Specification of validation criteria

Validation criteria for Arabic Broadcast News Speech Corpus

Niklas Paulsson

S. Haamid

M. Shaheen

2005

The NEMLAR project is supported by the INCO-MED programme

http://www.nemlar.org

Contents

The NEMLAR project is supported by the INCO-MED programme

http://www.nemlar.org

1 Executive summary

This document contains the specifications of the validation criteria for the Arabic Broadcast News Speech Corpus within the Nemlar project. This document gives an overview of the aspects of the database which are validated, such as, e.g. documentation, completeness of database items or acoustic quality of the speech signal. It presents a set of tolerance margins, otherwise called validation criteria, which are employed to accept or reject a database. It also outlines the validation procedure which is done in a number of stages.

2 Introduction

The aim of this document is to specify validation criteria that Nemlar databases should fulfill and give an overview of the aspects of the databases that are checked in the process of validation. This document lists the criteria against which the databases are checked and which are employed to accept or reject a database.

The principles of validation and the criteria listed here have evolved over a number of previous BNSC projects. Therefore, a number of documents have been used as a reference here ([1], [2], [3]). This means that the principles and the validation criteria have been extensively tested before. Arabic languages we are dealing with in Nemlar, however, pose new challenges that have to be accounted for during validation as well.

Apart from very specific validation criteria (like the allowed number of missing files) the databases should also fulfill a lot of other requirements that immediately follow from the specifications of the databases. These specifications are related to the database format and structure, to the transcription conventions, speaker demographics, environmental conditions, and the lexicon contents. A summary of or a reference to these specifications is contained in the present document, as their fulfillment is of immediate importance for the acceptability of a database.

The following aspects of the validation criteria are addressed in the sections below:

1. Documentation.
2. Database structure, file names and contents.
3. Database items and completeness.
4. Acoustic quality of the speech files.
5. Annotation files.
6. Lexicon.
7. Database design and collection.
8. Recording conditions.
9. Transcription quality.

3 Documentation

Each database must be accompanied by a DESIGN.DOC which is written in English and includes the following information:

- contact person: name, address, affiliation;
- distribution media

- number of disks;
- contents of each disk;
- layout of the disk file system;

- formats of speech and label files;
- file nomenclature and directory structure
- reference to the validation report VALREP.DOC
- source selection
- database design
  - hardware
  - software
  - recording platform
  - recording procedure
- database contents
  - topics
  - radio station
- description of the different recording environments and their distribution in the database
- speaker information:
  - a complete and comprehensive rationale of the regional pronunciation variants that are distinguished;
  - speaker name;
  - number of speakers;
- annotation information:
  - procedure used;
  - quality assurance;
  - a list of non-standard and alternative spellings (or reference to file SPELLALT.DOC);
  - standard character set used for transcription (ISO-8859<n> or other if needed)
  - transcription conventions such as abbreviations, proper name conventions, contractions (July or july, isn't, cannot or can not, etc.);
  - annotation symbols for non-speech acoustic events including the standard defined;
  - markers for mispronunciations, recording truncations, unintelligible speech;
  - language specific conventions;
- lexicon information:
  - procedures to obtain phonemic forms from orthographic input;
  - list of SAMPA phone symbols;
  - list of orthographic symbols for languages not using Latin alphabet;

- whether or not the transcription and the lexicon are case-sensitive;
- whether multiple transcriptions are supported;
- whether frequency information is provided;
- whether stress information is supplied;
- whether there are any tags, and if so, the tagging conventions used, e.g., record (noun) vs. record (verb);
- list of words that are from a foreign language;
- tables with frequency of occurrence of phones represented in the phonetically rich material (combined words and sentences) and in the full database (at transcription level);
- list of rare phonemes;
- any other language-dependent information or conventions;

- indication of quality assurance: estimated percentage of files that were double-checked by the producer;
- ratio in minutes: speech / music;
- any other information useful to characterize the database.

4 Database Structure, File Names and Contents

4.1 File names for label files and speech files and directory names

The databases should comply with the following directory structure:

\<database>\< data>

Where:

<database>	defined as: <DBname><#><language code>. where <DBname> is NMBCN <#> is 7 for Nemlar <language code> is a 2-letter language code, AR for Arabic.
<data>	Defined as DATA

Table 1: Directory structure

Both signal files and label files have to be put in the terminal node subdirectories.
In addition to the previous structures the following directories are used to store the other (non-speech data) files:

\ *(root)*	README.TXT file containing a short description of the database and the files, DISK.ID file and COPYRIGH.TXT
\<database>\DOC	documentation
\<database>\TABLE	speaker, session, recording condition, overlap and lexicon tables

Table 2: Directory structure for non-speech data files

The filenames should correspond to the following template:

DD_YYMMDD_SSS_LL.<ext>

where:

