Project no.:	FP6-506738
Project Acronym:	TC-STAR
Project Title:	Technology and Corpora for Speech to Speech Translation
Instrument:	Integrated Project
Thematic Priority:	IST

Deliverable no.: D30
Title: Evaluation report

Due date of the deliverable:	31st of March, 2007
Actual submission date:	14th of May, 2007
Start date of the project:	1st of April 2004
Duration:	36 months
Lead contractor for this deliverable:	ELDA D30
Authors:	D. Mostefa (ELDA), O. Hamon (ELDA), N. Moreau (ELDA) and K. Choukri (ELDA)

Revision:[Final 1.0]

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium(including the Commission Services)	
CO	Confidential, only for members of the consortium(including the Commission Services)	

Table of Contents

1 Introduction

This document reports on the evaluation activities conducted in the third year of the TC-STAR project. The TC-STAR project, financed by the European Commission within the Sixth Framework Program, is envisaged as a long-term effort to advance research in the core technologies of Speech-to-Speech Translation (SST). SST technology is a combination of Automatic Speech Recognition (ASR), Spoken Language Translation (SLT) and Text To Speech (TTS). The project targets a selection of unconstrained conversational speech domain (speeches and broadcast news) and three languages: European English, European Spanish, and Mandarin Chinese. To assess the advances in SST technologies, annual competitive evaluations are organised. The aim of the third evaluation campaign is to measure the progress made during the third year of the project in ASR, SLT, TTS and in the whole end-to-end Speech-to-Speech system. In addition to the measure performance, the infrastructure built in TC-STAR is also evaluated.

The third TC-STAR evaluation campaign took place during months 34-35 of the project, more precisely from 21 January 2007 to 15 March 2007. The results of the evaluation campaign are presented at the third TC-STAR evaluation Workshop in Aachen on March, 28-30 2007.

We first describe the evaluation tasks common to ASR, SLT and TTS. Then, we present the Automatic Speech Recognition evaluation. Then results of Spoken Language Translation are given. The third part of the document concerns the Text To Speech module evaluations. Finally the evaluation of the whole system, referred as end-to-end evaluations are presented.

1.1 Evaluation tasks

To be able to chain the ASR, SLT and TTS components, evaluation tasks are designed to use common sets of raw data and conditions. Three evaluation tasks are selected:

- **European Parliament Plenary Sessions (EPPS)**: the evaluation data consisted of audio recordings of the EPPS original channel of the parliamentary debates, and of the official documents published by the European Community, containing post-edited transcriptions of the sessions, in English and in Spanish. The focus is exclusively on the Parliament Members speaking in English and in Spanish, therefore the interpreters speeches are not used. These resources are used to evaluate ASR in English and Spanish and SLT in the English-to-Spanish (En→Es) and Spanish-to-English (Es→En) directions.

- **CORTES Spanish Parliament Sessions**: since there are few Spanish speeches in the EPPS recordings, audio recordings of the Spanish Parliament (Congreso de Los Diputados) are used. The data are used in addition to the EPPS Spanish data to evaluate ASR in Spanish and SLT from Spanish-to-English (Es→En).

- **Voice Of America**: the evaluation data consisted of audio recordings in Mandarin Chinese (Zh) of the broadcasted news of the Mandarin "Voice of America" (VOA) radio station. Those data are used to evaluate speech recognition systems in Mandarin Chinese and translation from Mandarin into English (Zh→En).

1.2 Participants

The list of TC-STAR participants in the third evaluation campaign is given below.

- internal TC-STAR participants:
  - IBM, Germany,
  - IRST, Istituto Trentino di Cultura - Il Centro per la ricerca scientifica e tecnologica, Italy,
  - LIMSI, Laboratoire d'Informatique pour la Mcanique et les Sciences de l'Ingnieur, France,

  - NOKIA, Finland,
  - RWTH, Rheinisch-westfische Technische Hochschule, Germany,
  - SIEMENS, Germany,
  - UKA, Universität Karlsruhe, Germany,
  - UPC, Universitat Politecnica de Catalunya, Spain.

- external institutions
  - NICT-ATR, Advanced Telecommunications Research Institute International, Japan:
  - CAS Chinese Academy of Science, China,
  - Daedalus Data, Decisions and Language, S. A, Spain,
  - JHU, John Hopkins University, United States,
  - LIUM, Laboratoire d'Informatique de l'Universit du Maine, France,
  - Translendium, Spain,
  - UDS, Universität Des Saarlandes, Germany,
  - VERBIO Speech Technology, Spain,
  - XMU, Xiamen University, China.

Table 1 gives an overview of participation for Automatic Speech Recognition, Spoken Language Translation and Text To Speech. Moreover, in order to compare SLT results with commercial products we have computed the SLT scores of commercial Systran [12] and Softissimo [11] products.

	ASR			SLT			TTS		
	En	Es	Zh	En→Es	Es→En	Zh→En	En	Es	Zh
IBM	X	X	X	X	X		X	X	
ITC-irst	X	X		X	X	X			
LIMSI	X	X	X	X	X				
NOKIA							X		X
RWTH	X	X		X	X	X			
SIEMENS							X	X	
UKA	X		X						
UPC		X					X	X	
ATR				X	X				
CAS						X			X
DAEDALUS	X	X							
JHU				X					
LIUM	X	X							
Translendium				X					
UDS				X	X	X			
Verbio							X	X	
XMU						X			

Table 1: Participants in the Third TC-STAR Evaluation Campaign

2 ASR evaluation

2.1 Tasks and conditions

There are three tasks and three different training conditions for each task.

- the EPPS task: automatic speech recognition systems are evaluated on recordings of the European Parliament's sessions in English and Spanish recorded in June-September 2006,
- the CORTES task: recordings from the Spanish Parliament of June 2006 are used for the evaluation,
- the VOA task: broadcast news recordings of December 1998 of the radio Mandarin *Voice of America* are used.

For each task, three training conditions are defined:

- Restricted training condition (participants can only use data produced within the TC-STAR project),
- Public data condition (all publicly available data can be used for training and has to be documented),
- Open condition (any data before the cut-off date can be used).

Cut-off. The cut-off date is 31st of May 2006 for English and Spanish. Systems are not allowed to use any training data (audio recordings, text data, etc) produced after the 31st of May 2006. For Chinese, a black-out period covering December 1998 is defined, rather than a cut-off date.

Segmentation. A manual segmentation is exploited for the EPPS task to separate the English (respectively the Spanish) part from non-English (respectively non-Spanish) part in the original channel recordings.

Metrics. Classical evaluation metrics are used: Word Error Rate (WER) for the EPPS task, Character Error Rate (CER) for the VOA task.

For Spanish and English, the scoring is done in four modes: with or without case, with or without punctuation. The error rates are computed on the best alignment between the reference (correct sentence) and the hypothesis (system output). The alignment is done by dynamic programming and minimises the misalignment of two strings of words [6].

Three kinds of errors are taken into account when computing the word error rate, i.e. substitution, deletion and insertion errors. Substitution occurs when a reference word is replaced by another word in the best alignment between the reference and the system hypothesis. Deletion happens when a reference word is not present in the system hypothesis in the best alignment. Insertion is when some extra words are present in the system hypothesis in the best alignment between the reference and the hypothesis.

2.2 Language resources

Three sets of data are used, corresponding to the three classical phases of an evaluation: training, development and test.

	Transcribed		Non transcribed	Total
	Politicians	Interpreters		
EPPS English	21h	70h	200h	291h
EPPS Spanish	10h	51h	230h	291h
CORTES Spanish	38h			38h

Table 2: Acoustic training resources for the restricted condition

Language	Usage	Domain	Epoch	Amount
English	Dev	EPPS	Oct04;Nov04;Jun05;Sep 05	12h
Spanish	Dev	EPPS	Oct04;Nov04;Jun05;Sept0;Oct05;Nov05	12h
Spanish	Dev	PARL	Dec04;Nov05	6h
Mandarin	Dev	BN	Dec 1998	12h
English	Eval	EPPS	Jun06-Sept06	3h
Spanish	Eval	EPPS	Jun06-Sept06	3h
Spanish	Eval	PARL	Jun06-Sept06	3h
Mandarin	Eval	BN	Dec98	3h

Table 3: Development and evaluation sets

2.2.1 Training data sets

Restricted condition. For the restricted condition, only data produced within TC-STAR could be used for training purposes. This data is produced on recordings of the European Parliament from 3 May 2004 to 18 May 2006. The audio files are recorded and provided by RWTH. The manual transcriptions of the English recordings are done and provided by RWTH, while those of the Spanish recordings are done and provided by UPC. In addition, for the EPPS tasks, the Final Text Edition (FTE) of the documents published by the European Commission, from April 1996 to May 2006, are downloaded and provided by RWTH. In addition to the EPPS data, 38 hours of the CORTES Spanish parliament are recorded and transcribed by UPC.

