DEVELOPMENT AND EVALUATION OF AN ITALIAN BROADCAST NEWS CORPUS

Marcello Federico, Dimitri Giordani, Paolo Coletti

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica
I-38050 Povo, Trento, Italy.

Abstract

This paper reports on the development and evaluation of an Italian broadcast news corpus at ITC-irst, under a contract with the European Language resources Distribution Agency (ELDA). The corpus consists of 30 hours of recordings transcribed and annotated with conventions similar to those adopted by the Linguistic Data Consortium for the DARPA HUB-4 corpora. The corpus will be completed and released to ELDA by April 2000.

1. INTRODUCTION

Broadcasting is rapidly moving toward digital standards and major broadcasting companies have been starting to digitize their archives. The availability of large multimedia digital libraries will increase the demand of services from content providers and the public. This will give rise to the need for technologies that make the management and access of multimedia archives easier.

In order to develop technology for the processing of audio archives (Brugnara et al., 2000), ITC-irst has started in Summer 1999, under a contract with the European Language resources Distribution Agency (ELDA), a project for the collection and packaging of an annotated speech corpus of broadcast news in the Italian language. The language resource consists of about 30 hours of radio news covering several years. Speech recordings present variations of topic, speaker, acoustic channel, speaking mode, etc. The corpus, called Italian Broadcast News Corpus (IBNC), has been segmented, labelled and transcribed with conventions similar to those adopted by LDC for the DARPA HUB-4 corpora. [1] The IBNC, which is at present in a final stage of development, will be released to ELDA by April 2000.

In Section 1, the initial requirements of the corpus are stated and the specifications used for the transcription and annotation are described. In Section 2, the corpus development activities are presented: data collection, transcription and annotation, and post-processing. Finally, in Section 3 issues concerning the corpus evaluation are addressed and preliminary results are provided.

2. REQUIREMENTS AND SPECIFICATIONS

The main objective of the project is the transcription and annotation of a collection of audio files. The purpose of transcription is the exact reproduction, under form of text, of the verbal and non verbal sounds of a speech recording. The purpose of annotation is to associate certain signal, speaker, and content conditions with the speech and its transcription. This is done using an XML formalism. The whole transcription and annotation task is performed manually by employing a specific software tool, called Transcriber (Barras et al., 1998). In the following, the annotation structure and the transcription conventions are described.

2.1. Annotation Structure

The annotation is organized in the following hierarchical elements: episode, section, turn, and segment.

Episode An episode identifies the recording of a particular broadcast of a program at a certain date and time.

Section A section denotes a particular portion of signal within an episode. Sections divide an episode into untranscribed portions, fillers (introductions, credits, etc.) and reports (single stories) which are identified by specific topics.

Turns A turn denotes a portion within an episode containing speech of a single speaker.

Segment A segment denotes a small portion of a turn, usually not longer than 10 seconds, that contains speech delimited by breaths. Segments are useful for the sake of the transcription itself.

Turns can be characterized by the following features: speaker, mode, channel, fidelity, and background (cf. Table 1).

Speaker Each speaker of an episode is identified by a name. If possible, speakers are identified by their name and surname. Otherwise, speakers are named either as `reporter_nn` or as generic `speaker_nn` where `nn` stands for a progressive number. Moreover, the speaker's gender, accent, and dialect are indicated.

Mode The speaking mode may assume one of two values: `spontaneous` or `planned`. In contrast to spontaneous speech, planned speech does not present much disfluencies, such as restarts and hesitations.

Channel The channel feature aims at identifying the transmission medium. It may assume one of two values: `telephone` or `studio`. By convention, the `telephone` value is also used to denote audio channels with a limited bandwidth, e.g. dictaphones.

Fidelity Fidelity aims at grading the recording and transmission quality. Hence, regardless of the channel, fidelity may assume one of three values: `high`, `medium`, or `low`.

[1] See Internet site www.ldc.upenn.edu.

