Multimed Tools Appl (2012) 60:327–346
DOI 10.1007/s11042-010-0610-3

LIT: transcription, annotation, search and visualization tools for the Lexicon of the Italian Television

Thomas M. Alisi · Alberto Del Bimbo · Andrea Ferracani · Tiberio Uricchio · Ervin Hoxha · Besmir Bregasi

Published online: 6 October 2010

Abstract LIT (Lexicon of the Italian Television) is a project conceived by the *Accademia della Crusca*, the leading research institution on the Italian language, in collaboration with CLIEO (Center for theoretical and historical Linguistics: Italian, European and Oriental languages), with the aim of studying frequencies of the Italian vocabulary used in television. Approximately 170 hours of random television recordings acquired from the national broadcaster RAI (Italian Radio Television) during the year 2006 have been used to create the corpus of transcriptions. The principal outcome of the project is the design and implementation of an interactive system which combines a web-based video transcription and annotation tool, a full featured search engine, and a web application for data visualization with text-video syncing. Furthermore, the project is currently under deployment as a module of the larger national research funding FIRB 2009 VIVIT (*Fondo di Investimento per la Ricerca di Base, Vivi l'Italiano), which will integrate its achievements and results within a semantic web infrastructure.*

Keywords Video transcription · Annotation · Streaming · Multimedia indexing · Retrieval · Semantic web · User experience design

T. M. Alisi (✉) · A. Del Bimbo · A. Ferracani · T. Uricchio · E. Hoxha · B. Bregasi
Media Integration and Communication Center, University of Florence, Florence, Italy
e-mail: thomasalisi@gmail.com

A. Del Bimbo
e-mail: delbimbo@dsi.unifi.it

A. Ferracani
e-mail: andrea.ferracani@unifi.it

T. Uricchio
e-mail: tiberio.uricchio@gmail.com

E. Hoxha
e-mail: ervin.hoxha@gmail.com

B. Bregasi
e-mail: besmir.bregasi@gmail.com

1 Background

A video is made up of audio and visual information, and providing direct access to content through video metadata is essential for the integration and study of multimedia materials, either in traditional desktop or rich internet applications. For this reason, annotated corpora are fundamental components of applications designed for linguistic research. The exploitation of research tasks on corpora of speech data relies at least on a basic transcription, while further linguistic analyses usually require that the transcription is aligned with the speech. These basic tools and functionalities are very important to achieve results in the field of computational linguistics [6].

Annotation tools, both manual and automatic, of audio-video materials have proliferated in recent years in different research fields (Praat [19], Transcriber [24], Anvil [4], ILSP [13], the NITE Workbench [18] and many others [1]), but the scattered set of domains in which they were involved and the lack of truly recognized standards did not help in the stabilization of emerging technologies. Moreover, the rapid evolution of applications from traditional desktop to rich internet and networked environments makes further developments and deployments of research products which could have a twist towards technology transfer, and thus maturation, even more difficult.

The research effort made towards the definition of standards has produced numerous results, and usually each of them applies to its specific domain. The common approach when it comes to research and development of tools for metadata handling is to have a decent compatibility with XML-based standards: performances are usually maximized with strictly proprietary applications and storage systems, while interoperability is guaranteed through the use of well known interchange protocols, such as SOAP/REST [21] web services (both based on the HTTP application layer of the Transfer Control Protocol), or RDF/SPARQL [20] endpoints for semantic web applications.

The underlying structure of annotation tools is thus delegated to protocols and procedures whose fortune and acceptance may vary, depending on research and technology trends. One of the most reliable and diffused projects for linguistic annotation of the early 2000s is the Architecture and Tools for Linguistic Analysis Systems (ATLAS) [8] which, as stated from the official website, *"addresses an array of applications [that] needs spanning corpus construction, evaluation infrastructure, and multi-modal visualization. The principal goal of ATLAS is to provide powerful abstractions over annotation tools and formats in order to maximize flexibility and extensibility"*. The project delivered an annotation tool which relied on its proprietary XML-based format (Atlas Interchange Format, AIF) and was intended to become a diffused standard, but its support was subsequently dismissed and parts of the project merged in the forthcoming technologies of automated recognition developed by the Multimodal Information Group. Several other projects has been developed in the field of language annotation, but most of them are either discontinued or based on old technologies and proprietary formats [15].

The gap in terms of standard definitions was progressively filled by two well known projects: TEI and MPEG-7, which were able of delivering open standards and spread them in different contexts, thanks to their versatility and easy of use. The Text Encoding Initiative (TEI) [5] is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. TEI is an XML vocabulary defined by an XML Schema which suggests a substantial implementation of structural and functional formalisms of literary

texts in the structural organization of the elements of a markup language, focusing on the description of logical and functional structures of documents.

MPEG-7 [22], formally named "Multimedia Content Description Interface", is a standard for describing the multimedia content data that supports some degree of interpretation of the information meaning, which can be passed onto, or accessed by, a device or a computer code. MPEG-7 is not aimed at any one application in particular; rather, the elements that MPEG-7 standardizes support as broad a range of applications as possible.

The two standards have reached wide diffusion and have been adopted by a large number of products and projects, the latter being designed mainly for the description of the properties and features of multimedia files, while TEI strictly focusing on linguistic issues. For this reason an application that deals with specific aspects of language analysis is more likely to find a friendly development environment using the TEI standard, while providing mappings for MPEG-7 conversions is still important in order to be compatible with such a large part of the research world.

Even assuming that an application is using either TEI or MPEG-7 as a data interchange format, evidences show that research trends recently moved towards automatic annotation of video content and speech recognition for automatic subtitling, both the techniques having principal outcomes in commercial environments. This assumption is reflected by the numbers of relevant publications and projects in the field of automated annotation and recognition, but unfortunately these important results in multimedia are progressively reducing the availability of updated resources in the linguistic research field of study.