Public condition. For the public condition, training data are data sets publicly available though various international Language Resources distribution agencies (ELRA, LDC, . . .).

This year a new corpus of 48 million words from the Hansard British Parliament has been released by ELDA and is used for language modelling in the public training condition.

Open condition. For the open condition, any data before the cut-off date could be used. The cut-off date is 31st of May 2006 for English and Spanish. For Chinese, December 1998 is a blackout period.

2.2.2 Development and evaluation data

Due to the short period between the second and the third evaluation campaign, it is not possible to have enough recordings from the European Parliament to produce 3 hours of development data and 3 hours of test data. Therefore, unlike the first and second evaluation campaigns, no new development data is produced this year. Nevertheless, all the previous development and evaluation sets could be used for system development. For the evaluation, the Parliament sessions from which the audio recordings are selected ran from June to September 2006 for the EPPS tasks. For the CORTES task, recordings from June 2006 are used as evaluation data. For Chinese, audio recordings of Voice of America between 26 and 27 December 1998 are selected. Table 3 gives an overview of the development and test data for each language.

Site	Open	Public	Restricted
IBM	7.1%	9.2%	9.8%
ITC-irst		9.5%	11.3%
LIMSI*		9.1%	10.0%
LIUM		22.1%	22.4%
RWTH		9.0%	9.7%
UKA		9.2%	
TC-STAR *	6.9%		

Table 4: Results in terms of word error rate for English for each training condition.

Validation of language resources. SPEX validated the transcriptions of the development and test sets in English and in Spanish. For that, they selected 2000 segments from each set at random. The development and evaluation transcriptions for Chinese, English and Mandarin are successfully validated by SPEX. More details can be found in [13].

2.3 Evaluation results

The ASR run took place from the 21st to the 28th of January 2007. There are 8 participating sites in the ASR evaluation, 6 from the TC-STAR consortium and 2 external participating sites. In addition to individual submissions, a ROVER combination is performed by the TC-STAR partners involved in speech recognition. Each participant had to submit for scoring the output of at least one system trained under one of the specified conditions (i.e. open, public, or restricted). In total there are 41 different submissions: 24 for English, 16 for Spanish and 1 for Mandarin. The detailed submissions for each training condition are listed in Table 72 in Annex 1.

2.3.1 Results for English

We received 24 different submissions from 6 participating sites. In table 4, we show the best result obtained by each site and in each training condition.

The best results are obtained by IBM in *open* training condition with a WER of 7.1%. The significant gain obtained by IBM compared to the public condition is due to a new approach of using very large web data (12 Giga words) for building language models.

Then the WER of TC-STAR partners submissions in public training condition range 9.0% to 9.5%. From this table we can see that there is an improvement of 0.6%-0.9% between the restricted and the public condition for TC-STAR participants.

The ROVER combination, noted as TC-STAR system in the table, performed with a word error rate of 6.9% which is close to the IBM best system. The TC-STAR combination uses the Recogniser Output Voting Error Reduction (ROVER) method [4]. The ROVER system is able to reduce error rates by exploiting differences in the nature of the errors made by multiple ASR systems.

2.3.2 Results for Spanish

There are 15 submissions from 7 institutions, 5 from TC-STAR and 2 external participating sites, Daedalus and LIUM. LIUM obtained a word error rate of 19.8% which is close to the results they obtained for English. Daedalus uses a commercial speech recogniser system which is not adapted to the task and so their results are quite bad. The best results are obtained by RWTH with a word error rate of 8.9%.

We can see that TC-STAR partners results are close with word error rates from 8.9% to 9.5% and most of their systems used only TC-STAR data (restricted condition).

Site	Open			Public			Restricted		
	EPPS	CORTES	**TOTAL**	EPPS	CORTES	**TOTAL**	EPPS	CORTES	**TOTAL**
Daedalus				45.6%	47.4%	**46.6%**			
IBM	6.9%	11.1%	**9.2%**				7.1%	11.3%	**9.4%**
ITC-irst				7.6%	11.2%	**9.5%**	7.7%	11.2%	**9.6%**
LIMSI							6.9%	11.1%	**9.2%**
LIUM							16.1%	22.9%	**19.8%**
RWTH							6.8%	10.7%	**8.9%**
UPC							20.3%	33.6%	**27.5%**
TC-STAR				5.8%	8.8%	**7.4%**			

Table 5: Results in terms of word error rate for Spanish for each training condition and for each domain (CORTES and EPPS).

Here we can observe a clear improvement of the ROVER combination with a word error rate of 7.4%, which is 1.5% better in absolute than the best system.

In table 5 the results are computed for the whole data sets but also separately for the EPPS data and the Spanish CORTES data. We can see that the results are much more better on EPPS than CORTES and this for each participating site. For example in restricted training condition, the WER is 30 to 40% lower on the EPPS than on the CORTES data. This can be explained to the availability of more training data (audio and text data) for EPPS Spanish.

2.3.3 Results for Mandarin Chinese

There is a common submission from LIMSI and UKA for the Mandarin Voice of America task. First, the UKA system produces a first hypothesis. This one is then used by the LIMSI system to adapt acoustic models and then to produce the final recognition output. For this task the Character Error Rate is 7.5%.

2.4 Progress over the years

To measure the improvement from the start of the project until now, we evaluated also the performance of 2005 and 2006 systems. Some sites (IRST, RWTH and UKA) ran their previous systems on the 2007 evaluation data. So the results shown here are obtained on the same evaluation data sets. Unfortunately, for many sites, their 2005 and 2006 systems are no longer available. Figure 1 and 2 show a comparison of 2005, 2006 and 2007 systems for IRST, RWTH and UKA. We can see that there are important improvements between 2005 and 2007. The improvement is higher between 2005 and 2006 than between 2006 and 2007. The improvements can be explained by the amount of training resources available each year but also improvements in the methods and systems.

2.5 Summary

All TC-STAR partners involved in speech recognition (IBM, IRST, LIMSI, RWTH, and UKA) participated in the ASR evaluations and sent system hypothesis on different conditions (open, public or restricted training data conditions). UPC who is officially involved in SLT and TTS also joined the ASR evaluations and submitted system outputs for the Spanish language. In addition to TC-STAR partners, some external partners joined the ASR evaluation campaign (Daedalus and LIUM). The best word error rate for English is obtained by IBM with 7.1% for a single system. For Spanish the best word error rate is obtained by RWTH with a word error rate 8.9% (6.8% on the Spanish EPPS only). For Chinese the combined system LIMSI/UKA performed with a character error rate of 7.5%. System combinations are performed for English and Spanish and lowered the error rate from 7.1% to 6.9% for English and from