DD	Database identification code (00-ZZ) For NEMLAR: N7
YYMMDD	Year, month, day of recording
SSS	Source of recording (three characters, e.g. 'RTM')
LL	Two letter ISO 639 language code, e.g. 'AR'
<ext>	File type code, i.e. .WAV = speech file .SAM = SAM annotation file .TRS = transcription in XML format

Table 1.3 - Filename convention

4.2 The DOC directory

The following files should be in \<database_name>\DOC:

- DESIGN.DOC
- PLATFORM.DOC
- TRANSCRP.DOC (optional)
- SPELLALT.DOC (optional)
- SAMPALEX.PS
- ISO8859<n>.PS
- SUMMARY.TXT
- SAMPSTAT.TXT
- VALREP.DOC
- TRANS-13.DTD

The validation of the DESIGN.DOC main documentation file is described in section 3. PLATFORM.DOC contains platform specifications. TRANSCRP.DOC contains transcription instructions to the transcribers (in the native language and/or in English). ISO8859<n>.PS is a postscript file containing the ISO-8859-<n> character table used for orthographic transcription. The SAMPALEX file lists the SAMPA symbols used for the phonemic transcriptions in the lexicon together with an example. SUMMARY.TXT contains an overview of all items recorded for each session. SAMPSTAT.TXT is the output of the acoustical check on the speech files performed by each partner. The file VALREP.DOC which contains the validation report is created by the validation centre.

4.3 The TABLE directory

Tables should be in \<database>\TABLE

- LEXICON.TBL
- SPEAKER.TBL
- SESSION.TBL

4.4 Other directories

The root directory should contain the files:

- README.TXT: ASCII text file containing a description of the files in the database

- README.HTM: with browser access to all documentation directories (optional)
- COPYRIGH.TXT: copyright statement in ASCII
- DISK.ID: 11-character string with volume name

4.5 Other requirements

All text files should have <CR><LF> at line ends. This concerns all label files, all table (.TBL) files, all index (.LST) files, and all (.TXT) files.
All table files and index files (but *not* SUMMARY.TXT) should report the field names collected in each record as the first row (header) of the file. In this header tabs should be used to separate the fields just like in the rest of the file.
Empty files are illegal. This is of special relevance for speech and label files.
For each label file there must be one corresponding speech file and vice versa.
Obviously the database should not be infected by any viruses.

5 Database Items and Completeness

5.1 Mandatory items specifications

It will be checked if all mandatory items are recorded:

- Topics
- Segmentation
- Speaker information
- Audio source information

5.2 Validation of missing items

For each database it will be checked if all mandatory items are present in sufficient quantities.

6 Acoustic Quality of the Speech Files

All speech files should be recorded with 16 kHz and 16 bit PCM WAV format.

7 Annotation Files

Checks will be performed to make sure that:

- Correct labels and correct accompanying values are used
- There are no empty label files
- Each line is delimited by <CR><LF> (DOS format)

8 Lexicon

8.1 Format checks

For the lexicon table the following checks are carried out:

- Format check
- All and only SAMPA phoneme symbols are used

- The lexicon contains all words in the transcriptions except distorted words (i.e., mispronounced or truncated words)
- If tagging is supplied, check that all tag symbols are defined and only those symbols are used

The lexicon should be complete. The completeness check is carried out on orthographic transcriptions in the label files in order to find out if all the transcribed words are in the lexicon. Under completeness is not permitted, over completeness is.

8.2 Validation of phonemic transcriptions

1000 lexicon entries will be checked for phonetic correctness by native speaker phoneticians that were not involved in the original transcription process, or by comparing with other available pronunciation lexicons.
The validation of the phonemic correctness of the lexicon entries is organized as follows:

- 1000 entries are randomly extracted from the lexicon;
- Only the first phonemic transcriptions is kept in case of multiple transcriptions;
- The check is carried out at the segmental level only (not at syllable boundaries or stress marks, if provided)
- The check is carried out by a phonetically trained person who is a native speaker of the language
- The given transcription receives the benefit of the doubt
- The given transcription is correct if it represents a possible pronunciation of the word (which is not necessarily the most common)
- Each transcription is rated on a 3-point scale: OK; Minor error; Major error
- A max. of 10% minor errors are allowed; and a max. of 5% major errors are allowed
- A minor error occurs if only one symbol in the transcription is wrong
- A major error occurs if more than one symbol is wrong

Since only a sample of 1000 entries is evaluated, the detected errors give the following confidence intervals when extrapolated to the entire DB.

Error percentage	**Confidence interval**
5%	3.6% - 6.4%
10%	8.1% – 11.9%

Table 5: Confidence intervals

9 Database design and collection

The database should follow the general rules for Broadcast News Speech Corpus:

- max 1 recording per day
- recorded by means of a radio receiver
- a scheme for selecting topics

10 Recording Conditions

Selection of news, spacing in time, no 2 recordings from same day, recorded from a radio source, Arabic news, news content to be defined.