Video annotation and speech recognition, in conjunction with other technologies like natural language understanding, are indispensable for multimedia indexing and search, nonetheless all research fields related to computational linguistic usually require a fine-grained identification of paragraph levels and a definition of corresponding metadata descriptors: these outcomes are hardly achieved through automatic image and speech analysis. Speech recognition is error prone and its accuracy dramatically depends on quality of acoustic conditions and type of data sets: it can be used for library indexing and retrieval in specific contexts—some experiments show that data retrieval from transcripts of spoken documents can be just 3–10% worse than information retrieval on perfectly human transcribed data [11]—but it is not mature enough to meet the specific needs of linguistic research. In addition, multilingual speech recognition raises the issue of training the detectors (e.g. the Sphinx-III continuous speech recognition system developed at Carnegie Mellon) on acoustic and lexicon data models of different languages or dialects in order to achieve higher performances, thus increasing the overall complexity of systems and algorithms due to the analysis of the phonology, morphology, syntax and prosody of specific languages.

Recent experiments have shown that, in addition to an accurate speech transcription, it is possible to get some specific information such as: the gender of the speaker, a rough categorization of the type of acoustic environment (e.g. music, noise, clean, etc.), the beginning and the end of new stories, the top level story topics and a rough speaker clustering [3]. Unfortunately these results still present high error rates and these tools are not capable of achieving good performances, if not in scenarios with little variability.

Extracted data based on image features, color histograms or optical flow analysis are recently achieving encouraging results and can help building language models from visual features [3] and lately the traditional machine learning techniques have been reinforced with semantic web algorithms as in [7]. The american National Institute for Standards and Technology (NIST), which is also currently hosting the evolution of ATLAS with its

spawned sub-projects, has even established the TREC conference series, whose goal is to encourage research in information retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results, still the level of precision required by linguistic research and the price/performance ratio provided by manual transcription is unachievable by automatic systems supported by image analysis.

Another important aspect considered during the development of the project presented in this paper was the availability of online annotation and transcription tools, since YouTube has recently added new features for video annotation and transcription. Users can send invitations for shared annotations and add three types of information: speech bubbles, notes and spotlights. YouTube allows users to upload captions and subtitles and files with videos transcription, even if still in beta and only for English language. Although very useful, the system is still too rough and generic for the specific domain of linguistic research, since it does not allow necessary customizations for specific contexts and user defined taxonomies.

For these reasons this paper presents a set of tools where up to date techniques for rich internet applications are integrated with video streaming, content indexing and multimedia retrieval, targeting the specific sector of web applications for linguistic research. The project is part of a series of activities defined along with the *Accademia della Crusca* and CLIEO for the design and implementation of a set of integrated tools for linguistic annotation handling. The *Accademia* is a leading institution in the field of research on the Italian language: presently, a principal activity is the support of scientific research and the training of new researchers in Italian linguistics and philology through its Centres and in cooperation with Universities. The CLIEO (Center for theoretical and historical Linguistics: Italian, European and Oriental languages) was founded in order to provide a confluence into a single research and higher education entity of different institutions previously active in Florence in the field of Linguistics: University structures (Department of Italian Studies; Department of Middle Age and Renaissance Studies; Department of Linguistics; Inter-Universitary Center for the Geolinguistic Study of Proverbs), the Accademia della Crusca, the Opera del Vocabolario Italiano—Italian Dictionary (OVI, a CNR Institute), and the Institute of Legal Information Theory and Techniques (ITTIG, a CNR Institute).

2 Architecture

Since LIT is designed in order to satisfy specific requirements of its users, a use case scenario is the first step for the creation of a correctly targeted application: two classes of users are identified for the system were transcriptionists and researchers, and Fig. 1 reports a simplified version of the tasks they have to undertake.

Expanding each task into a complete use case scenario (where each action is described in details, e.g. type of connection, subtasks with detailed description of sequence diagrams, etc.), leads to the definition of specifications and requirements, where key factors are:

- Large use of standards. Using diffused standards for storing annotated data is a key factor to allow interoperability with other systems and, if needed, achieve simplicity in creating software adaptors to allow sharing data with other systems.
- User experience design (UXD) [10, 17]. User experience highlights the experiential, affective and meaningful aspects of Human-Computer Interaction (HCI), but also covers people's perception of practical aspects such as utility, ease of use and efficiency.

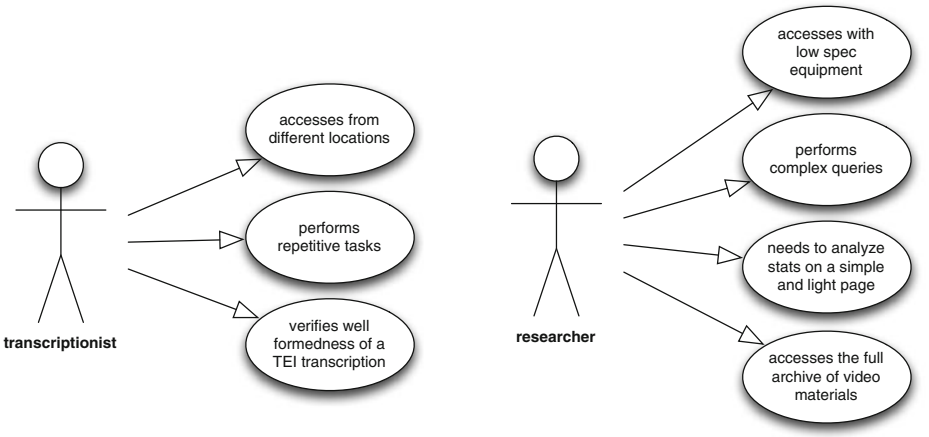

Fig. 1 Use case scenarios

The use case scenario for LIT was initially developed taking into account the pool of researchers and users targeted for the system, considering they have mainly a linguistic background, and have to deal with delicate and repetitive tasks: a typical scenario for scientists and technicians. Some aspects and interactions have been thus designed minimizing users' cognitive effort while approaching to a complex interface.

