Validation of language resources in TC-STAR

Henk van den Heuvel, Eric Sanders

SPEX/CLST Radboud University Nijmegen, Netherlands

CLST, Radboud University, Erasmusplein 1, 6525 HT Nijmegen, Netherlands
E-mail: H.vandenHeuvel@let.ru.nl

Abstract

In TC-STAR a variety of Language Resources (LR) are being produced. In this contribution we address the validation of resources that were created and used for the second Evaluation Campaign of the project. For the three types of topics covered by the project (ASR, SLT, TTS) the validation of both development and evaluation sets is described. For each type we successively address the description of the data, the validation procedures and the validation results. It is concluded that validation constitutes an important and useful element in the production of high quality TC-STAR language resources.

1. Introduction

The TC-STAR project [1] aims to achieve major breakthroughs in the field of speech-to-speech translation (SST), more specifically automatic speech recognition (ASR), spoken language translation (SLT) and text-to-speech (TTS). TC-STAR focuses at the translation of unconstrained conversational speech as it appears in broadcasted (parliamentary) speeches and meetings. The project started in April 2004 and lasts for a period of three years.

To encourage significant advances in all SST technologies, annual competitive evaluations are organised. ASR, SLT and TTS are evaluated independently and within an end-to-end system. The project targets a selection of unconstrained conversational speech domains—speeches and broadcast news—and three languages: European English, European Spanish, and Mandarin Chinese. For each of these evaluations, development and test databases are produced and validated in the TC-STAR project.

This contribution deals with the validation of the language resources (LR) produced for the second Evaluation Campaign of TC-STAR which took place in March 2006. We will successively address the validation procedure, criteria and results for the three types of LR concerned. More specifically we will only deal with the development and evaluation test sets produced for this second evaluation campaign. Other publications address the training LR for ASR (Van den Heuvel et al., 2006) and the training LR for TTS (Bonafonte et al., 2006) that were produced in TC-STAR.

In TC-STAR the validation of the LR is carried out by SPEX[2].

2. Validation of ASR annotations

2.1 Data description

English and Spanish speeches from the European Parliament Plenary Sessions (EPPS) were obtained via Europe by Satellite and recorded by RWTH[3]. Care was taken that only recordings from politicians (and not from interpreters) were selected for transcription. To this end, for Spanish recordings from the Spanish Parliament and the Spanish Congress (PARL) were added. Although most of the speeches are planned, almost all speakers exhibit the usual effects found in spontaneous speech (hesitations, false starts, articulatory noises).

The text compilation of the speeches given by members of the European Parliament in plenary sessions (translated into all official languages of the EU) is known as the Final Text Edition (FTE). The EUROPARL web site provides all of these reports since April 1996. The FTE aims for high readability and differs notably from the verbatim transcript. Transposition, substitution, deletion and insertion of words can be observed in the reports; for transcription purposes, these could only be used as source for the speaker's identity and the spelling of proper names. The Spanish Parliament also provides session reports. In this case the reports were close to what the speaker has said and were used as starting point for the transcriptions.

The transcriptions were performed with Transcriber[4], a tool for assisting the manual annotation of speech signals. It provides a user-friendly graphical user interface for segmenting long duration speech recordings, transcribing them, and labelling speech turns, topic changes and acoustic conditions.

[1] http://www.tc-star.org

[2] http://www.spex.nl

[3] Rheinisch-Westfälische Technische Hochschule, http://www.rwth-aachen.de

[4] http://trans.sourceforge.net/

The manual annotations of the data include:

- Sections and segments
- Breakpoints: placed at each (grammatical) sentence boundary.
- Speaker information regarding type of speaker (politician, interpreter), name, gender, mastering of English (native/non-native/heavily non-native)
- Non-speech noises of various categories
- Orthographic transcriptions
- Markers for spontaneous speech phenomena: filled pauses, hesitations, mispronunciations, false starts.
- Lexical tags for unintelligible parts, foreign words and words of unknown spelling (e.g. neologisms)

The following events are excluded from the transcription procedure by segmentation: Music, Cross talk, Unintelligible speech, Speech in languages other than the target language, Applause, Programs other than parliament.