11 Transcription

For transcription validation 30 minutes are selected randomly and their transcription is checked manually. Transcription validation of speech is carried out by a trained native speaker of the language concerned, who did not participate in the original transcription process. The transcription validation of the non-speech symbols is not necessarily done by a native speaker of the language, but by someone experienced in listening to background noises and capable of deciding which noises should be transcribed or not. The transcriptions in the label files are checked by listening to the corresponding speech files and by correcting the transcriptions if necessary. As a general rule it is maintained that the delivered transcription should always receive the benefit of the doubt and that only overt errors should be corrected.

For further details concerning the transcriptions, see “nemlar_transguide_v1.2.doc”.

11.1 Annotation levels

The following annotation levels are considered of primary interest and should be included:

- Orthographic transcription of speech (in news, not in music, commercials etc.), including Named Entities
- Speakers and speaker turns
- Segment markers (portions of max. 10 s.)
- Topic/story boundaries
- Background noises (stationary and instantaneous noise events)
- Music / Noise

Other annotation levels are considered of secondary interest and may be added later:

- Change of background
- Speaking mode (i.e. “spontaneous” vs “planned/read”)
- Word boundaries

The structure of the annotation is based on the following hierarchical elements: episode, section, turn, and segment.

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

Section - A Section denotes a particular portion and/or story 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 (see also section 2.4).

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.

11.2 Speaker information

A list of speakers and attributes is shown at the beginning of each transcription file. Speakers are identified by name and surname, when possible, or by a progressive code of one of the categories reporter and speaker. The list of speakers is tagged with the <speakers> spanning tag, while each entry is marked with the non-spanning <speaker/> tag. Here is a definition of the attributes of the latter tag.

- **id:** The speakers code that is used in the successive turns.
- **name:** The speaker's name. If a speaker cannot be identified, then it is classified either as reporter or as speaker. In that case, a numeric code is appended to the name.
- **check:** One of the values "yes" or "not". The speaker's name has been checked verified.
- **type:** One of the values "male" or "female" or "unknown".
- **dialect:** One of the values "native" or "non native".
- **accent:** The accent of speakers with a prominent accent. In general, the accent is indicated by specifying by a geographic area.
- **scope:** One of the values "local" or "global". Speakers which have not been identified have a local scope, while speaker's with name and surname have a global scope.

11.3 Transcription conventions

The following validation criteria are applied to orthographic transcriptions:

- Transliterations are case-sensitive unless specified otherwise in the documentation
- Punctuation marks should not be used in the transliterations
- Digits and numbers must appear in full orthographic form
- Character set and line formatting
- Names, Numbers, Acronyms and Abbreviations
- Accents and apostrophes
- Special Bracketing Conventions
- Noise markers
- Pronunciation markers
- Lexical markers
- Language markers
- Event markers

These criteria are checked both automatically on the *full* database, and by the native speaker on the *subset* for transcription validation.

11.4 Type of errors

Two types of errors are distinguished:

1. Errors in the transcription of speech
2. Errors in the transcription of non-speech (background noises)

Errors in the transcription of truncations, mispronunciations, word fragments and unintelligible fragments are counted as errors in the transcription of speech. Only errors in the transcription of non-speech acoustic events are counted as non-speech transcription errors.

11.5 Criteria for validation

The main criteria for the validation of the transcriptions by the expert are:

- For speech a maximum of 5% of the validated utterances (=files) may contain a transcription error.
- For non-speech a maximum of 20% of the validated utterances (=files) may contain a transcription error.

Only erroneous omissions of noise symbols are considered errors in non-speech.
All non-speech symbols are mapped onto one during validation, i.e. if a non-speech symbol was at the proper location then it is validated as correct, regardless if it is the *correct* non-speech symbol or not. Only stationary noise may not be confused with another type of noise.
The error percentage is determined at item level, not at word level.

11.6 Statistical reliability

A random sample of 1000 utterances is checked for the complete database.
For each set of 1000 items the (95%) confidence intervals for varying error percentages are:

Error percentage	**Confidence interval**
5%	3.6% - 6.4%
10%	8.1% – 11.9%
50%	46.9% - 53.1%
95%	93.6% - 96.4%

Table 8: Confidence intervals for 1000 items

11.7 Spelling check

A formal spelling check of the orthographic transcriptions will not be carried out by the validation centre. It is recommended that partners report the results of a spelling check that they carried out themselves in the documentation of the database.

12 Validation procedures

A database is validated in at least three stages: pre-validation, validation and pre-release validation.

13 References

[1] van den Heuvel, H.*: Validation criteria*. Orientel. Technical Report D6.2. Version 1.2, 2002.