- Rich internet application (RIA). A RIA is a web application executed by a browser plug-in, providing functions and interactions usually associated to desktop applications. The key advantages of using a RIA paradigm while deploying LIT are given by centralization and the availability of advanced functionalities through a web browser, such as: asynchronous communication, improved performance due to local processing on the client, drag and drop, transitions, sliders, interactive grids and tables, client side form checking, stacked or tabbed collapsible panes, modal and non-modal dialog boxes. The use case scenario defined for the project clearly shows that users have remote access to the system, henceforth the video database should allow distributed access: these requirements are enough to determine a RIA environment, in order to adopt a "deploy once, run anywhere" strategy.
- Modularity. LIT is designed to be part of a bigger picture, where results coming from annotation have to be analyzed in a semantic web environment; in this sense, it has to be easily plugged in a larger context, and thus provide a set of endpoints where common protocols are used.

The design of the system focused on effectiveness and efficiency while performing tasks required for the specific application domain: usability was constantly evaluated taking into account comparisons between how targeted users felt while performing specific tasks and how technical requirements were achieved. The system thus offers friendly interfaces, optimized to minimize the learning curve for targeted users, both for annotation and retrieval, easy to learn and responsive, allowing users to rapidly annotate and search multimedia material without a specific knowledge of technical jargons or engineering background. Thanks to the exploitation of RIAs potential, users can annotate a transcription of a television broadcast using uniquely a graphical interface, without need of manually editing xml files or specific tags. A large amount of information can be presented using RIA components (Flex allows easy interface design providing an XML-based development

platform) and data are shown graphically and can be manipulated interactively. Every task is accomplished on demand and in realtime without multiple steps or pages being reloaded, thus reducing servers load.

The project presents two different interfaces: a search engine, based on classical textual input forms, and a multimedia interface, used both for data visualization and annotation (latter functionalities being activated after authentication) and the whole systems relies on a backend implemented to handle TEI transcriptions and provide the necessary indexing and search functions. The interface for multimedia data handling, thanks to the use of the RIA paradigm, is multi-platform (Windows, Linux, Mac), runs on any browser equipped with a standard Adobe Flash player [14], conforms to general hardware requirements for the targeted audience (thus avoiding possible performance issues) and has extended support of transcriptions and annotations using any encoding format (ANSI, UTF-8, UTF-16, ASCII, etc.). Using a multi-encoding enabled environment means that indexing and search functions can be easily extended to support all kind of language and characters (the system currently works with all western languages, with special functionalities defined for accented characters).

The requirements and specifications described have been taken into account and detailed in the following sections, with regards to three main actions:

1. definition of the interchange format, conforming to a standard schema;
2. learnability, memorability and satisfaction of an annotation and visualization web interface;
3. efficiency of a fully featured indexing and search engine.

2.1 Standard in the interchange format

Since LIT has to focus more on linguistic issues than general multimedia handling approach, the data interchange format chosen is XML-TEI, the international standard mentioned above. TEI can be used for describing humanistic and historical documents in digital format: key features are that it is XML-based, portable and customizable. Moreover, TEI has already been used for another project funded by the *Accademia* and thus presents an interesting outcome for the integration of data provided by different contexts, which will be taken into account during the integration of the system in a larger research project funded by the Italian ministry of education and research, the FIRB 2009 Vivit.

The LIT project is the first web-based tool that allows video annotations in this format: since it has to deal with its own definition of contexts in annotations, using a standard which allows schema customizations is mandatory, hence TEI represents a perfect choice. The revision chose for the annotation system is the latest P5, and section 8 of TEI guidelines define specifically the structure of speech transcription (a complete reference can be found at http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html). TEI offers great flexibility when defining the structure of a text transcription, and assuming the basic XML schema is respected, specific elements and attributes for the definition of a proprietary structure can be defined: an online tool then can be used for compliancy verifications.

A TEI document is composed of blocks where global metadata definitions are generally placed at the beginning of the document inside a <teiHeader> element, while subsequently a set of one or more <TEI> elements define the content itself, made of local metadata instances.

For this reason, the structure of a document can be schematized as in Fig. 2.

The standard is extensible with regard to the definition of the taxonomy used, which can be modified without affecting backward compatibility of records in the database. The categories,

defined as identifiers in the general teiHeader block, are then referred in the specific local instances of TEI blocks by elements defining a specific recording present in the annotation.

The categories and sub-categories defined in LIT are:

- advertising
- fiction (sub-categories: TV film, short series, series, serial)
- entertainment (sub-categories: variety show, game show, reality show, humour and satire)
- information (sub-categories: news, reportage, in-depth report, live report)
- scientific and cultural (sub-categories: documentaries, magazines)
- talk shows (sub-categories: political, cultural, sports)

The cinema category is totally missing because, due to its nature of being programmatically prepared for a specific distribution, it does not present evidences of language evolution interesting for research as in the rest of the footage.

The recording element contains all the data useful for the identification of a specific broadcast, including the name of the broadcaster, date and time, transmission and category. The most important part of a single TEI block is the set of utterances, which clearly defines the transcription of the recording and is accompanied by additional definitions of production details and references to cue-points defined in the overall timeline in the teiHeader global definitions. The details used for the specification of an utterance are:

- type of communication (can be: monologue, dialogue)
- type of speech (can be: improvisation, programmed, executed)

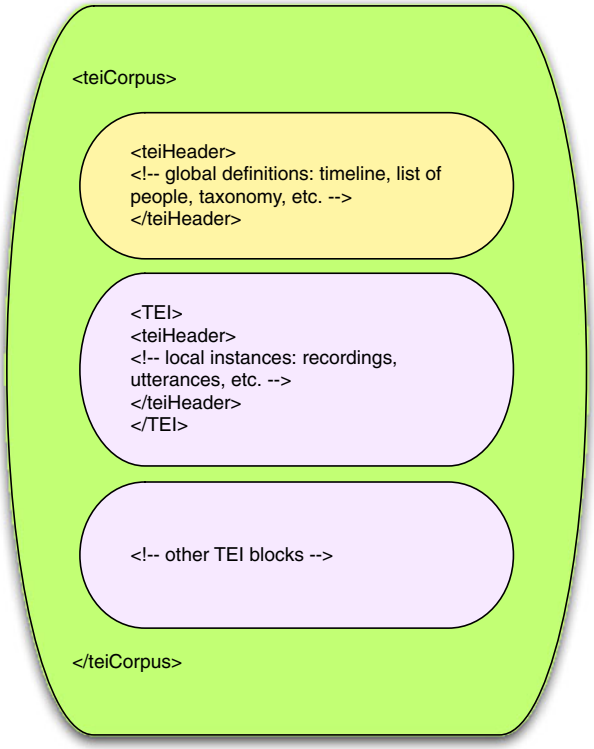

Fig. 2 Structure of a TEI document

- gender of speaker (can be: male, female)
- type of speaker (can be: professional, non professional)
- speech technique (can be: on scene, voice-over)