Another factor that can influence the acoustic quality of turns, or even smaller portions of signal, is the **background**. Background serves to identify noises produced by the environment, namely speech, music, or other. Background noise can be indicated by selecting and labelling portions of the audio signal. As a general rule, background noises are indicated only if they are significant, i.e. persistent and easy audible.

2.2. Transcription Conventions

Transcription is only applied to segments containing speech. The transcription text consists of mixed-case ASCII characters of the ISO-8859-1 extended set. Only alphabetic characters and punctuation marks are used, along with bracketing characters. Punctuation is used only to improve text readability. Explicitly uttered punctuation marks are transcribed in words. Uppercase characters are only used in proper names. Line breaks do not belong to the transcription, but are used to separate XML tags indicating annotation elements.

Proper names Proper names are transcribed with upper-case initials, acronyms are transcribed as upper-case words, and numbers are transcribed as words (e.g. novantanove instead of 99). Abbreviations are not used. However, words that are spoken as abbreviated (e.g. "Aut. Min." rather than "Autorizzazione Ministeriale") are spelled that way. When said as a sequence of letters, acronyms are properly tagged. Moreover, spelled out letters and name initials are transcribed as upper-case words.

Accents and apostrophes Accented characters are used only for final vowels. Apostrophes are used for initial or final elisions. Words with apostrophes are isolated as well as all the other words.

Special Bracketing Conventions Bracketing conventions are used according to the Transcriber's native XML format. [2] Specific tags are used to annotate local or extended events. For each event type, the range of possible descriptions is now given:

- Noise: `respiro` (breath), `tosse` (cough), `labbra` (lips), `starnuto` (sneeze), `microf` (microphone), `risata` (laugh), `fruscio` (rustle), `parlato` (conversation), `rumore` (other noise).
- Pronounce: `errore` (error), `nonverb` (non verbal sound), `ride` (laugh), `roca` (hoarse), `sigla` (spelling).
- Lexical: `dubbia` (uncertain transcription), `scono` (unknown word), `neol` (neologism), `()` (broken syntax).
- Language: foreign language words are tagged with a language identifier: e.g. `ar` stands for Arabic. In general, foreign words are indicated when uttered within an Italian spoken turn. Foreign language turns are not transcribed, but the idiom, if known, is annotated by a language event.

Feature	Values
Speaker	$< name - surname >$
Gender	$female/male/unknown$
Accent	$< description >$
Dialect	$native/nonnantive$
Mode	$planned/spontaneous$
Fidelity	$high/medium/low$
Channel	$studio/telephone$
Background	$clean/music/speech/other$

Table 1: Features available for each speech turn.

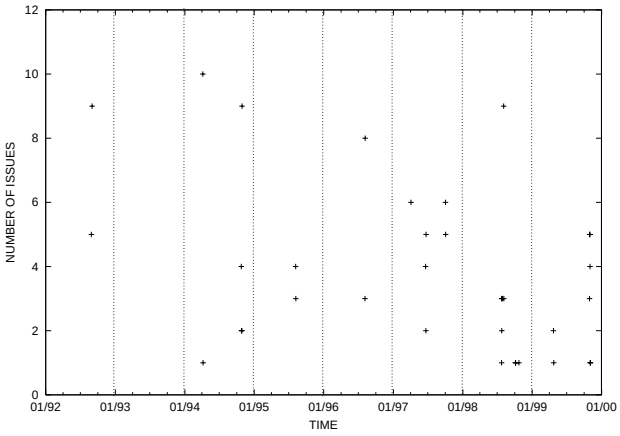

Figure 1: Distribution of news program issues over time.