For Mandarin, the development set of the first Evaluation Campaign could be re-used. As a result, only an evaluation test set was compiled for this language.

Recording dates for the various data sets are presented in Table I.

Language	Data type	Dev. set	Eval. set
English	EPPS	6-9 June 2005	from 7, 8, and 25 Sep. 2005
Spanish	EPPS	6-9 June 2005 and 4-7 July 2005	5-8 Sep., 26-29 Sep. 24-27 Oct., 14-17 Nov. 2005
Spanish	PARL	N/A	24 Nov 2005
Mandarin	Voice of America	N/A	23-25 Dec. 1998

Table I: recording dates for ASR development sets and test sets

Each development set and evaluation test set was about 3 hours in duration comprising some 25,000 words each. The evaluation test set for Spanish comprised both 3 hours of EPPS speeches and 3 hours of PARL speeches.

2.2 Procedure and results

The correctness of the manual annotations of the material was validated. Below follows a brief account of procedure and criteria for the validation of transcription quality.

- 2000 segments are randomly selected for the validation (including very short ones with only noise), with a maximum of 50 segments/speaker

- Segments are grouped per speaker and offered as such to the validator; this facilitates the speaker verification.

- A native speaker of the language performs the check on the speech part of each segment. 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, the delivered transcription should always receive the benefit of the doubt; only overt errors should be corrected.

Non-speech events are not corrected if the validator preferred another symbol, but considered the given symbol as one of a similar kind.

The following validation criteria are used:

- A max. of 2% of the segments may contain an error in the attribution of speaker characteristics: wrong speaker (mismatch of speaker with speaker name), wrong speaker gender; wrong nativeness classification;

- A max. of 5% of the segments may contain an error in the segment boundaries: no boundary at the end of a sentence, boundary in the middle of a sentence without a natural breakpoint such as a pause, breath pause etc., extremely long segment (> 10s), more than 1 speaker in segment;

- A max. of 5% of the segments may contain an error in the attribution of lexical tags;

- A max. of 5% of the segments may contain an error in the transcription of speech;

- A max. of 10% of the segments may contain an error in the transcription of non-speech events.

These validation criteria were adopted and modified from data collections conducted in SpeechDat and SpeechDat-like projects[5]. The validators were not aware of the validation criteria (the permitted maximum percentage of errors); the percentage errors were computed afterwards by SPEX staff, and compared to the criterion values.

The annotations of all data sets were approved. Apart from the orthographical transcription of the speech parts approval was given after the first check. For orthographical transcriptions of the evaluation test sets approval was obtained after correction of the transcriptions and ensuing re-validation. More detailed results are presented in the Appendix (Table I).

For the third Evaluation Campaign in TC-STAR, SPEX will take care to compute transcription errors in term of Word Error Rates (rather than segment error rates). Further specific attention will be given to avoid that speech in another than the target language is included in the data sets.

[5] http://www.speechdat.org

3. Validation of SLT translations

3.1 Data description

English to Spanish and Spanish to English

Both SLT development and evaluation sets used the same data as the ASR development and evaluation sets, in order to enable end-to-end evaluation. For each development and evaluation test set, subsets of 25,000 words were selected from the EPPS manual transcriptions, and from the FTE documents, in English and in Spanish. EPPS English verbatim transcriptions and FTE documents were translated into Spanish by 2 different translation agencies. EPPS Spanish verbatim transcriptions and FTE documents were translated into English by 2 different translation agencies. For the translations from Spanish into English both EPPS and PARL speeches were used totalling 100,000 words of the source texts. Care was taken that the source texts of each language were from the same speeches as transcribed for the ASR development and evaluation test set.

Mandarin to English

Texts up to a total of 42,000 characters (about 30,000 'words') were selected from the Voice Of America verbatim transcriptions and translated into English by 2 different translation agencies.