A single utterance in a TEI transcription appears thus as follows:

```
<u who="#Person49" start="#t1" end="#t2" comtype="dialogo" spokentype="esecutivo"
env="esterno" voice="incampo">che c'è Rex ? che succede ?</u>
```

where the 3 attribute parameters "who", "start" and "end" are defined as references to identifiers declared in the global header of the file as:

```
<timeline origin="" unit="s" id="timeline1">
<when id="t1" absolute="00:00:02:665"/>
<when id="t2" absolute="00:00:05:174"/>
</timeline>
```

and

```
<person id="Person49">
<persName>
<forename>pm1</forename>
</persName>
<sex>Maschio</sex>
<provenance>interno</provenance>
</person>
```

As it clearly appears from the example reported, TEI allows to define specific needs for annotation in terms of attributes and elements of the XML schema, which are then validated through an online tool. Its flexibility is a strong plus for the adoption of the standard in the linguistic field of research, and the schema generated for the transcription becomes an important building block of the search engine, where optimized XML based solutions can be effectively used for information retrieval, as shown in following sections.

2.2 Annotation and visualization

The term "linguistic annotation" refers to any analytical or descriptive note that can be applied to linguistic data collected in the form of textual data. There are essentially two methods for annotation of linguistic corpora: automatic and manual, the latter being more diffused mainly for its formal correctness of contents where this feature is a major requirement. In recent years there has been a growing interest in statistical-based language processing: these kind of systems need a training set provided by manual annotators, and results provided by manual annotations usually work as ground truth for performance evaluation of automatic systems. Furthermore, automatic systems normally present some limitations and high error rate in text to speech alignment due to speaker's voice characteristics, poor signal-to-noise ratio and overlapping speaker's voices. While it is fairly easy to automatically obtain time aligned transcriptions, it is far more difficult, if not impossible, to determine and identify other crucial characteristics such as:

- single utterances, due to speech elliptics, overlappings, disfluencies et alia;
- speaker's identity and gender;
- the characteristics of her/his pronunciation;

- the environment in which the characters are speaking;
- the exact punctuation, capitalization and prosodic features of the text from a linguistic point of view [12].

For these reasons the choice made for LIT focused on a manual annotation system, in order to deliver and manage speech to text transcriptions far more accurately than automated systems allow.

It has already been outlined how annotation tools should focus on specific classes of annotation problems, in order to make the process of annotation more efficient. Such issues in fact have great influence on the design of specialized tools used to manually create annotations. From this point of view it is useless, if not impossible, to develop a tool that covers all specific needs of different annotations, but it is necessary to design a system that effectively complies with the specific requirements of a domain [9]. If correctly annotated, a corpus of multimedia materials can be reusable by other subjects and research projects independently from the specific use case scenario, henceforth LIT is implemented as a specialized tool that makes use of a data model based on a standard interchange format. Developing a manual annotator therefore was a decision dictated by specific needs of linguistic research. Even if this decision led to a fairly high cost in terms of human resources (several linguists have transcribed and annotated all the multimedia materials hosted by LIT over a period of 18 months approximately), it was still acceptable due to the foreseen outcomes, being mainly the accuracy of results, the novelty in the definition of a web based architecture for annotation, and the usability of the system for non-expert users. Using a semi-automated speech transcription system was out of scope for LIT, since this important step was almost completed during the early phases of the project. Furthermore, a set of syntactic, prosodic and phonological metadata was manually added to transcription, thus making impossible the use of automated software.

The “Collected Requirements for annotations tools”, defined by the ISLE Natural Interactivity and Multimodality Working Group report [16], have been taken into account for usability and functionality evaluations of LIT. The tool was designed according to the points outlined by the Group:

- *Portability:* it can be used on different platforms.
- *Source code:* it is web based and its source code can be open to the research community.
- *Flexible architecture:* adding new components can easily extend it.
- *Three layered structure:* the user interface layer is separated from the logic layer and the data representation layer. Each layer can be changed independently.
- *I/O Flexibility:* the annotation schema is available and the output format is compatible with other tools.
- *Robustness and stability:* the tool has been tested by developers, then extensively used by transcriptionists, it is robust, stable and can be used in real time.
- *Audio/Video Interface:* it provides an easy to use method to view and play audio / video segments. It supports large files and controls of streaming multimedia material.
- *Flexibility in the coding scheme:* the scheme is extendible. The tool allows for custom taxonomies.
- *Easy to use interface:* the interface is intuitive and easy to use. Users are always informed on available actions through proper feedback, thus letting them to understand how it works simply using it. Similarly, the interface implements concepts familiar to targeted users, and provides interactions typical of a desktop application.
- *Learnability:* it is easy to learn and annotations can be performed with very few actions. Learnabilty has been addressed according to several principles: predictability (users should be able to foresee how the system responses to actions), synthesizability (users should be able to understand which actions have led to the current state), consistency

(same or similar components have to look alike and to respond similarly on user input), generalizability (annotations tasks should be grouped together when possible and respond to the same principles) and familiarity (annotation tasks in the web application domain should correspond to annotation activities in the real world)

- *Attractiveness:* it is attractive to the user and uses the modern Rich Internet Applications paradigm. The application has a clean and minimal design, where users can 'play' with it, running and scrubbing videos, displaying frames in which sentences are pronounced and searching their favourite television characters.
- *Transcription support:* it is based on speech transcription.
- *Marking:* it allows the marking of segments of any length and also overlapping or discontinuous fragments.
- *Meta-data:* it supports metadata referring to related annotations.
- *Annotation process:* it supports selection-based annotation with appropriates tags.
- *Visualization:* the annotated information is visible for all annotation elements in real time and in the form of text.
- *Documentation:* there is a user manual.
- *Querying, extraction:* a search engine is integrated with the tool.
- *Data Analysis:* it supports the estimated actual duration of speech.