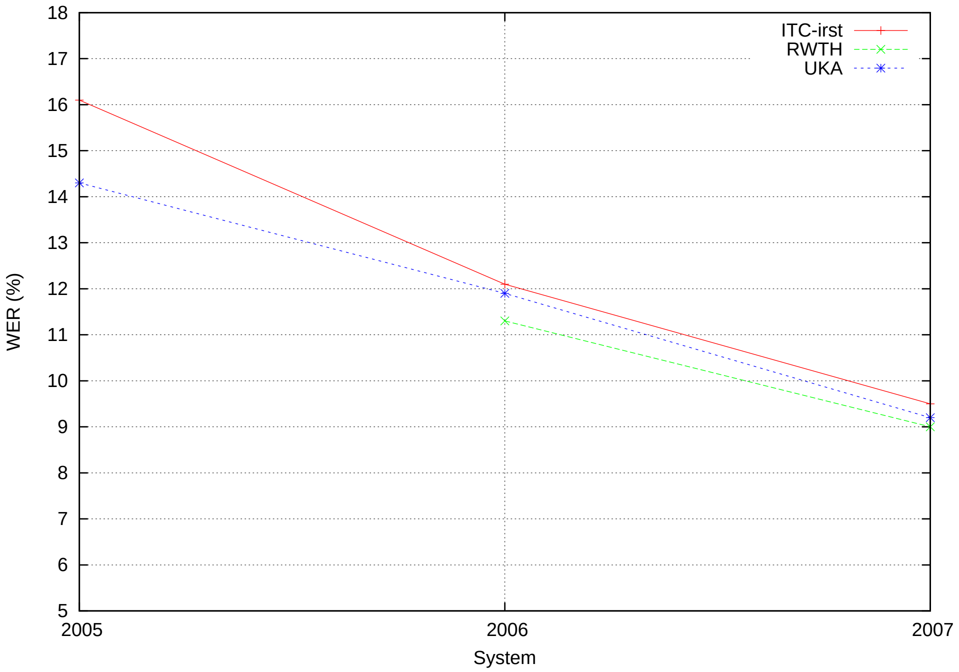


Figure 1: Progress on English for ITC, RWTH and UKA

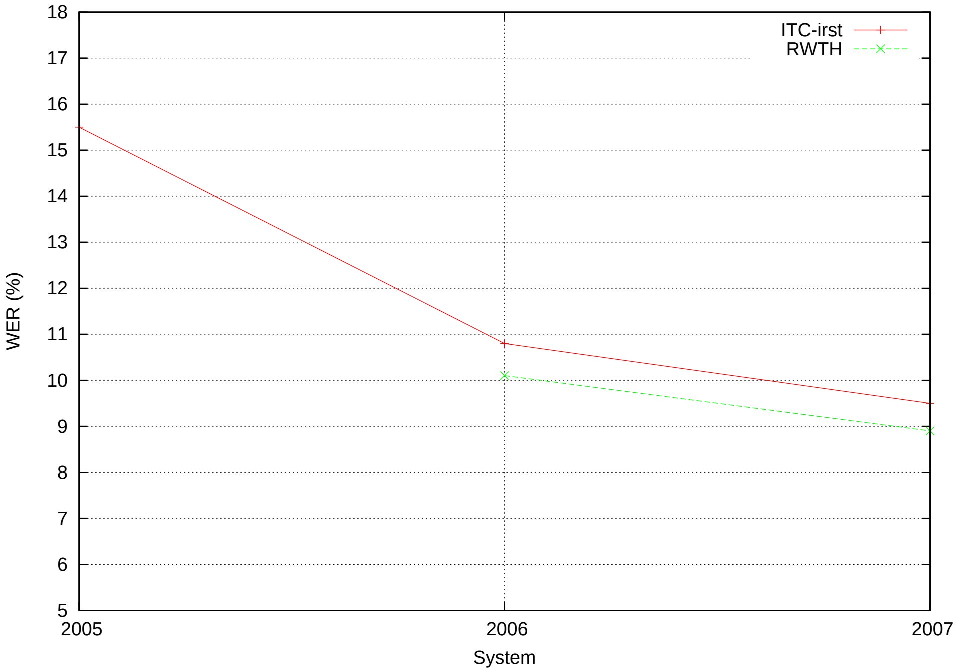


Figure 2: Progress on Spanish for ITC and RWTH

8.9% to 7.4% for Spanish. Three sites (IRST, RWTH and UKA) sent us outputs of their 2005 and 2006 systems. Comparison of results over the years show that word error rates have been largely reduced (see figures 1 and 2).

2.6 Evaluation packages

The data used for the third evaluation campaign in ASR are available as evaluation packages. An evaluation package which includes resources, protocols, scoring tools, results of the ASR official campaign, etc., those are used or produced during the campaigns are available and distributed by ELDA. The aim of this evaluation package is to enable external players to evaluate their own system and compare their results with those obtained during thecampaignitself. Three evaluation packages (one per language) are available on ELRA's catalog of language resources [3].

3 SLT evaluation

3.1 Tasks and conditions

Three different tasks and three translation directions have been considered for the evaluation of the SLT technology: the first one is the EPPS task. Text data from the debates that took place at the European Parliament between the 12^{th} of June and the 28^{th} of September 2006 are used. This task includes two translation directions, English-to-Spanish and Spanish-to-English. An additional CORTES task has been used for the Spanish-to-English direction: text data (manual transcriptions, automatic transcriptions, final text editions) from the debates of the Spanish Parliament that took place on the 14^{th} and the 20^{th} of June 2006 have been added to the Spanish-to-English EPPS data. The third task is the VOA task for the direction Mandarin-to-English. Transcriptions of Mandarin Chinese audio recordings of the Voice of America radio channel are used to evaluate translation systems in the Chinese-to-English direction.

For Spanish-to-English and English-to-Spanish directions, three kinds of text data are used as input:

- The first one is the output of a combination of some automatic speech recognition systems. The ASR ROVER combination, which gives the lowest error rate, is used. The text is in true case and punctuation marks are provided. This year no manual segmentation in sentences is provided and the SLT systems have to segment the ASR output automatically. Then the SLT output data is automatically aligned to the reference translations, in order to produce the segmentation for scoring. This type of data is called “ASR” in the results parts.
- The second type of data is the verbatim transcription. These are manual transcriptions produced by ELDA. These transcriptions include spontaneous speech phenomena, such as corrections, false-starts, etc. The annotations are produced for English and Spanish. As for the ASR output, the text data is provided with punctuation and in true case. This type of data is called “Verbatim” in the results parts.
- The last one is the text data input. Final Text Editions (FTE), provided by the European Parliament and the Spanish Parliament, are used for the EPPS and CORTES tasks. These text transcriptions are edited and differ slightly from the verbatim ones. Some sentences are rewritten. The text data include punctuation, uppercase and lowercase and do not include transcription of spontaneous speech phenomena.

An example of the three kinds of inputs is shown in Table 6 in Annex B.

FTE	*President-in-Office, you mentioned the issue of data retention.*
Verbatim	*you mentioned , President-in-office , about the issue of data retention.*
ASR output	*you mentioned the president in office about the issue of data retention*

Table 6: Example of ASR, verbatim and FTE inputs

For Chinese-to-English direction, two kinds of text data are used as input.

- The first one is the output of the automatic speech recognition systems. The common submission from LIMSI/UKA is used. No punctuation marks are provided. Again this year no manual segmentation in sentences is provided and the SLT output data is automatically aligned to the reference translations for scoring.
- The second type of data is the verbatim transcriptions. These are manual transcriptions produced

by ELDA. These transcriptions include spontaneous speech phenomena, such as hesitations, corrections, false-starts, etc. As for the ASR output, the text data is provided without punctuation.

As for the ASR evaluations, different training conditions are distinguished. The first one is the primary condition (EPPS-Only track) in which systems can only use the data produced within TC-STAR and the LDC Large Data listed in the Table 73. The aim is to have strict comparisons of systems. No additional bilingual data is allowed, but monolingual tools (e.g. POS-taggers) and publicly available monolingual data can be used.

In the secondary condition (Public Data track), any publicly available data before the cut-off date (May 31, 2006) can be used for training purposes.

3.2 Language resources

Three sets of data are used, corresponding to the three standard phases of an evaluation: training, development and test.

3.2.1 Training data sets

The training data for the VOA task are data sets publicly available through various international Language Resources (LR) distribution agencies (LDC, ELRA) and correspond to the training data of the second evaluation campaign.

For the EPPS task, the training data consists of the same data as for ASR training: the Final Text Editions (FTE), in Spanish and English, from April 1996 to May 2006, provided by RWTH and ELDA. They are considered as reference translations from each other to train the systems. The EPPS data is sentence-aligned. Additionally, the manual verbatim transcriptions of the EPPS recordings in English and Spanish from May 2004 to January 2005 are provided by RWTH (English) and UPC (Spanish). Additional data have been provided: EU Bulletin Corpus, JRC-Acquis Multilingual Parallel Corpus and UN Parallel Corpus.

3.2.2 Development and evaluation data sets

The SLT development set corresponds to the development and test data of the first and the second evaluation campaigns. It consists the same data as the ASR development data set, in order to enable end-to-end evaluation.

Subsets of 25,000 words are selected from the EPPS verbatim transcriptions, from the CORTES verbatim transcriptions, from the EPPS FTE documents and from the CORTES FTE documents, in English and in Spanish. Subsets of 25,000 words are selected from the VOA verbatim transcriptions which correspond to the test data of the first evaluation campaign.

ELDA subcontracted professional translation agencies to get reference translations of the data. EPPS English verbatim transcriptions and FTE documents are translated into Spanish by 2 different agencies; EPPS Spanish verbatim transcriptions and FTE documents are translated into English by 2 different agencies; VOA verbatim transcriptions are translated into English by 2 different agencies; CORTES Spanish verbatim transcriptions and FTE documents are translated into English by 2 different agencies.

All source text sets and reference translations presented above are formatted using the same SGML DTD that has been used for the NIST Machine Translation evaluations.

The development data for the ASR task are provided using the outputs of the ASR systems. A ROVER combination has also been provided. The corresponding references are those of the verbatim development data. All source text sets are formatted using the CTM format that has been used in the ASR evaluation.

A summary of the development data used can be seen in Table 74 in Annex B.

As for development, the same procedure is followed to produce the test data. The corresponding data sets used are summarised in Table 75.

As a whole, we have 39 data sets. For a given set, there are:

- the data to be translated, in the source language, organised in documents and segments, except the ASR input which is in CTM format,
- two reference translations of the source data, issued by professional translators, also organised in documents and segments,
- several candidate translations produced by the participants in the evaluation, following the same format of the source and reference sets.

3.2.3 Validation of language resources

SPEX validated the reference translations of the development and test sets for all three translation directions.

For each translation direction and for each reference translation (each set is translated by 2 translation agencies, to produce 2 reference translations) they extracted 1,200 words from contiguous segments selected at random from the source text (except for Mandarin, where they are taken from the target text). Half of the 1200 words are selected from the FTE sources and half from the VERBATIM sources.

The validation criterion is that a reference translation must have less than 40 penalty points to be considered valid. Translation errors are then scored using the following penalty scheme.