A "text" version of the VOA data was made (to have a similar text condition to that of the EPPS FTE data). In this "text"-version the punctuation marks and capitalizations (in English) were retained in the transcriptions, whereas these were removed from the "verbatim" version. Since the development set for the first Evaluation Campaign could be re-used, validation was restricted to an evaluation test set only, covering verbatim transcripts from VOA broadcasts of 23-25 December 1998 (which is identical to the ASR test set).

3.2 Procedure and results

About 1200 words of the source text were selected for validation. Since for Mandarin there is no unequivocal opinion on what words are, the selection was made on counting some 1200 English words in the translated texts. It was warranted that a continuous part from the beginning and from the end of each source text (of the complete file with 25,000 words) was selected. (The translation agencies were of course not aware of this selection procedure). The corresponding part of both translations was then retrieved. The translations from the two agencies were offered to the validators in different files. The validators were native in the target language and at least near-native in the source language.

To ensure consistency from one review to another, the following system was adopted for judging translations.

Error	**Penalty**
Syntactic	4 points
Deviation from guidelines (under Translation Quality)	3 points
Lexical	2 points
Poor English (resp. Spanish) usage	1 point
Significant spelling or punctuation error	½ point (to a maximum of 10 points)

"Poor usage" is like "awkward", i.e. when it reveals a low literacy quality or non-nativeness in style. This penalty system was directly adopted from LDC[6] where it is used as a standard way to validate human translations.

It is essential that the given translation received the benefit of the doubt. Only clear errors should be indicated.

For each error found, the corresponding penalty points were counted. If less than 40 penalty points were counted for the 1200-word sample, the translation was considered acceptable. The validators were not aware of the penalty counting system; they only knew the error categories. The final penalty score was computed afterwards by SPEX staff, and compared to the criterion values.

Since each text was translated by two agencies, two validations per source text (one for each agency) were carried out. Verbatim and FTE translations were distinguished as well. All translations were approved but all evaluation test sets were fully or partially corrected before approval could be given. A more detailed overview of validation results is presented in Table II in the Appendix.

For future translation evaluations we will revise the scoring scheme outlined above, together with our team of validators. For example one might question whether a syntactic error will have a more severe impact on intelligibility of a translation than a lexical error.

4. Validation of TTS annotations

4.1 Data description

Assessment of speech synthesis is needed to determine how well a system or technique performs in comparison to previous versions as well as other approaches (systems & methods). Apart from testing the whole system, all components of the system are evaluated separately. This approach grants better assessment of each component as well as identification of the best techniques in the different speech synthesis processes.

For the second Evaluation Campaign there was no development data especially produced and validated for the project.

For the evaluation of the prosodic and acoustic synthesis modules series of data were produced for all three languages English, Spanish and Mandarin. For the evaluation of the prosody the test material typically was naturally spoken from parliamentary speeches (English, Spanish) and from the National High-Tech program 863 TTS evaluation in 2003 (Mandarin). For the evaluation of the acoustic quality, artificial semantically unpredictable

[6] http://www.ldc.upenn.edu/

sentences were used. The table (adopted and modified from Bonafonte et al. 2006) below shows which modules were tested. We have indicated in the table which data were validated.

Module 1: Text analysis (for English)	
Test M1.1	Evaluation of text normalization and end of sentence detection
Test M1.2	Evaluation of word segmentation (Mandarin)
Test M1.3	Evaluation of POS tagger
Test M1.4	Evaluation of Pronunciation
Module 2: Prosody (for English, Spanish, and Mandarin)	
Test M2.1	Evaluation of prosody (using segmental information, resynthesis)
Module 3: Acoustic generation (for English and Mandarin)	
Test M3.1	Intelligibility (functional test)
Test M3.2	Naturalness
System evaluation (data not validated)	
Test S	System evaluation (based on ITU P.85), MOS Evaluation in end-to-end system, including ASR and translation
Voice conversion (data not validated)	
Test VC.1	Voice conversion *removing* prosody effect
Test VC.2	Voice conversion *including* prosody
Expressive speech (data not validated)	
Test E	Judgment test about speech expressivity