The system consists of two views:

- an interface to browse the corpus and view the selected videos, along with their transcription and metadata;
- an interface to create and edit video annotations on the transcription source files.

The browsing interface (Fig. 3) shows the video collection present in the model. Users can select a video and play it immediately, and read the associated metadata and speech transcription in sync. Each record in the list of videos provides a link to the raw annotation in XML-TEI format. The annotation can be opened directly inside the browser and saved on the local systems. Subtitles are displayed at the bottom of the video while segments in

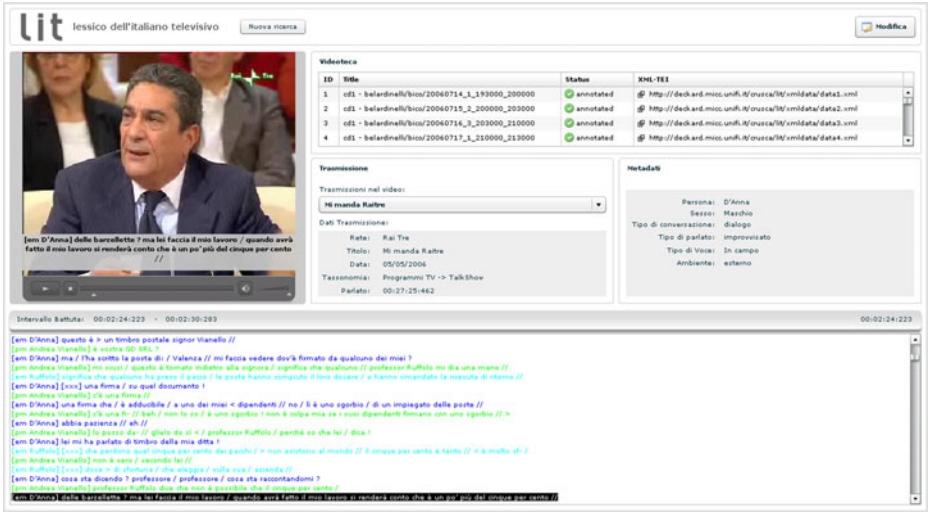

Fig. 3 Browse interface

the transcription area are automatically highlighted during playback and metadata are updated accordingly. When the text-to-speech alignment is completed through annotation activities, users can select a unit of text inside the transcription area and the video cue-point is aligned accordingly; on the contrary, scrolling the trigger on the annotated video segment highlights the corresponding segment of text.

The annotation interface (Fig. 4) is accessed by transcriptionists after authentication, and allows the association between transcriptions and the corresponding video sequences. Annotators can set, using the tools provided by the graphic user interface, the cue points of speech on the video sequences and assign them an annotation without having prior knowledge of the format used. The tool provides functionalities for the definition of metadata at different levels, or multiple "layers": features can be assigned to the document as a whole, to individual transmissions, to speakers in the transmissions and to each single segment of the transcription.

The interaction on the interface is based on the metaphor of the accordion, in order to guide users through all the necessary steps, built on three subsequent panels, each corresponding to a specific layer of the annotation (Transmission, People and Transcription). The goal of using this metaphor is to logically distribute information through different levels and avoid stress that usually emerges when users are exposed to an excessive cognitive load: the accordion switches between different views gradually sliding panels, in order to cover the area displayed by the current view. This gives the user a smooth and clean look and feel of the navigation process.

- The first panel (Transmission) allows annotators to define metadata for the "transmission" layer (title, broadcaster, date, time of screening, and a category from the custom taxonomy tree representation).
- In the second panel (People) users can define metadata for the "people" layer (identity, gender, type of speaker and a colour used to highlight utterances within the transcription area).
- In the third panel (Transcription) the text of the transcription is either pasted or directly typed in, and then utterances and their specifications are assigned to different segments of the video.

Fig. 4 Annotation interface

The utterance metadata, associated to individual segments, is defined selecting a line in the Transcription panel and clicking the button labelled "edit utterance". The action opens a specific annotation panel for the definition of specifications: the identity of the speaker, type of speech (improvisation, programmed, executed), speech technique (on scene, voice-over) and type of communication (monologue, dialogue), start and end points of the segment (in milliseconds).

The annotation panel provides also an additional trigger to control the video playback: users can move through the video sequence with fine grain precision and set a value in decimals to control the size of the step used for moving the cue-point forward and backward; this allows rapid and accurate identification of the segments' key frames uttered by speakers. At this stage the user manages the non-trivial task of text-to-speech alignment, establishing biunivocal relations between units of text and units of speech.

Annotations can be modified or deleted in case of an error; the browsing view is notified and automatically updated after each modification. When modifications are saved, the XML-TEI file of the transmission is automatically generated with the structure described in the previous section, and is ready for the indexing.

Both the annotation and the browse interfaces are developed in Adobe Flex and Actionscript 3. The videos displayed in the system are streamed with RTMP protocol (Real Media Transfer Protocol), using the free developer edition of Adobe Media Server.

2.3 Indexing and search engine

The problem of efficiently indexing and storing a large amount of data in XML can be solved in several ways: the main problem for the LIT project was to find a method that could obtain good performance while processing data, and provide the level of granularity for data access required by the engine specifications.

In a simple environment, developed for an untrained user, most of the search functions implemented would be completely useless and would extremely decrease performances. In the case of LIT, the users targeted for the project set the requirements which, in some cases, are radically different from the kind of functions that a common search engine usually provide. This aspect represents an important novelty in the effort of the scientific community aimed at the definition of standard tools for linguistic analysis. The main features are described in the following list.