Error	**Penalty points**
Syntactical	3 points
Lexical	3 points
Poor usage	1 point
Capitalisation	1 point
Punctuation or spelling errors	0.5 point (to a maximum of 10 points)

Table 7: translation errors penalties

All translations are successfully validated.

Direction	**FTE**		**Verbatim**	
	Ref 1	**Ref 2**	**Ref 1**	**Ref 2**
En◊Es	35	14	40	17
Es→En (EPPS)	18	38	20	40
Es→En (CORTES)	34	35	26.5	22.5
Zh→En	N/A	N/A	27	37

Table 8: Validation results for translation (N/A means that no traslation was available).

The detailed validation results for the reference translations are reported in Table 8.

3.3 Schedule

The development phase took place from September 6, 2006 to January 31, 2007. The evaluation run took place from January 31 to February 7, 2007. The scoring is done in 2 phases. Automatic evaluation is released on February 23 2007. Human evaluation is organised from February 19 to March 20 2007.

3.4 Participants and submissions

The total number of participants in this third evaluation campaign is 12: 6 from the TC-STAR consortium and 6 external participants. External participants are Institute of Computing Technology, China (ICT), The John Hopkins University, United States (JHU), National Institute of Information and Communications Technology - Advanced Telecommunications Research Institute International, Japan (NICT-ATR), Translendium SL, Spain (Translendium), Universität Des Saarlandes, Germany (UDS) and Institute of Artificial Intelligence - Xiamen University, China (XMU)

All participants are allowed to submit for both conditions (Primary and Secondary), and various versions of their systems. The total number of submissions is 176, 57 Submissions for English-to-Spanish, 64 Submissions for Spanish-to-English and 34 Submissions for Chinese-to-English.

There have been 12 submissions for the SLT ROVER (English to Spanish and Spanish-to-English for the three tasks FTE, Verbatim and ASR).

In order to make a comparison with real market products, we ran the evaluation for English-to-Spanish, Spanish-to-English and Chinese-to-English directions with *Systran Professional Premium 5.0* and for English-to-Spanish and Spanish-to-English directions with *Reverso, intranet and development version, Softissimo*. Two translations have been done for Systran, the first one with no specific tuning on the software and the second one in adding the “Business” dictionary. The results obtained with the two systems are shown in the following section, together with those of the other systems.

The submissions received for both condition types are detailed in Table 76 in Annex B.

In order to measure the improvement made within TC-STAR, some 2005 and 2006 systems have been evaluated on the 2007 data. Submissions of 2005/2006 systems are depicted in Table 77.

Thus, 27 additional submissions have been evaluated.

3.5 Evaluation results

The following conditions are applied for evaluation. The same ASR input is used for all systems. It is the result of the ROVER combination of ASR hypotheses, except for the Chinese to English for which the common submission from LIMSI/UKA is used. Case information is used by evaluation metrics. Punctuation marks are present in all the inputs, except Chinese inputs.

3.5.1 Human evaluation

Protocol. The evaluation is carried out on the English to Spanish direction only. All kinds of input (ASR, Verbatim, and FTE) are evaluated in this direction. The primary outputs of all the systems are evaluated as well as the reference translations produced by professional translators. For comparison purposes, we have also added the translation provided by the *Systran* and *Softissimo* products. Furthermore, 2006 and 2005 systems have also been evaluated on the 2007 evaluation data sets.

Each segment is evaluated in relation to adequacy and fluency measures. For the evaluation of adequacy, the target segment is compared to a reference segment. For the evaluation of fluency, only the syntactical quality of the translation is evaluated. The evaluators grade all the segments firstly according to fluency, and then according to adequacy, so that both types of measures are done independently, but making sure that each evaluator does both for a certain number of segments.

For the evaluation of fluency, evaluators have to answer the question: “Is the text written in good Spanish?” A five-point scale is provided where only extreme marks are explicitly defined, ranging from “Perfect Spanish” to “Non understandable Spanish”.

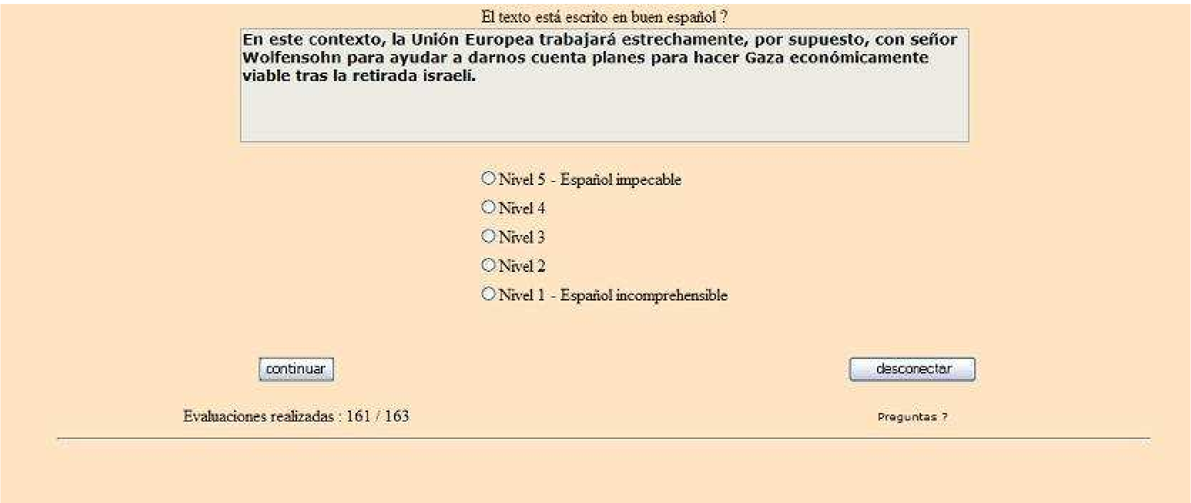
For the evaluation of adequacy, evaluators have to answer the question: “How much of the meaning expressed in the reference translation is also expressed in the target translation?”.

A five-point scale is also provided to the evaluators, where, once again, only extreme cases are explicitly defined, going from “All the meaning” to “Nothing in common”.

Two evaluations are carried out per segment, they are done by two different evaluators, and segments are distributed to evaluators randomly, because evaluators should not build a “storyline” and preserve information between two adjoining segments.

Evaluators are native speakers of the target language educated up to university level.

Evaluation interface. In order to perform the evaluation, we re-used a specific web interface which has already been used for the human evaluation of the French CESTA project [5]. This has been adapted to the Spanish language. This web interface allows for online evaluation, which means that the judges can work at home. This interface has been developed in PHP/MySQL and can be used with a standard browser on Windows or Linux. Figure 1 shows the evaluation page for fluency.



El texto está escrito en buen español ?

En este contexto, la Unión Europea trabajará estrechamente, por supuesto, con señor Wolfensohn para ayudar a darnos cuenta planes para hacer Gaza económicamente viable tras la retirada israeli.

○ Nivel 5 - Español impecable
○ Nivel 4
○ Nivel 3
○ Nivel 2
○ Nivel 1 - Español incomprehensible

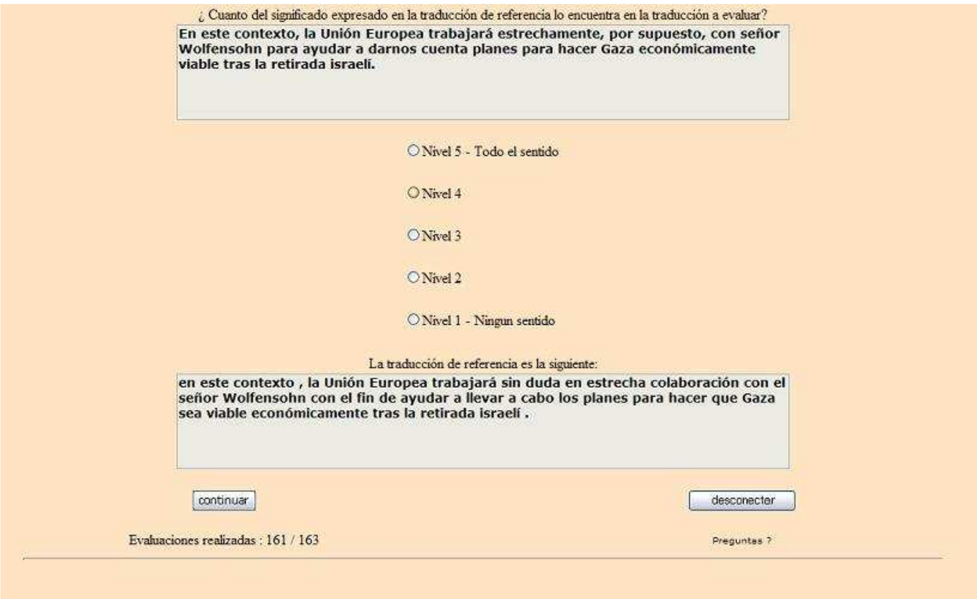
Evaluaciones realizadas : 161 / 163

Figure 3: Fluency evaluation.