3. DEVELOPMENT

3.1. Data Collection

RAI, the major Italian broadcast company, supplied studio quality recordings of radio news programs sampled from its internal digital archive. The sampling strategy was the following. Starting from year 1992 until 1999, some days per year were chosen and several news recordings were taken of those days and possibly of one or two days after. For some reason, no recordings of year 1993 were taken. The distribution of news programs against each day of the considered time interval is shown in Figure 1. The collection consists of 150 programs, for a total time of about 30 hours, issued in 36 different day, between 1992 and 1999.
Recordings were supplied by RAI on Digital Audio Tapes (DAT), with 44kHz sampling rate and 16 bit resolution. Each DAT was manually processed to transfer each single program issue into a single file. During this operation, the signal was down-sampled to 16kHz with a resolution of 16 bits, and encoded into the NIST Sphere PCM format.

3.2. Transcription and Annotation

The transcription and annotation work involved five people with different tasks: two transcribers, two intermediate supervisors, and one final supervisor. Before starting, an annotation manual was written and people were trained on few audio samples. As many new issues concerning the annotation arose during the first weeks of work, a Web site was set up containing the updated manual and useful resources to solve the spelling of foreign proper names.

[2] See Internet site www.ldc.upenn.edu/mirror/Transcriber.

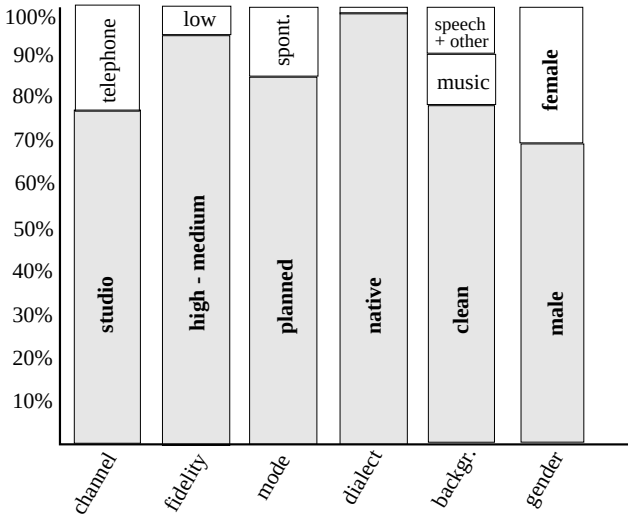

Figure 2: Statistics about six recording conditions.

The work was organized into the following steps:

1. signal segmentation
2. transcription and annotation
3. transcription alignment check
4. annotation structure check (I)
5. annotation structure check (II)
6. lexical check

Steps 1-2 were performed by the transcribers. Steps 3-4 by the intermediate supervisors. Steps 5-6 by the final supervisor. Hence, each transcript was manually checked by at least three different people.

3.3. Post-processing

A post processing on the whole transcripts was finally performed to improve the annotation accuracy. This work was carried out manually by using automatic tools.
Topics of sections and speaker labels were revised to uniform their layout. Hence, consistency of speech-to-speaker assignments was also manually verified. A tool was used that extracts, for a given speaker label, a speech sample of it from each episode of the whole collection.
Another issue was the attempt to identify some of the speakers labelled as `reporter`. Two methods were combined. First a speaker recognition algorithm was trained on the most frequent known speakers and then applied to the speech segments labelled as `reporter`. The algorithm provides potential speaker names ranked according to a matching score. Identification was then attempted by manually comparing speech samples of each unknown speaker with those of the known speakers. The automatically generated rank list helped to focus the manual search. In this way, about 50% of the unknown `reporter` speakers were identified.
Finally, the consistency of speech-to-channel assignments was verified. A Gaussian Mixture Model (GMM) classifier was trained (Cettolo, 2000) and let run on the whole