- Case sensitiveness. Can be switched on and off by users.
- Accented characters. The presence of characters like "à", "è", "é", "ò", etc. in the Italian language is very frequent, so it is fundamental to make queries that allow defining the sensitiveness to accented characters, in order to determine the evolution of uses of particular forms.
- Frequency analysis of root form expansion. It is of fundamental importance for linguistic research to analyze the frequency of use of some words, and results are usually compared using a fixed root of a word to see how many times each possible expansion of that root is used.
- Jolly characters. Querying with a "?" to specify a generic character and a "*" to specify a generic sequence of characters is very common for most search engines.
- Ordered or unordered word sequence with defined distance. A query for an "exact phrase" search is very common; it is although very rare the option of specifying the distance between each couple of words in a sequence.
- Fine grain specification of the context specified for the single utterances.

The use of XML-native database for the implementation of the functions described, initially appeared to be the best solution for index creation and implementation of the search functions; there were nonetheless some critical issues that made the solution infeasible with this technology such as, among the others, the complexity of creating structured query which had to deal with case sensitiveness and unordered sequences. This led to the definition of an object oriented mapping of the TEI structure in use with its defined custom fields, and the subsequent adoption of an open source engine for storage, indexing and retrieval of the data objects. All the architecture is developed in Java.

The library used for XML-TEI parsing is Apache Digester, which provides serial access to XML files in a way quite similar to the standard Java SAX parser, adding the additional strength of reading elements and attributes from the source file and instantiating them as objects with properties in the Java virtual machine, thus unleashing the power of object oriented programming for data handling. This allows the definition of a mapping between TEI elements and objects as described by the UML class diagram in Fig. 5.

Next step in the implementation is to store the objects and make them accessible for retrieval: in order to maximize performances and implement all the search functionalities, LIT needs more to rely on established and consolidated indexing and querying systems, than to implement a classic data persistence mechanism. The solution adopted is henceforth based on an open source indexing and search engine: Apache Lucene. This engine has an indexing method based on data fragments and stores the documents on the filesystem, while integer offsets are used to refer words contained in them. The indexing algorithms are based on a fundamental data fragment: the document. Each document can be defined independently and can contain a set of different fields. Each field contains the instance of a property defined in the diagram of Fig. 5 and can be used either as a search parameter or as a fragment of content which will be displayed in the list of results after a query has been performed. The fields implemented for each Lucene documents in LIT are 23, described as follows:

1. file full path and name, used for reference;
2. file extension: as the engine could index different types of documents, it is used as a constrain for selecting just the XML-TEI sources in this specific application;
3. last update: can be used as reference to selectively update the index;
4. transcription: the utterance as written in the TEI file, with notes removed;
5. transcription, lower case: same utterance, lower case to perform case insensitive search, since it is more efficient to save an extra field in the database for lower case search, using little extra disk space, rather than performing functions related to case sensitiveness;

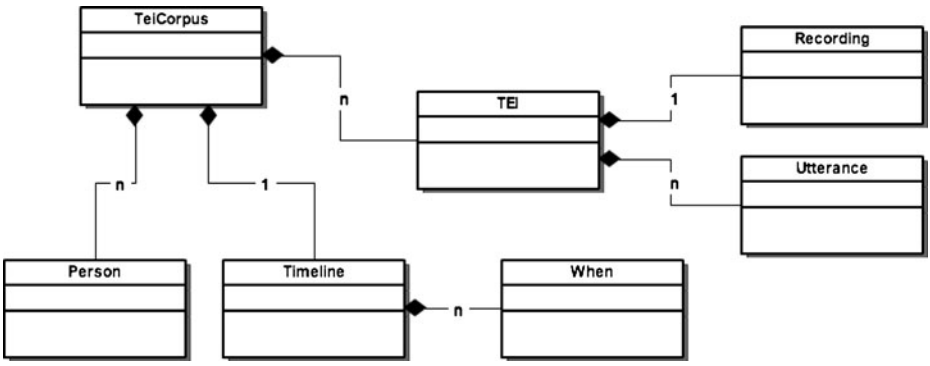

Fig. 5 XML to object mappings

6. transcription, full text with notes: from time to time some notes are included in the utterance by the transcriptionists, and this text must not appear in the statistical data of the corpora;
7. recording title: name of the transmission;
8. recording author: name of the broadcaster;
9. recording category: one of the categories for transmission classification defined in the taxonomy (see section 2.1);
10. recording date, text: date of the recording, in human readable form;
11. recording date, sortable: used for date range queries;
12. recording time: recordings are extracted from afternoon and evening television slots;
13. recording duration: several recordings can appear inside a single extracted hour of broadcast;
14. recording type: can be audio or video, for further development;
15. video ID: refers to the video file handled by the streaming server;
16. time start: beginning of an utterance, in milliseconds;
17. time end: end point of an utterance, in milliseconds;
18. utterance, type of communication: can be either monologue or dialogue;
19. utterance, type of speech: can be improvisation, programmed or executed;
20. utterance, type of speaker: can be either professional or non-professional;
21. person, gender: can be male or female;
22. person, speech technique: can be on scene or voice-over;
23. person ID and name.

When a query is performed, the Lucene API accesses the specified repository and performs the search on the indexes, returning a set of hits ordered by relevance. It should be clear enough at this point that the trickiest task for the application is to construct a meaningful query in order to provide significant results. Lucene provides a query construction kit based on a function called *BooleanQuery*, which allows to programmatically add pieces of query, coming from different search fields of the interface, and compose them logically. Each piece of query is thus added to the complete pool using a set of specific functions, which can better express the search field used (i.e. a *MultiFieldQueryParser* function is used for the "all these words" input field, while a *SpanNearQuery* function is used for the definition of word sequences with ranges). After the boolean query is composed, a set of filtering parameters is added, in order to constrain the query and give results conforming to what was specified in the interface. Thread safe java objects then make data access and searches are based on integer sums and subtractions, thus resulting in a very high performance framework, even when handling large amount of data. Results are sent back to the interface in a proprietary format in order to correctly render data on the frontend.

The search interface (Fig. 6) is based on standard text input fields. It provides a JSP frontend to the search functions defined for the engine and uses the Lucene query syntax for the identification of HTML elements. The interfaces recall a common "advanced search" form, providing all the boolean combinations usually present in search engines and, for this reason, making users comfortable with basic features. Notably, some uncommon features appears among other fields, such as:

- the "free sequence" field, with option for defining it exact, ordered or unordered;
- the "distance" parameter, where free sequences can appear within specified ranges inside a single utterance;
- the date range parameter.

