TC-STAR: New language resources for ASR and SLT purposes

Henk van den Heuvel (1), Khalid Choukri (2), Christian Gollan (3), Asuncion Moreno (4), Djamel Mostefa (2)

(1) SPEX/CLST Radboud University Nijmegen, Netherlands; (2) ELDA, Paris, France;
(3) RWTH, Aachen, Germany; (4) UPC, Barcelona, Spain

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) is being produced. In this contribution we address the resources that have been created for Automatic Speech Recrognition and Spoken Language Translation. As yet, these are 14 LR in total: two training SLR for ASR (English and Spanish), three development LR and three evaluation LR for ASR (English, Spanish, Mandarin), and three development LR and three evaluation LR for SLT (English-Spanish, Spanish-English, Mandarin-English). In this paper we describe the properties, validation, and availability of these 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 speech synthesis (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. Automatic Speech Recognition (ASR), Spoken Language Translation (SLT) and Text-To-Speech (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.

At present, a range of Language Resources (LR) has been produced. In this contribution we address the resources that have as yet been produced for ASR and SLT. These are 14 LR in total: two training LR for ASR (English and Spanish), three development LR and three evaluation LR for ASR (English, Spanish, Mandarin), and three development LR and three evaluation LR for SLT (English-Spanish, Spanish-English, Mandarin-English). These LR are addressed in this paper. Another contribution to this conference will deal with the TTS databases produced in TC-STAR (Bonafonte et al, 2006).

2. SLR for ASR training

So far, two LR for ASR training purposes have been produced in TC-STAR, one for European accented English and one for European Spanish. Both LR contain speeches from the European Parliament Plenary Sessions (EPPS), obtained via Europe by Satelite and recorded by RWTH. Additionally, the Spanish LR includes recordings from the Spanish Parliament and the Spanish Congress. Until now over 300 hours of speech per LR have been recorded and the recording process is ongoing. Each audio file is monaural with 16-bit resolution at a sample rate of 16kHz. The EPPS recordings include speeches from politicians who speak in the targeted language and from interpreters. Today 20 official languages are spoken within the European Parliament. Therefore, the larger amount of speeches contained are from interpreters whose speaking style is monotonous as compared to that of the politicians. The speaking style of the polticians can be categorized by their accent or dialect (nativeness). Although most of the speeches are planned, almost all speakers exhibit the usual effects known from spontaneous speech (hesitations, false starts, articulatory noises).

The compilation of texts of the speeches given by members of the European Parliament in plenary sessions (translated in 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.

For the first TC-STAR Evaluation Campaign (March 2005) approximately 40 hours of speech per LR were manually transcribed as training data. Further exact transcriptions were produced for the second Evaluation

Campaign in February 2006. These LRs (speech data and transcripts) are available from ELRA.

The **English** database comprises 102 hours of transcribed recordings leading to almost 800k running words and a 19k vocabulary. The additional 75 hours of untranscribed speech from the EPPS can be used for unsupervised training. The recordings were made between May 2004 and May 2005. The database was transcribed and packaged by RWTH, Aachen (Gollan et al., 2005).

The **Spanish** database comprises both, recordings of members and interpreters of the European Parliament speaking in the parliamentary plenary sessions (EPPS) and recordings of the Spanish Parliament (PARL). Transcription was performed by Applied Technologies on Language and Speech, S.L. (ATLAS), from Spain. The owner of the transcriptions is Universitat Politecnica de Catalunya, from Spain.

Spanish EPPS recordings ranged from May 2004 to January 2005 and were provided by RWTH Aachen.
Spanish recordings of the Spanish Parliament Plenary Session ranged from July 2004 to December 2004 and were provided by Universitat Politècnica de Catalunya.
The recordings were made by internet reception and satellite reception from Europe by Satellite. Satellite recordings were decoded and audio streams resampled to WAV files. Internet recordings were provided as RealMedia streams. Recordings were not processed in any way, including Plenary Session pauses and segments of untranslated speech (language different from target language). The “Day's schedule of EbS" to select the raw segments of a Plenary Session.