More specifically the evaluation LR can be described as follows:

For English:

- text M1: 500 sentences, 10,000 words (POS tags), M1.4 (1000 names and words with phonemic transcription)
- prosody M2: 6 paragraphs, 74 sentence segments, 509 words (POS tags, Phon. transcriptions)
- naturalness M3.2: as M2
- artificial sentences M3.1: 60 sentences, 469 words (POS tags, phon. transcriptions)

For Spanish:

- prosody M2: 18 paragraphs, 32 sentences, 830 words (POS tags, phon. transcriptions)

For Mandarin:

- prosody M2: 6 paragraphs, 58 sentence segments, 422 'words' (POS tags, Phon. transcriptions in pinyin)
- naturalness M3.2: 6 paragraphs, 437 'words' (POS tags, Phon. transcriptions in pinyin)
- artificial sentences M3.1: 50 sentences, 295 'words' (POS tags, phon. transcriptions)

For the English and Spanish resources there was also a phoneme labelling and phoneme segmentation for the speech of the M2 corpus.

4.2 Procedure and results

For the English Text module M1 the following validation criteria were used:

- For a maximum of 5% of the sentences the end points may be judged as erroneous (sample of 500 sentences)
- A maximum of 5% of the POS-tags may be judged as erroneous (sample of 1200 contiguous POS tags)
- Phonemic transcriptions: A max. of 3% minor errors is allowed; and a max. of 2% severe errors is allowed. This holds for each of the two levels (segmental quality and supra-segmental quality; 'segmental' refers to the phone symbols, and 'suprasegmental' refers to the symbols for syllable boundaries and for stress markers).

The first phonemic transcription of all 1000 entries was validated. The context was preserved, i.e. previous and next words and the full sentence were supplied to the validator, who was a native speaker of British English and a phonetic expert.

The task of the validator was defined as follows:

- The given transcription gets the benefit of the doubt
- The given transcription is correct if it represents a possible pronunciation of the word for common words, and a possible or probable pronunciation for proper names
- Each transcription is rated for both segmental and supra-segmental quality. The segmental quality refers to the phoneme symbols used; the supra-segmental quality refers to the use of syllable boundary markers and stress markers (and tone or morphological markers, if provided).
- A 3-point scale is used for each transcription: OK; Minor error; Severe error. This scale is used twice per transcription, once for segmental quality and once for supra-segmental quality.
- A minor error occurs if only one symbol in the transcription is wrong.
- A severe error occurs if more than one symbol is wrong

The validators were not aware of the validation criteria (the permitted maximum percentages of errors); the percentage errors were computed afterwards by SPEX staff, and compared to the criterion values.
The English Text module was approved according to these criteria (see Table III in the Appendix).

As for the validation of the prosodic and acoustic modules similar criteria were used for the validation of sentence segments, POS tags, and phonemic transcriptions. For the Mandarin phonemic transcriptions in pinyin, no supra-segmentals (except tone) were validated. The Spanish and English evaluation data were approved on these criteria without need for correction and re-validation. For Mandarin, only the POS-tags needed revision.

For English and Spanish phoneme labels and phoneme segmentations of the M2 corpus (1500 phonemes) were validated by comparing the annotations (labels and segment boundaries) with the corresponding speech signal. As criteria were used:

- A maximum of 5% PER is allowed for the phoneme labels

- A maximum of 5% of the phonemes may have a temporal deviation exceeding 25 ms.

It was concluded that both English and Spanish fulfilled these criteria.

More detailed TTS validation results can be found in Table III of the Appendix.

5. Conclusion

In this paper we have presented the validation of the development and evaluation LR that were used in the second Evaluation Campaign of the TC-STAR project. LR were distinguished according to the three main research topics in the project (ASR, SLT and TTS). We have presented the data sets and the procedures employed to validate them. From our findings it becomes evident that validation constitutes an important element in the production of high quality LR. The LR in TC-STAR are optimised through an iterative process of quality assessment and improvement (validation, correction and re-validation).