From top to bottom of Figure 3, the following items are displayed on this page: the key question for the evaluation of fluency, the text to evaluate, 5 radio-buttons for the 5-point scale measuring fluency, a button to continue the evaluation and move on to the next segment (“continuar”), a button to leave the evaluation (“desconectar”), the number of evaluations done and the total of evaluations to do (“Evaluaciones realizadas”) and a link allowing the evaluator to ask for help should he/she have any questions or problems (“Preguntas?”).

The evaluator reads the text to evaluate in the editing window and can click with the mouse on one of the five radio-buttons proposed. When the evaluation of the text is completed, he/she can move on to the next evaluation. The evaluation is saved automatically and the evaluator does not need to do anything else.

From top to bottom of 4, the following items are displayed on this page: the question for the evaluation of adequacy, the text to evaluate, 5 radio-buttons for the 5-point scale measuring adequacy, the reference text to compare to the text to evaluate, a button to continue the evaluation and move on to the next segment (“continuar”), a button to leave the evaluation (“desconectar”), the number of evaluations



¿ Cuanto del significado expresado en la traducción de referencia lo encuentra en la traducción a evaluar?

En este contexto, la Unión Europea trabajará estrechamente, por supuesto, con señor Wolfensohn para ayudar a darnos cuenta planes para hacer Gaza económicamente viable tras la retirada israelí.

○ Nivel 5 - Todo el sentido

○ Nivel 4

○ Nivel 3

○ Nivel 2

○ Nivel 1 - Ningun sentido

La traducción de referencia es la siguiente:

en este contexto , la Unión Europea trabajará sin duda en estrecha colaboración con el señor Wolfensohn con el fin de ayudar a llevar a cabo los planes para hacer que Gaza sea viable económicamente tras la retirada israelí .

Evaluaciones realizadas : 161 / 163

Figure 4: Adequacy Evaluation.

done and the total of evaluations to do ("Evalautciones realizadas") and a link allowing the evaluator to ask for help should he/she has any questions or problems.

The evaluator reads the text to evaluate, then, compares it to the reference text and finally assigns a score to the segment by clicking with the mouse on a radio-button. When the evaluation is completed the evaluator can move on to the next evaluation. The evaluation done is also registered automatically.

Set up

Data. Taking into account all the different SLT tasks considered (FTE, Verbatim, ASR), the ROVERs, the Systran and Softissimo products, the human reference translations (for Verbatim/ASR and FTE) and the 2005/2006 systems, there are 14 ASR outputs, 16 Verbatim outputs and 15 FTE outputs to evaluate. A subset of around 350 sentences or segments is extracted for evaluation from each output, which corresponds to one third of the whole output. The subset corresponds to a selection of 20 speeches common with the manual end-to-end evaluation.

Evaluators. The number of evaluators (i.e. judges) is restricted to the number of segments to be evaluated and the duration of the evaluation. Two evaluations are done per segment, and both are done by two different judges. 100 evaluators are recruited and are native speakers of Spanish. Table 9 provides a summary of the details for human evaluation.

Number of evaluators	**number of evaluation / segment**	**Task**	**Number of segments**	**Number of systems**	**Total number of evaluations**	**#Evaluation segments / Evaluator**
100	2	FTE	339	15	10 170	317.7
		Verbatim	360	16	11 520	
		ASR	360	14	10 080	

Table 9: Statistics on the human evaluation

The 100 evaluators have to evaluate around 300 segments which correspond to around 5 hours of evaluation (according to a time of one minute by sentence).

Evaluators have been mainly taken from our internal list of contacts. Some other means of recruitment are employed as posters and leaflets distributed, snowball recruitment, contacts with universities.

The segments are shared with the manual end-to-end evaluation. Therefore segments taken from 20 excerpts of speech have been selected. It represents 30% of each system.

Evaluators agreement. Each segment within the human evaluation has been evaluated twice, so as to measure consistency in the evaluations carried out and to have significant number of judgements. This is done by first computing the ratio between those scores which are identical for two evaluations and the total number of segments.

The total agreement between the evaluators has proven to be rather good, about one third of the segments obtain identical evaluations with the two evaluators. This is similar with the last year evaluation. The agreement on FTE data is slightly higher than the agreement on Verbatim. And the agreement on ASR data is much lower than the two others. The FTE and Verbatim tasks are more or less equally difficult to evaluate, while the ASR task is much more difficult.

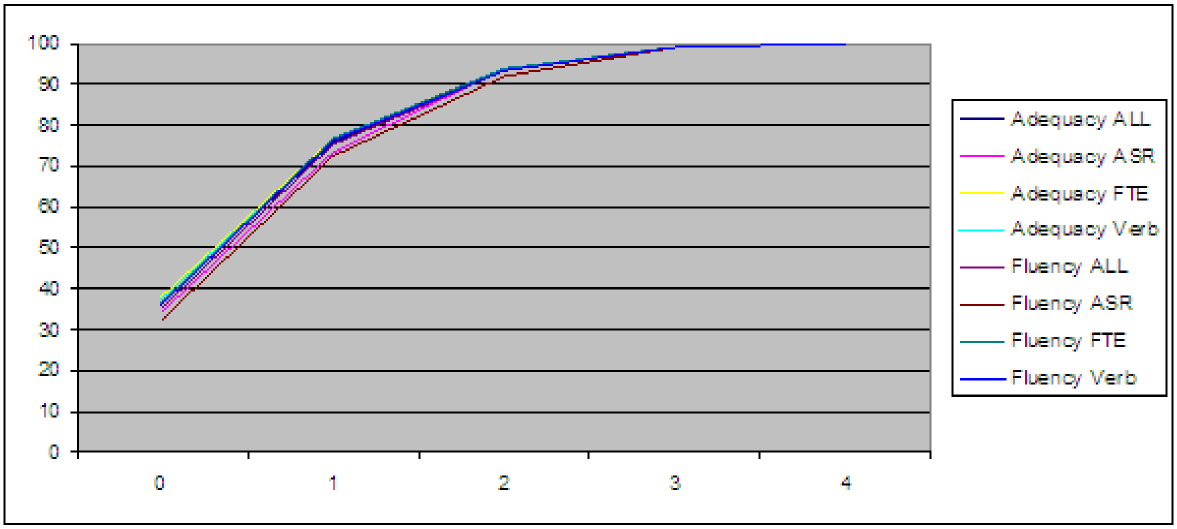


Figure 5: Total agreement between the 1st and 2nd evaluation.

Each segment has been evaluated twice by two different people. The evaluators have to score the adequacy and fluency on a five-point scale. Figure 5 shows the percentage of sentences that have a score difference of less than the value on the x-axis.

We can see that more than 30% of the segments have obtained exactly the same score and than more than 70% have obtained a score that do not differ more than 1 point between the first evaluation pass and the second one.

Table 10 shows the mean of the deviance between two evaluations of a same segment (done by two different evaluators) and Table 11 shows the standard deviance of the deviances. Both tables permit to give an impression of the disagreement between the human evaluators.

	FTE + Verb. + ASR	**FTE**	**Verbatim**	**ASR**
Fluency	0.98	0.94	0.96	1.05
Adequacy	0.95	0.92	0.94	0.99

Table 10: Mean of the deviance

	FTE + Verb. + ASR	**FTE**	**Verbatim**	**ASR**
Fluency	0.93	0.91	0.93	0.94
Adequacy	0.92	0.91	0.93	0.92

Table 11: Standard deviance of the deviances

Table 10 and Table 11 lead us to the same conclusions, except for the fact that the standard deviance of the deviance (between two evaluations of the same segment) is more significant. Thus, the two evaluators certainly assessed differently, although it should also be considered that human evaluation is subjective.

In a general trend, evaluators' agreements are better than last year.

Results. The results obtained for the different tasks are detailed below.

FTE task. First, evaluation scores have been computed and, then, the ranking of the participating systems has been established.

	Fluency score	Adequacy score	Fluency rank	Adequacy rank
Human Reference	4.49 ±.02	4.49 ±.02	1	1
IRST	3.57 ±.05	3.71 ±.04	2	6
SLT ROVER	3.50 ±.05	3.78 ±.04	3	2
UPC	3.47 ±.05	3.77 ±.05	4	3
UKA	3.43 ±.04	3.72 ±.04	5	5
IBM	3.42 ±.05	3.59 ±.05	6	7
RWTH	3.38 ±.05	3.75 ±.04	7	4
Reverso	3.30 ±.04	3.59 ±.04	8	7
UDS	3.22 ±.05	3.37 ±.05	9	9
Systran	3.12 ±.04	3.37 ±.05	10	9

Table 12: Human scoring and ranking for the FTE task

Table 12 shows the ranking of the systems that have participated in the FTE task. It also details the specific scores obtained by each system, which range between 5 (good) and 1 (bad), and the confidence interval.