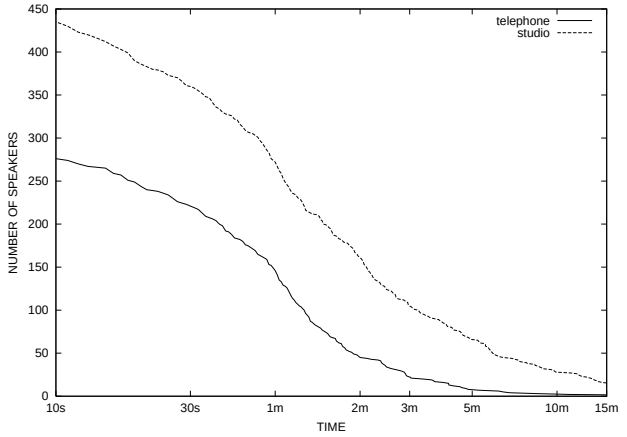

Figure 3: Number of speakers for which a given amount of speech seconds is at least available.

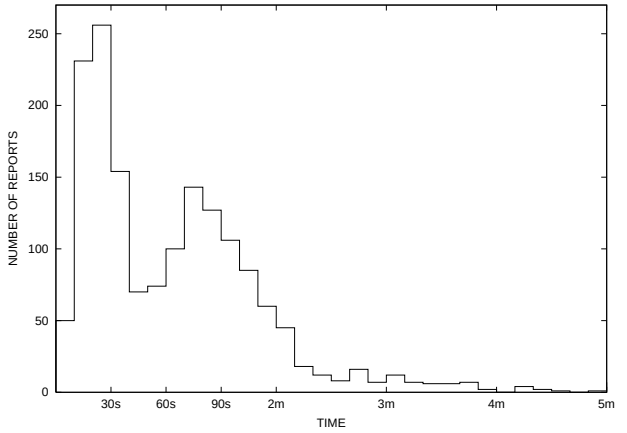

Figure 4: Duration distribution of the `report` sections.

database. The classifier was forced to classify speech segments either as telephone or studio. The frame error rate of the GMM classifier on a manually verified test set is 0.4%. The frame by frame mismatch between the manual and automatic channel classification was 1.1%. Speech segments with suspect labels were manually checked by inspecting the time-frequency spectrogram.

4. EVALUATION

Evaluation of the IBNC focuses on two issues: content richness and annotation accuracy. While there are definite results about the first issue, only preliminary results are available about the accuracy of the annotation. As a matter of fact, measures about the accuracy are still used to correct annotation errors within the corpus.

4.1. Content Analysis

A desirable feature of the corpus is to be rich in terms of acoustic and linguistic content. Hence, statistics related with these concepts have been extracted and analysed.
Six independent features that somehow describe the acoustic content of the corpus have been measured: channel type, recording fidelity, speaking mode, background noise, and speaker gender. Figure 2 reports the IBNC composition of each feature. It results that 76.7% of material contains

studio quality speech (23.3% contains telephone speech), 95.0% of the material is of high/medium fidelity, 85.0% contains planned speech, 78.9% is free of background noise, 70.4% is uttered by male speakers, and 98.9% by native speakers. By considering all the features together, the IBNC presents conditions of clean, studio-quality, and planned speech, about 57.0% of the time. However, the corpus provides significative samples of other conditions: e.g. about 7h of telephone speech and 4h:30' of spontaneous speech.

Another important issue concerning the acoustic content could be the number of speakers represented in the corpus. The total number of identified speakers is 677, which cover about 96.5% of the total speech time. The remaining speech is classified either as `reporter` (2.5%) or as generic `speaker` (1.5%). With respect to the audio channel, there are 451 identified speakers with studio speech and 298 with telephone speech. The number of speakers which are recorded in both conditions is 72. The number of speakers for which a given amount of speech is at least available is plotted in Figure 3. It results that there are at least 146 telephone speakers and 272 studio speakers for which at least one minute of speech is available.

Other statistics on the transcribed texts were computed to evaluate the linguistic content of the corpus. As already mentioned, IBNC contains 150 news programs spanning over the period 1992-1999. The news programs globally contain 1618 sections classified `report`, and 402 sections `filler`. The distribution of the lengths of the report sections is shown in Figure 4. The two peaks in the distribution are probably due to the presence of short and long program issues in the collection.

