of Illinois Press, Champaign-Urbana 1983.

[22] D. Schüller, *Die Schallaufzeichnung als historisches Dokument*. In: *Fortschritte der Akustik - DAGA, 90. Plenarvorträge und Kurzreferate der 16. Gemeinschaftstagung der Deutschen Arbeitsgemeinschaft für Akustik*. Deutsche Physikalische Gesellschaft e.V., Bad Honnef/Wien 1990, 71-87.

[23] A. Seeger, L.S. Spear, *Early Field Recordings: A Catalogue of Cylinder Collections at the Indiana University Archives of Traditional Music*. Indiana University, Bloomington 1987.

Per ordini d'acquisto rivolgersi a:

BARDI EDIZIONI srl
Corso del Rinascimento, 24 – 00186 Roma
tel.: +39 06 6877815
fax: +39 06 68807739

segreteria@bardiedizioni.it
www.bardiedizioni.it