EPPS transcriptions consist of 61:53 hours of speech and PARL transcriptions consist of 38:24 hours of speech. The total amount of audio recordings including non transcribed sections is 143:10 hours. Speech files are encoded with 16 kHz, 16 bits, single channel. Format is raw PCM (.WAV) without header information. RealMedia streams were converter to WAV files using WinAmp and RealPlayer software with “Tara Audio Video Plugin for WinAmp”. Each speech file has an accompanying file with the transcription in xml format (extension .TRS).

Table 1 shows the hours of transcribed speech as a function of speakers’ characteristics such as gender, if the speaker is a politician or an interpreter, if the speaker is native or non native both for the EPPS and PARL recordings

The transcriptions were performed with Transcriber, 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.

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, control 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 English, Applause, Programs other than parliament.

Categorization of speakers			EPPS English	EPPS Spanish	PARL Spanish
male	interpreter	native	40:22	22:03	0:00
		non-native	0:54	1:50	0:00
	politician	native	11:02	8:49	27:15
		non-native	6:19	0:16	1:26
female	interpreter	native	26:00	24:24	0:00
		non-native	3:25	3:01	0:00
	politician	native	2:54	1:15	9:43
		non-native	0:38	0:15	0:00

Table 1: Distribution of transcribed speech in terms of speakers’ characteristics [hh:mm].

2.1 Validation

Both the English and Spanish SLR for ASR training were thoroughly validated. At validation it was tested whether the SLR met the minimum requirements imposed by the original specifications. The validation criteria were related to the following properties of the SLRs:

1. Documentation: correctness and completeness
2. Database structure, formats and file names
3. Corpus items: design and completeness
4. Acoustical quality of the speech data
5. Formal correctness of the annotation files
6. Speaker qualifications
7. Recording conditions
8. Annotation quality

The correctness of the manual annotations is considered of primary importance for validation of a SLR. Below follows a brief account of procedure and criteria for the validation of transcription quality.

- 2000 segments are 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.

The following validation criteria are used:

- A max. of 2% of the segments may contain an error in the attribution of speaker characteristics: not same speaker (within same speaker block), wrong speaker gender; wrong nativess 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.

As a general rule, the given transcription should get the benefit of the doubt. Only obvious 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.

Both training SLR met the validation criteria on all domensions tested.

3. ASR Evaluation Suites

ELDA has created an ASR Evaluation Suite for the English and Spanish development and evaluation EPPS speech data. For each language the Evaluation Suite contains the signal files, all transcripts and segmentations, protocols, scoring tools, a proper documentation, and the ASR recognition results of all participants in TC-STAR's first Evaluation Campaign. The aim of the Evaluation Suites is to enable external players to evaluate their own systems and compare the results with those obtained during the first TC-STAR Evaluation Campaign in February 2005.

3.1 English and Spanish

The development data consisted of EPPS recordings (in English and Spanish) from 25 to 28 October 2004, manually transcribed by ELDA. In each language, 3 hours of recordings were selected and transcribed, corresponding to approximately 35,000 running words in English and 33,000 running words in Spanish. Contiguous audio segments were transcribed, up to 3 hours, without special focus on the English- (resp. Spanish-) speaking politicians. ELDA also provided the corresponding Final Text Editions (FTE), which are the official transcriptions of the parliamentary debates, published by the EC in English and Spanish.

For the Spanish and English evaluation sets, the Parliamentary sessions recordings from 15 to 18 November 2004 were manually transcribed. The English test set is made of about 34,000 running words while the Spanish test set contains about 32,000 running words.