The human reference gets from afar the best results even they are not so perfect than we could estimate. Regarding the general performance of the systems, after the human reference, the automatic system obtaining the highest score is IRST, surprisingly higher than the ROVER scores for Fluency. For fluency evaluation, following systems are UPC, UKA, IBM and RWTH close from each others. Conclusions are the same for adequacy, except for IBM who obtains subsequently lower results (identical to Reverso). Finally, Reverso, UDS and Systran get the lower results, for both fluency and adequacy.

The difference between the human reference and the automatic systems is still considerable. When considering the performance of systems for fluency and adequacy, all of them obtain higher scores for adequacy than for fluency.

The ranking shows some differences between fluency and adequacy: Higher differences are for IRST (2^{nd} position for fluency, 6^{th} position for adequacy) and RWTH (7^{th} position for fluency, 4^{th} position for adequacy).

Verbatim task. We first compute the scores and establish the ranking of the systems.

	Fluency score	Adequacy score	Fluency rank	Adequacy rank
Human Reference	4.24 ±.03	4.39 ±.03	1	1
RWTH	3.39 ±.05	3.61 ±.05	2	5
SLT ROVER	3.37 ±.05	3.71 ±.05	3	2
IRST	3.35 ±.05	3.60 ±.04	4	6
LIMSI	3.32 ±.04	3.57 ±.05	5	7
UKA	3.31 ±.05	3.64 ±.04	6	3
UPC	3.25 ±.05	3.62 ±.04	7	4
IBM	3.24 ±.05	3.54 ±.05	8	8
Reverso	3.08 ±.05	3.39 ±.05	9	9
UDS	3.07 ±.05	3.24 ±.04	10	10
Systran	2.84 ±.04	3.18 ±.05	11	11

Table 13: Human scoring and ranking for the Verbatim task

Here again the human reference gets the best results. Fluency scores are very close for all the TC-STAR systems, ROVER included (0.15 points between the 1^{st} automatic system – RWTH – and the 7^{th} automatic system –IBM). Anyway, ROVER has higher results for adequacy. But the others TC-STAR systems obtain again scores far from each others. Reverso, UDS and Systran have the lower results, in particular Systran with the fluency evaluation. Rankings are quite different according to the fluency or the adequacy evaluation.

As we can see, even for the human translators, FTE is easier to translate than Verbatim, according to the difference of 0.25 in the scores.

ASR Task. Table 14 outlines the scores and establishes the ranking of systems.

	Fluency score	Adequacy score	Fluency rank	Adequacy rank
IRST	3.09 ±.05	3.19 ±.05	1	1
SLT ROVER	3.04 ±.04	3.15 ±.04	2	4
LIMSI	2.99 ±.04	3.17 ±.04	3	3
RWTH	2.95 ±.05	3.11 ±.05	4	5
IBM	2.91 ±.04	3.06 ±.05	5	6
UKA	2.89 ±.05	3.18 ±.05	6	2
UPC	2.87 ±.05	3.04 ±.04	7	7
Reverso	2.51 ±.04	2.53 ±.04	8	8
Systran	2.42 ±.04	2.50 ±.04	9	9

Table 14: Human scoring and ranking for the ASR task

IRST gets the higher scores for the ASR evaluation for both fluency and adequacy. Next is the

ROVER combination system for fluency, and so on the others TC-STAR systems which are very close: LIMSI, RWTH, IBM, UKA and UPC. Reverso and, in a most important way, Systran get lower results for fluency. Conclusions are rather the same for adequacy, except ranking which is quite different and lower IBM and UPC results.

2006 systems. Table 15 shows the scores and establishes the ranking of systems.

	Fluency score	Adequacy score	Fluency rank	Adequacy rank
RWTH-2006 (fte)	3.41 ±.05	3.64 ±.04	1	1
IRST-2006 (fte)	3.32 ±.05	3.62 ±.04	2	2
IBM-2006 (fte)	3.32 ±.05	3.55 ±.05	2	4
RWTH- 2006 (Verb)	3.26 ±.04	3.57 ±.05	4	3
IBM-2006 (Verb)	3.20 ±.04	3.50 ±.04	5	5
IRST-2006 (Verb)	3.11 ±.05	3.47 ±.05	6	6
IBM-2006 (ASR)	2.90 ±.04	2.98 ±.05	7	7
IRST-2006 (ASR)	2.81 ±.04	2.83 ±.05	8	8

Table 15: Human scoring and ranking for the 2006 systems

As previously said, he difference between the human reference and the automatic systems is still considerable, but as we can observe in Table 15, TC-STAR systems improve performance over the years. Best 2006 FTE system is lower of 0.16 points than the best 2007 FTE systems (resp. 0.13 points for Verbatim and 0.19 points for ASR). Moreover, each 2007 system obtain higher score than its corresponding 2006 system.

Summary. As a general comment, the previous results show that the FTE scores are globally better than the Verbatim scores, and both are better than the ASR scores. Figure 6 sums up the differences.

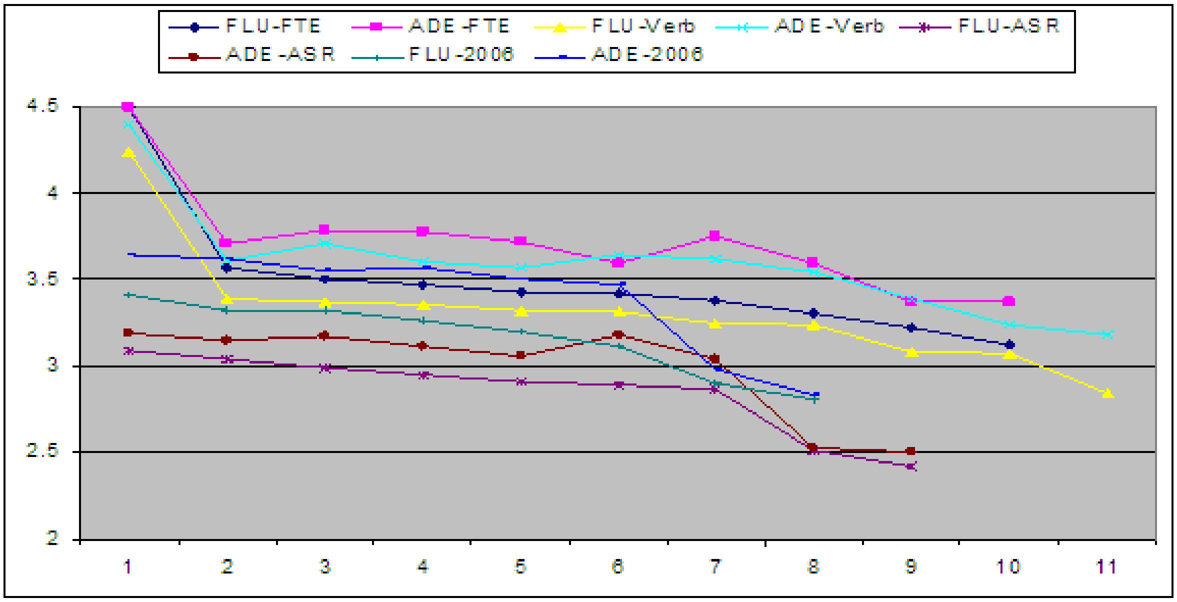


Figure 6: Differences between FTE, Verb. and ASR scores.

Finally, Table 16 summaries the overall ranking of the whole human evaluation for both fluency and adequacy (from the higher scores to the lower).

Fluency ranking	Adequacy ranking
Human reference (FTE)	Human reference (FTE)
Human reference (Verbatim)	Human reference (Verbatim)
IRST (FTE)	SLT ROVER (FTE)
SLT ROVER (FTE)	UPC (FTE)
UPC (FTE)	RWTH (FTE)
UKA (FTE)	UKA (FTE)
IBM (FTE)	SLT ROVER (Verbatim)
RWTH-2006 (FTE)	IRST-BEST (FTE)
RWTH (Verbatim)	UKA (Verbatim)
RWTH (FTE)	RWTH-2006 (FTE)
SLT ROVER (Verbatim)	IRST-2006 (FTE)
IRST (Verbatim)	UPC (Verbatim)
IBM-2006 (FTE)	RWTH (Verbatim)
IRST-2006 (FTE)	IRST (Verbatim)
LIMSI (Verbatim)	IBM (FTE)
UKA (Verbatim)	Reverso (FTE)
Reverso (FTE)	LIMSI (Verbatim)
RWTH-2006 (Verbatim)	RWTH-2006 (Verbatim)
UPC (Verbatim)	IBM-2006 (FTE)
IBM (Verbatim)	IBM (Verbatim)
UDS (FTE)	IBM-2006 (Verbatim)
IBM-2006 (Verbatim)	IRST-2006 (Verbatim)
Systran (FTE)	Reverso (Verbatim)
IRST-2006 (Verbatim)	UDS (FTE)
IRST (ASR)	Systran (FTE)
Reverso (Verbatim)	UDS (Verbatim)
UDS (Verbatim)	IRST (ASR)
SLT ROVER (ASR)	Systran (Verbatim)
LIMSI (ASR)	UKA (ASR)
RWTH (ASR)	LIMSI (ASR)
IBM (ASR)	SLT ROVER (ASR)
IBM-2006 (ASR)	RWTH (ASR)
UKA (ASR)	IBM (ASR)
UPC (ASR)	UPC (ASR)
Systran (Verbatim)	IBM-2006 (ASR)
IRST-2006 (ASR)	IRST-2006 (ASR)
Reverso (ASR)	Reverso (ASR)
Systran (ASR)	Systran (ASR)

Table 16: Overall ranking of the evaluation

This table allows to observing the general trend of the scores: FTE results are higher than Verbatim ones, which are closer to the 2006 FTE results. 2006 Verbatim results are almost higher than 2007 ASR results.