Fig. 6 Search interface

Advanced search features are shown inside dedicated panels (Fig. 7), which can be expanded if necessary. These panels give all the options for specifying the constraints of a query, as defined for the XML-TEI custom fields used in LIT. The extended parameters allow to:

- set the case sensitiveness of a query;
- perform a word root expansion of jolly characters present in the query;
- set the constraint for specific categories defined in the taxonomy;
- select specific parameters for utterances, such as type of speech (improvisation, programmed, executed), speech technique (on scene, voice-over), type of communication (monologue, dialogue), speaker gender and type (professional, non professional).

When a query is executed, results are presented with a header that gives a quick glance at the distribution within the set of categories and broadcaster (Fig. 8).

Fig. 7 Advanced search fields


La ricerca ha prodotto **13** risultati. Dettagli query.

Emittente	Occorrenze
Rai Uno	4
Rai Due	3
Rai Tre	6
Non classificati	0

Categoria	Occorrenze
Altro	4
talkshow	4
Cinema	2
cinema	2
Divulgazione	0
Fiction	2
filmtv	1
serie	1
Informazione	0
Intrattenimento	3
gameshow	3
Pubblicità	0

Fig. 8 Stat summary

The complete list of results is shown just beneath the stats. Each result contains the recording data (title, broadcaster, timing and category) and a preview of the utterance with its corresponding metadata (Fig. 9).

Selecting the title of a recording opens directly the visualization tool, and the referred utterance is automatically marked up (as in Fig. 3), the two modules of the system in fact, search engine and annotation tool, are connected using HTTP GET so that the user can select a text result from the results view provided by the search engine and open the transcription with the corresponding video sequence on the first view of the annotation tool.

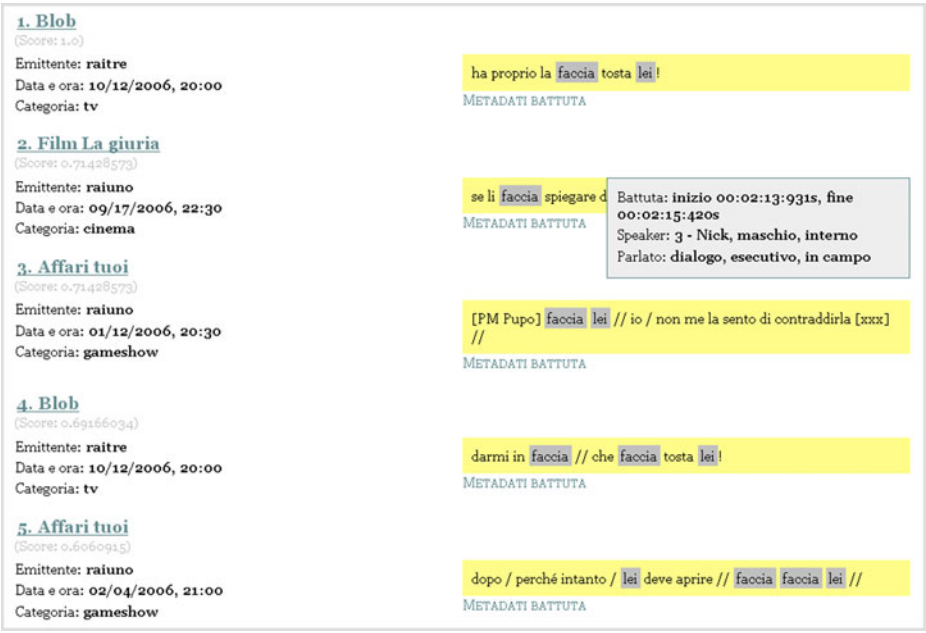

Fig. 9 Utterance metadata

3 Results and developments

This paper shows the results of a collaborative work between computer engineers and social scientists, aimed at the development of a speech transcription and annotation tool for linguistic research. Providing these tools is greatly useful in the field of linguistic computational methodologies, primarily because they allow the systematic management of large amounts of data. It is henceforth essential for such tools to be open to the online community, and to use advanced and extensible data interchange formats. The purpose of LIT is to provide technical means aimed at improving efficiency in annotation, search and analysis of multimedia data to linguistic researchers, as well as accessing, visualizing and sharing materials. The tool provides a specialized domain-oriented architecture that significantly increases the productivity of the researchers and is currently running on MICC servers at the address http://deckard.micc.unifi.it:8080/litsearch/ where the complete corpus of annotated XML-TEI files can be downloaded as well.

The system contains 168 hours of RAI (Radio Televisione Italiana) broadcasts, aired during the year 2006. All the annotations were created by researchers of the *Accademia della Crusca* while LIT was under development, in late 2009. They helped to improve the software reporting bugs and problems, proposing tips and suggestions and contributing proactively to the definition of the initial set of specifications: the application was thus developed in an Agile programming framework in order to allow maximum flexibility and seamless integration of features while the system was already running with limited functionalities, and progressively achieving an easy to use interface, efficient, usable and ergonomic. The system has approximately 20.000 utterances stored and using Lucene for search and retrieval does not raise any performance issue. Optimization of indexes can scale easily up to 2M documents without significant loss of performance, as demonstrated in the testing environment, providing results in few milliseconds even for complex queries. Both the Java application (Tomcat 6) and the streaming server run on a single Ubuntu virtual machine on VMWare ESXi with 2 dedicated 64bit cores and 4GB RAM.

The tool currently supports a content-based annotation but a new release is under development and will be included in the VIVIT project [23]. Upcoming features support additional types of annotation, such as structuring (syntactic and rhetorical structures), signal-spectrographic analysis, automatic suggestion of metadata and semantic structures. In particular the system will be extended to allow term suggestion functionalities for semi-automatic definition of ontologies. Ontologies can provide a framework for semantic classification of concepts, through which instances of each class can be inferred using RDF as a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed. This specification defines the syntax and semantics of the SPARQL query language for RDF. SPARQL can be used to express queries across diverse data sources, whether the data is stored natively as RDF or viewed as RDF via middleware. This approach allows using machine-learning techniques in order to help users while building the concept classification and refine search results by automatically adding related concepts.