Due to high time pressure on the production of the development and evaluation data, validation is at times scheduled after data release. This is problematic for development data, and to a lesser extent for evaluation transcripts, if scores can be recomputed later on the corrected and approved transcripts/translations. For the third Evaluation Campaign more effort is needed to schedule validations and corrections before data release.

Regardless of this, the validations are valuable for the production of the Evaluation Suites which are part of the deliverables of TC-STAR. These Evaluation Suites will be made available through the European Language Resources Association (ELRA[7]) and require validated transcripts and translations, as explained above, as a minimum quality demand.

6. Acknowledgement

The TC-STAR project is supported by the EC in the sixth Framework Programme under contract FP6-506738.

7. References

Bonafonte, A., Höge, H., Kiss, I., Moreno, A., Ziegenhain, U.,Van den Heuvel, H., Hain, H.U., Wang, X.S., Garcia, M.N. (2006). TC-STAR:Specifications of Language Resources and Evaluation for Speech Synthesis. In : Proceedings LREC2006, Genoa, Italy.

Gollan, C., Bisani, M., Kanthak, S., Schlüter, R., Ney, H. (2005): Cross Domain Automatic Transcription on the TC-STAR EPPS Corpus. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, March, 2005.

Van den Heuvel, H., Choukri, K., Gollan, Chr., Moreno, A., Mostefa, D. (2006): TC-STAR: new language resources for ASR and SLT purposes. In: Proceedings LREC2006, Genoa, Italy.

[7] http://www.elra.info

APPENDIX : VALIDATION RESULTS

Language	Data type	Data set	Speaker (max 2%)	Segment (max. 5%)	Lex.Tags (max. 5%)	Speech (max. 5%)	Non-Speech (max. 10%)
English	EPPS	Dev. set	0	1	0	5	7
	EPPS	Eval. set	1	2	0	4 R	9
Spanish	EPPS	Dev. set	0	0	0.1	5	7
	EPPS+PARL	Eval. set	0	0	0	5	8
Mandarin	Voice of America	Dev. set	N/A	N/A	N/A	N/A	N/A
	Voice of America	Eval. set	1	1	0	5 R	7

Table I : Validation results in error percentages (segment level) for the ASR data sets. More detailed descriptions of the validation criteria are given in section 2.2. Only end results are presented ; R means that the results was obtained after correction and re-validation of the annotations

Language pair	Data type	Data set	Agency 1	Agency 2
English-Spanish	EPPS	Dev. set	52.5	30.5
	EPPS	Eval. set	40 R	28.5 R
Spanish-English	EPPS	Dev. set	34.5	40
	EPPS	Eval. Set	39.5 R	38 R
	PARL	Dev. set	72	72
	PARL	Eval. set	0 R	0 R
Mandarin-English	Voice of America	Dev. set	N/A	N/A
	Voice of America	Eval. set	39.5 R	38 R

Table II : Validation results in penalty points for the SLT data sets. The maximum number of penalty points allowed is 40. More detailed descriptions of the validation criteria are given in section 3.2. Only end results are presented ; R means that the results was obtained after correction and re-validation of the annotations

Language	Data type	Sentence breaks (max. 5%)	POS-tags (max. 5%)	Segmental transcriptions (max. 5%)	Suprasegmental transcriptions (max. 5%)	Phoneme labels (max. 5%)	Phoneme segments (max. 5%)
English	Text analysis	0.6	1.0	0.6	0.1	N/A	N/A
	Input ac. and pros. modules	9.5	0	1.0	0.2	-	0.5
Spanish	Input ac. and pros. modules	0	2.7	3.4	2.0	3.9	3.7
Mandarin	Input ac. and pros. modules	0	9.0	2.9	N/A	N/A	N/A

Table III : Validation results in error percentages for the TTS data sets. More detailed descriptions of the validation criteria are given in section 4.2. Only end results are presented ; R means that the results was obtained after correction and re-validation of the annotations.