While he development material for English and Spanish is mainly composed of interpreters' speeches, the strategy selection for the evaluation data consisted of first transcribing all available English- (resp. Spanish-) speaking politicians, then transcribing up to 3 hours of interpreters' speeches

3.2 Mandarin

For Mandarin Chinese , the development data consisted of 3 hours of audio recordings from the broadcasted news of Mandarin Voice of America between 1 and 11 December 1998 (http://www.ldc.upenn.edu/: Mandarin TDT3 1 Dec 98 to 11 Dec 98, LDC2001S95 & LDC2001T58). It corresponded to approximately 42,000 Chinese characters. ELDA produced the manual transcriptions.

The test set is made of audio recordings between 14 and 22 December 1998, manually transcribed.

		Size in hours	**Number of words or char.**	**Interpeters**	**Politicians**
Development	English	3.75	35,635 words	93,6%	6.4%
	Spanish	3.75	33,101 words	85,9%	14.1%
	Chinese	3.2	43,501 characters	N/A	
Evaluation	English	3.5	34,420 words	39.6%	60.4%
	Spanish	3.75	32,381 words	77%	23%
	Chines	3.2	44,103 characters	N/A	

Table 3 Development and evaluation data statistics

3.3 Validation

The TC-STAR ASR Evaluation suites were validated along the same 8 dimensions as presented in section 2 for the ASR training SLR. Again the validation of the manual annotations was given special attention. The suites were all approved.

4. SLT Evaluation Suites

The transcripts of the ASR development and evaluation sets were translated by professional agencies, maintaining the original segmentations. The three translation directions included were: English-Spanish, Spanish-English, Mandarin-English. Each text was translated by two different agencies.

ELDA prepared Evaluation Suites for the SLT translation data as well. These suites contain all translations, scoring tools, protocols and translation results from the TC-STAR partners during the first TC-STAR Evaluation Campaign.

4.1 English to Spanish and Spanish to English

The SLT development (resp. evaluation) set was built upon the ASR development (resp. evaluation) set, in order to enable end-to-end evaluation. 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 agencies. EPPS Spanish verbatim transcriptions and FTE documents were translated into English by 2 different agencies

4.2 Mandarin to English

Subsets of 25,000 characters were selected from the Voice Of America verbatim transcriptions and translated into English by 2 different agencies.

A "text" version of the VOA data was made (to have a

similar text condition to that of the EPPS's FTE data) preserving punctuation and capitalization (in English) in the transcriptions, while the "verbatim" version was stripped of these features

4.3 Validation

All translations produced by ELDA for the SLT development and test sets were validated by SPEX.. The following properties of the LRs were validated:

1. Documentation: correctness and completeness
2. Database structure, formats and file names
3. Translation quality

Special attention was given to translation quality. About 1,200 words of the source text are selected for validation. It is warranted that a continuous part from the beginning and from the end of each text is selected. The corresponding part of both translations is then retrieved. The translations from the two agencies are offered to the validators in different files.

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

Error	Penalty points
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.

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

For each error found, the corresponding penalty points are counted. If less than 40 penalty points are counted for the 1200-word sample, the translation is considered as acceptable.

Since each text is translated by two agencies, two validations per source text (one for each agency) are carried out. As a result the total number of SLT validations performed was 3 (language combinations) * 2 (LR types: development and evaluation) * 2 (agencies) = 12. Except for one, all 12 translations were approved.

5. Conclusion and future work

In this paper we presented the production and validation of 14 LR that have so far been created in the TC-STAR project: two training SLR for ASR (English and Spanish), three development LR and three evaluation LR for ASR (English, Spanish, Mandarin), and three development LR and three evaluation LR for SLT (English-Spanish, Spanish-English, Mandarin-English). All these LR are made available through ELRA in 2005 and 2006.

For the second Evaluation Campaign the same number of evaluation resources for ASR and SLT (with similar content) is currently under production. Furthermore, training and evaluation resources are presently also produced for TTS purposes (see also Bonafonte et al. 2006).

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:Speci_cations of Language Resources and Evaluation for Speech Synthesis. In : Proceedings LREC2006, Genoa, Italy.

C. Gollan, M. Bisani, S. Kanthak, R. Schlüter, and H. Ney (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.

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