Acknowledgments The LIT project was initially funded by CLIEO, the "Center for theoretical and historical Linguistics: Italian, European and Oriental languages", in collaboration with the Accademia della Crusca, the leading research institution on the Italian language, and we owe a debt of gratitude to Prof. Nicoletta Maraschio, who allowed to kick off this research. Most of the computer engineering work done at

the Media Integration and Communication Center had a continuous feedback from researchers of the Accademia and would not have been possible with the precious support of Marco Biffi and Vera Gheno. Luckily enough, the initial financial support of CLIEO has been extended for a 3 years project funded by the Italian Ministry of Education, University and Research, which will allow integrating semantic web functionalities.

Besmir Bregasi, Ervin Hoxha and Tiberio Uricchio are three M.Eng. students who merely knew how things could get complicated when projects are delivered by the Media Integration and Communication Center, but they had the chance to learn a lot while working on their B.Eng. dissertation assignment. Freshness of young minds is always a plus when efforts can be focused and welded with senior experience.

References

1. A complete list is available at http://annotation.exmaralda.org/index.php/Tools
2. Agile software development refers to a group of software development methodologies based on iterative development, where requirements and solutions evolve through collaboration between self-organizing cross-functional teams. http://agilemanifesto.org/
3. Amaral R, Meinedo H, Caseiro D, Trancoso I, Neto JP (2006) Automatic vs. Manual Topic Segmentation and Indexation in Broadcast News. In: Proc. of the IV Jornadas en Tecnologia del Habla
4. Anvil: http://www.anvil-software.de/ for more details
5. at the moment of this writing, definitions of TEI standard DTD are those of rev. 5, released Nov. 2007: http://www.tei-c.org/Guidelines/P5/
6. Bender EM, Langendoen DT (2010) `Computational linguistics in support of Linguistic Theory. Linguistic Issues in Language Technology 3(2):1–31
7. Bertini M, Cucchiara R, Del Bimbo A, Grana C, Serra G, Torniai C, Vezzani R (2009) Dynamic pictorially enriched ontologies for video digital libraries. IEEE Multimedia (MMUL)
8. bits of memory for the ATLAS project definition: http://xml.coverpages.org/atlasAnnotation.html and official website of the Multimodal Information Group, the research board which is currently hosting parts of the project: http://www.nist.gov/itl/iad/mig/
9. Dybkjaer L, Berman S, Kipp M, Olsen MW, Pirelli V, Reithinger N, Soria C (2001) Survey of existing tools, standards and user needs for annotation of natural interaction and multimodal data. Technical report, January
10. Garrett J (2002) Elements of user experience: user-centered design for the web. New Riders Press, USA
11. Hauptmann AG, Jin R, Ng TD (2003) Video retrieval using speech and image information. In Storage and retrieval for multimedia databases 2003. EI'03 Electronic Imaging, pp 148–159
12. Helena Moniz, Fernando Batista, Hugo Meinedo, Alberto Abad, Isabel Trancoso, Ana Isabel Mata da Silva, Nuno J (2010) Mamede, Prosodically-based automatic segmentation and punctuation, In Speech Prosody 2010, ISCA, Chicago, USA, May
13. ILSP: http://www.ilsp.gr for more details
14. in terms of diffusion, Flash player is supported by over 95% of market share: http://www.statowl.com/custom_ria_market_penetration.php
15. it is fairly difficult to find a complete and updated reference of annotation tools and applications. The following, dated 2001, even if a little bit old, is almost complete and can be useful as a starting point: http://www.ldc.upenn.edu/annotation/
16. Kristoffersen S (2008) Learnability and robustness of user interfaces: towards a formal analysis of usability design principles. In: Proceedings of the ICSOFT 2008: 3rd International Conference on Software and Data Technologies. vol. SE/GSDCA/MUSE: Institute for Systems and Technologies of Information, Control and Communication, pp 261–268
17. Kuniavsky M (2003) Observing the user experience—A practitioner's guide to user research. Morgan Kaufmann Publishers, Elsevier Science, USA
18. NITE: http://www.dfki.de/nite/ for more details
19. Praat: http://www.fon.hum.uva.nl/praat/ for more details
20. RDF standard official website: http://www.w3.org/RDF/ and SPARQL recommendation for RDF queries: http://www.w3.org/TR/rdf-sparql-query/
21. SOAP protocol standard definition at W3C: http://www.w3.org/TR/soap/ and definition of REST web services: http://en.wikipedia.org/wiki/Representational_State_Transfer

22. the MPEG-7 standard definition and overview: http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm
23. The VIVIT (Vivi l'italiano) project is funded by the FIRB, the Investment Fund for Basic Research of the Italian Ministry of University and Research and has the objective of creating an integrated digital archive of teaching materials, texts and iconographic documents and media for knowledge dissemination of Italian language and cultural history
24. Transcriber: http://trans.sourceforge.net/en/presentation.php for more details

Thomas M. Alisi is self employed in the digital media industry. He worked for a long time as project manager and researcher at the Media Integration and Communication Center, where he took also a Ph.D. in Computer Engineering, Multimedia and TLC, and became skilled in semantic web and interactive environments. Then he discovered that the time had come to speak of many other things, but he still collaborates with the University, where he occasionally manages to write a paper and a project proposal.

Alberto Del Bimbo is Professor of Computer Engineering at the Faculty of Engineering, University of Firenze. His research interests address analysis and interpretation of images and video and their applications, with particular interest at content-based retrieval in visual and multimedia digital archives, advanced videosurveillance and target tracking, and natural man-machine interaction assisted by computer vision.

Andrea Ferracani is working as researcher in the Visual Information and Media Lab at the Media Integration and Communication Centre, University of Florence. His research interests focus on web applications, collective intelligence and semantic web. He is principal lecturer for "Web languages and programming" at the Master in Multimedia Content Design and "Editoria Multimediale" at the Faculty of Architecture of the University of Florence. When the evening comes, he plays classic guitar, perfectly aware of the fact that mastering an instrument is an endless challenge.