3.5.2 Automatic evaluations

We use five different automatic metrics for the evaluation of the translation output.

BLEU. BLEU, which stands for BiLingual Evaluation Understudy, counts the number of word sequences (n-grams) in a sentence to be evaluated, which are common with one or more reference translations. A translation is considered better if it shares a larger number of n-grams with the reference translations. In addition, BLEU applies a penalty to those translations whose length significantly differs from that of the reference translations.

NIST. NIST is a variant metric of BLEU, from NIST, which applies different weight for the n-grams, functions of information gain and length penalty.

IBM. IBM is a variant metric from IBM, with a confidence interval.

mWER. mWER, Multi reference Word Error Rate, computes the percentage of words which are to be inserted, deleted or substituted in the translation sentence in order to obtain the reference sentence.

mPER. mPER, Multi reference Position independent word Error Rate, is the same metric as mWER, but without taking into account the position of the words in the sentence.

WNM. The Weighted N-gram Model is a combination of BLEU and the Legitimate Translation Variation (LTV) metrics, which assign weights to words in the BLEU formulae depending on their frequency (computed using TF.IDF [10]). We only give in this report the f-measure which is a combination of the recall and the precision.

AS-WER. The AS-WER is the Word Error Rate score obtained during the alignment of the output from the ASR task with the reference translations.

All scores are given in percentages, except NIST. For IBM, BLEU, NIST, WNM/F-measure the higher values mean better translations. On the other hand, for mPER and mWER, which are error rates, the lower values mean better translations.

Automatic results for English-to-Spanish. The statistics for the source documents are the following:

- Verbatim: 27 056 words for 1 167 sentences
- Text: 24 711 words for 1 130 sentences.
- ASR: 26 732 words.

As it can be seen, there is a higher number of words in the manual transcription (27 056) than in the final text edition (24 711). This is due to the hesitations, repetitions, etc. that can be found in the transcriptions. The number of words in the automatic transcription is slightly lower than the manual one (26 732 versus 27 056).

The ratio between the source text in English and the reference translation in Spanish is 0.92, which outlines a strong correlation between the length of the source sentence and its corresponding translation. IRST and UKA systems which strongly move away from this point of balance should be penalised by automatic metrics (at least by the NIST metric). The same occurs with the verbatim output, as the IRST and UKA outputs are 1.04 rather than 0.98 for the reference. All the other outputs from all the tasks are close to the reference file, and then are not penalised too much.

Table 17 presents the scoring results of primary systems for English-to-Spanish EPPS

Task	Site	NIST	BLEU	IBM	mWER	mPER	WNM	AS-WER
FTE	IBM	10.08	50.91	50.86	38.46	30.23	51.31	-
	IRST	10.37	52.26	51.52	35.92	28.84	51.78	-
	RWTH	10.38	52.49	52.01	36.83	28.71	51.21	-
	UKA	10.54	54.11	52.97	35.25	27.99	52.06	-
	UPC	10.43	53.29	52.80	36.09	28.75	51.65	-
	UDS	8.90	40.99	40.90	45.77	35.07	46.50	-
	ROVER	***10.44***	***53.85***	***53.85***	***35.70***	***28.34***	***53.43***	-
	Reverso	*8.39*	*36.82*	*36.83*	*50.45*	*40.64*	*43.54*	-
	Systran	*8.43*	*36.29*	*36.28*	*46.53*	*37.33*	*42.87*	-
Verbatim	IBM	9.84	48.24	48.12	40.86	30.99	49.41	-
	IRST	10.25	50.55	49.46	37.90	29.17	50.96	-
	LIMSI	10.29	51.53	51.04	37.86	28.76	50.44	-
	RWTH	10.13	50.06	49.26	39.13	30.07	50.95	-
	UKA	10.33	50.74	49.77	37.14	28.93	50.43	-
	UPC	9.99	48.93	48.67	39.99	30.65	49.83	-
	UDS	8.38	37.36	37.39	51.44	38.25	45.94	-
	ROVER	***10.43***	***52.63***	***51.46***	***36.74***	***28.82***	***52.06***	-
	Reverso	*8.27*	*35.54*	*35.57*	*52.44*	*41.19*	*42.01*	-
	Systran	*8.25*	*34.79*	*34.78*	*48.66*	*38.10*	*41.15*	-
ASR	IBM	8.62	37.15	36.43	50.49	38.48	46.17	49.66
	IRST	9.03	39.32	38.78	46.52	36.91	48.69	45.84
	LIMSI	8.94	38.29	37.56	47.61	37.55	47.34	46.87
	RWTH	8.92	39.66	38.69	48.16	36.91	47.69	47.45
	UKA	8.70	36.55	36.18	47.19	38.51	47.06	46.77
	UPC	8.65	36.43	35.82	49.63	38.87	46.08	48.92
	ROVER	***9.19***	***40.61***	***40.12***	***45.22***	***36.28***	***49.01***	***44.74***
	Reverso	*7.10*	*25.36*	*25.36*	*63.52*	*50.24*	*38.92*	*59.58*
	Systran	*7.06*	*24.74*	*24.35*	*60.33*	*46.94*	*37.63*	*60.44*

Table 17: Evaluation results of primary systems for the English-to-Spanish

Table 18 presents the ranking results of primary systems for English-to-Spanish EPPS.

Task	Site	NIST	BLEU	IBM	mWER	mPER	WNM	AS-WER
FTE	IBM	6	6	6	6	6	5	-
	IRST	5	5	5	3	5	3	-
	RWTH	4	4	4	5	3	6	-
	UKA	1	1	2	1	1	2	-
	UPC	3	3	3	4	4	4	-
	UDS	7	7	7	7	7	7	-
	ROVER	***2***	***2***	***1***	***2***	***2***	***1***	-
	Reverso	*9*	*8*	*8*	*9*	*9*	*8*	-
	Systran	*8*	*9*	*9*	*8*	*8*	*9*	-
	IBM	7	7	7	7	7	7	-
	IRST	4	4	4	4	4	2	-

Verbatim

	LIMSI	3	2	2	3	1	4	-
	RWTH	5	5	5	5	5	3	-
	UKA	2	3	3	2	3	5	-
	UPC	6	6	6	6	6	6	-
	UDS	8	8	8	9	9	8	
	ROVER	***1***	***1***	***1***	***1***	***2***	***1***	-
	Reverso	*9*	*9*	*9*	*10*	*10*	*9*	-
	Systran	*10*	*10*	*10*	*8*	*8*	*10*	-
ASR	IBM	7	5	5	7	5	6	7
	IRST	2	3	2	2	2	2	2
	LIMSI	3	4	4	4	4	4	4
	RWTH	4	2	3	5	2	3	5
	UKA	5	6	6	3	6	5	3
	UPC	6	7	7	6	7	7	6
	ROVER	***1***	***1***	***1***	***1***	***1***	***1***	***1***
	Reverso	*8*	*8*	*8*	*9*	*9*	*8*	*8*
	Systran	*9*	*9*	*9*	*8*	*8*	*9*	*9*

Table 18: Ranking of primary systems for the English-to-Spanish EPPS task

The SLT ROVER system combination got the best results for all the tasks. UKA has the best results for FTE task, LIMSI and UKA share the best results for the Verbatim task, and IRST, LIMSI and RWTH share the best results for the ASR task. The non-TCSTAR systems (UDS, Reverso and Systran) are really lower than the TC-STAR systems. Except a likely lower quality, it should be put into perspective that Reverso and Systran products are used as is, without any tuning. Moreover, it is well-known that n-gram metrics favour statistical systems. Within the TC-STAR consortium, we observe several groups for each task. For the FTE task, UKA got clearly higher results and IBM clearly lower results than the other participants. For the Verbatim task, UKA is also the leader, while IBM and UPC are clearly lower. Finally, for the ASR task, IBM, UKA and UPC are quite lower than IRST, LIMSI and RWTH.