The transcripts contain 318K words, for a vocabulary size of 23K words. An interesting curve that describes the language variability is the vocabulary growth function with respect to increasing amounts of texts. Figure 5 shows the vocabulary growth function of the IBNC and other three corpora: an excerpt of the Italian newspaper corpus *La Stampa*, the English spoken inquiry corpus ATIS, and the sub-corpus IBNC II that considers just one program issue (the longest) for each available day. In particular, IBNC II contains 36 issues and a total of 138K words. It results that the vocabulary growth of IBNC is significantly smaller that that of the newspaper corpus. The explanation could be twofold: the kind of language used by the different media and the redundancy of news in the IBNC, caused by the used sampling strategy (see Figure 1). However, the relatively small difference between the curves of IBNC and IBNC II, which is much less redundant, suggests that most of the gap is due to the different languages of radio and newspapers.

4.2. Annotation Accuracy

Some preliminary evaluation of the transcription and annotation accuracy has been carried out. The goal is to estimate statistically reliable upper bounds of different kind of annotation errors.

A panel of evaluators was used to spot some transcription and annotation errors within a set of 125 segments, randomly extracted from the transcripts. Each expert received

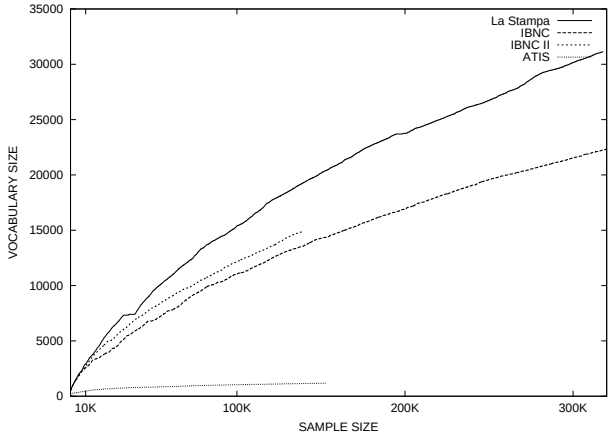

Figure 5: Vocabulary growth of IBNC and other three corpora: an Italian newspaper (*La Stampa*), English spoken queries (*ATIS*), and a sub-corpus if IBNC containing just one edition per day (*IBNC II*).

written instructions about the kind of errors to find out: annotation errors of acoustic condition (cf. Table 1), transcription errors of words, pronunciation errors of words, and annotation errors of non-verbal sounds.

Instructions specified to only consider "relevant" phenomena, i.e. easy audible. This requirement, which is of course subject to personal interpretations, was introduced in conformity with the annotation manual.

The results of the test were used to compute accuracy estimation on the whole collection with a 95% confidence level. Concerning the transcription error rate on words the estimate error was below 0.5%. The pronunciation error of words was below 1%. The annotation error of non verbal phenomena was below 1%. The annotation error on acoustic conditions of segments was below 12%. [3]

These estimates refer to a first evaluation session, aimed at finding out major problems to be addressed in the post-processing step, described before. Finally, it is expected that the final accuracy estimates will be higher than those reported here.

5. Acknowledgements

The authors thank Gaia Dondana, Barbara Mezzabotta, Mauro Cettolo, and Fabio Brugnara for their contribution to the development of the IBNC corpus.

6. References

Barras, C., Edouard G., Zhibiao, W., and Liberman, M., 1998. Transcriber: a free tool for segmenting, labeling and transcribing speech. In *Proc. of LREC*. Granada, Spain.

Brugnara, F., Cettolo, M., Federico, M., and Giuliani, D., 2000. A system for the segmentation and transcription of Italian radio news. In *Proc. of RIAO*. Paris, France.

Cettolo, M., 2000. Segmentation, classification and clustering of an Italian broadcast news corpus. In *Proc. of RIAO*. Paris, France.

[3] The largest part of errors was found in the fidelity condition.