Contrary to the last year reflection, FTE and Verbatim scores are quite similar for most of the participants. Last year, FTE scores are substantially higher than Verbatim scores and it seems the gap between the two kinds of data tended to be cut down (3.20 BLEU points between the last year best FTE and Verbatim systems instead of 2.58 BLEU points for this year). ASR scores follow the same way but no so manifestly (15.27 BLEU points between the last year best FTE and ASR systems instead of 14.45 BLEU points for this year). In addition, ROVER permits to reduce in a more important way the differences: 1.22 BLEU points of difference between FTE and Verbatim and 13.24 between FTE and ASR.

We can notice that BLEU metric gave higher scores than IBM metric. We also observe that results for mPER are approximately 8-9% higher than for mWER. That is slightly higher than last year results (10% higher): we can conclude the participants have improved the word reordering of their systems.

Table 19 presents the results of secondary systems for English-to-Spanish EPPS task.

Task	**Site**	**NIST**	**BLEU**	**IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**
FTE	UPC	10.20	51.79	51.29	37.31	29.94	51.25	-
Verbatim	UPC	8.97	40.40	39.83	43.82	35.15	44.61	-
ASR	UPC	8.15	33.36	33.07	49.74	41.09	44.72	49.50

Table 19: Evaluation results of secondary systems for the English-to-Spanish EPPS task

Since only one system participated in the secondary track, it is quite difficult to draw any conclusion. Anyway, we can notice that results are slightly lower for this track compared to the primary track. The differences between FTE, Verbatim and ASR tasks are more marked than for the primary track.

Automatic results for Spanish-to-English. Data statistics for Spanish-to-English source documents are the following:

- Text: 50 311 words, for 1 470 sentences whereof
  - CORTES: 25 084 words, for 642 sentences
  - EPPS: 25 227 words, for 828 sentences
- Verbatim: 56 884 words, for 1 342 sentences whereof
  - CORTES: 30 223 words, for 596 sentences
  - EPPS: 26 661 words, for 746 sentences
- ASR: 59 770 words whereof
  - CORTES: 31 734 words.
  - EPPS: 28 036 words.

There are fewer words in the manual transcriptions (56 884 words for Verbatim CORTES and EPPS) than in the automatic ones (59 770 words for ASR CORTES and EPPS).

The same remarks as for English-to-Spanish can be outlined. The ratio between the source text and the reference translation is very close to 1.

Table 20 shows the scoring results of primary systems for Spanish-to-English for the whole (EPPS+CORTES) corpus:

Task	**Site**	**NIST**	**BLEU**	**IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**
FTE	IBM	10.59	48.38	48.38	41.71	29.20	45.25	-
	IRST	10.33	47.27	47.27	42.72	30.37	45.24	-
	RWTH	10.47	47.72	47.73	42.11	29.71	44.33	-
	UKA	10.67	48.89	48.89	40.95	28.84	45.85	-
	UPC	10.38	47.35	47.36	42.82	30.55	43.94	-
	JHU	9.89	43.95	43.95	45.58	32.24	41.51	-
	NICT-ATR	10.35	46.48	46.48	43.01	30.05	42.42	-
	Translendium	9.65	41.80	41.81	48.01	34.08	40.55	-
	UDS	8.60	33.88	33.25	55.98	37.00	36.93	-
	ROVER	***10.60***	***49.05***	***49.05***	***40.99***	***29.22***	***46.60***	-
	Reverso	*8.33*	*34.45*	*34.45*	*60.45*	*47.22*	*37.27*	
	Systran	*9.35*	*39.52*	*39.54*	*48.69*	*34.96*	*37.83*	-
Verbatim	IBM	10.69	49.60	49.60	39.74	27.73	45.95	-
	IRST	10.29	47.46	47.46	41.86	29.76	45.89	-
	LIMSI	10.67	49.19	49.19	39.78	27.44	46.03	-
	RWTH	10.42	48.11	48.11	40.67	29.15	46.21	-
	UKA	10.82	49.87	49.30	38.99	27.64	45.63	-
	UPC	10.52	48.46	48.46	40.87	29.04	43.96	-
	JHU	9.81	43.17	42.95	44.91	31.74	41.95	-
	ROVER	***10.68***	***49.96***	***49.96***	***39.21***	***28.10***	***47.75***	-

	Reverso	*8.16*	*34.09*	*34.11*	*61.03*	*48.55*	*36.53*	-
	Systran	*9.48*	*40.56*	*40.56*	*46.91*	*33.31*	*37.79*	-
ASR	IBM	9.46	38.89	38.89	49.31	33.39	43.09	48.08
	IRST	9.42	38.95	38.95	49.74	34.30	43.04	47.99
	LIMSI	9.51	39.01	38.81	48.72	33.45	43.37	47.48
	RWTH	9.23	37.81	37.81	51.02	35.70	43.25	49.43
	UKA	9.54	38.65	37.77	48.32	33.70	42.92	47.18
	UPC	9.32	37.86	37.74	50.40	35.13	40.90	49.34
	JHU	8.93	34.75	33.92	52.18	36.37	40.17	51.09
	ROVER	***9.64***	***40.39***	***40.32***	***48.05***	***33.16***	***44.70***	***46.69***
	Reverso	*7.01*	*25.19*	*25.19*	*72.47*	*54.89*	*34.04*	*61.00*
	Systran	*8.31*	*30.73*	*30.73*	*57.33*	*39.63*	*34.95*	*59.87*

Table 20: Evaluation results of primary systems for the Spanish-to-English

Table 21 shows the ranking of primary systems for Spanish-to-English for the whole (EPPS+CORTES) corpus:

Task	**Site**	**NIST**	**BLEU**	**IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**
FTE	IBM	3	3	3	3	2	3	-
	IRST	7	6	6	5	6	4	-
	RWTH	4	4	4	4	4	5	-
	UKA	1	2	2	1	1	2	-
	UPC	5	5	5	6	7	6	-
	JHU	8	8	8	8	8	8	-
	NICT-ATR	6	7	7	7	5	7	-
	Translendium	9	9	9	9	9	9	-
	UDS	11	12	12	11	11	12	-
	ROVER	***2***	***1***	***1***	***2***	***3***	***1***	-
	Reverso	*12*	*11*	*11*	*12*	*12*	*11*	
	Systran	*10*	*10*	*10*	*10*	*10*	*10*	-
Verbatim	IBM	2	3	2	3	3	4	-
	IRST	7	7	7	7	7	5	-
	LIMSI	4	4	4	4	1	3	-
	RWTH	6	6	6	5	6	2	-
	UKA	1	2	3	1	2	6	-
	UPC	5	5	5	6	5	7	-
	JHU	8	8	8	8	8	8	-
	ROVER	**3**	**1**	**1**	**2**	**4**	**1**	-
	Reverso	10	10	10	10	10	10	-
	Systran	9	9	9	9	9	9	-
ASR	IBM	4	4	3	4	2	4	5
	IRST	5	3	2	5	5	5	4
	LIMSI	3	2	4	3	3	2	3
	RWTH	7	7	5	7	7	3	7
	UKA	2	5	6	2	4	6	2
	UPC	6	6	7	6	6	7	6

	JHU	8	8	8	8	8	8	8
	ROVER	***1***	***1***	***1***	***1***	***1***	***1***	***1***
	Reverso	*10*	*10*	*10*	*10*	*10*	*10*	*10*
	Systran	*9*	*9*	*9*	*9*	*9*	*9*	*9*

Table 21: Ranking of primary systems for the Spanish-to-English task

Again, the SLT ROVER system combination gets the best results for all the tasks, even if it is less marked. UKA has the best results for FTE task, IBM and UKA share the best results for the Verbatim task, and IBM, IRST and LIMSI share the best results for the ASR task. The non-TCSTAR systems (JHU, NICT-ATR, Translendium, Reverso and Systran) are really lower than the TC-STAR systems, except NICT-ATR for the FTE task which has higher score than one TC-STAR system, IRST. The same reason for the English-to-Spanish direction can explain those scores (although JHU is a statistical-based system, the other ones are phrase-based and rule-based systems).

Within the TC-STAR consortium, we can chunk the systems as previously, but differences between systems are less marked. For the FTE task, UKA is higher, then IBM, RWTH and finally IRST and UPC. For the Verbatim task, IBM is higher, next LIMSI and UKA, UPC and RWTH, while IRST is quite lower. Finally, for the ASR task, IBM, IRST and the LIMSI share the best scores, then UKA, and UPC and RWTH further.

Verbatim scores are slightly higher than FTE ones, but less than last year. Last year scores difference between the best FTE and Verbatim systems is 4.38 BLEU points while it is 0.89 for this year. However, the difference between the best FTE and ASR systems is 8.75 last year while the difference is 9.88 for this year. But regarding the reduction of the FTE/Verbatim differences, we can only conclude that ASR systems did not improve as well as Verbatim and FTE systems. As for English-to-Spanish direction, the differences has been reduced, but in a contrary way than for the English-to-Spanish direction. ROVER combination permits strongly to reduce the tasks differences: 0.08 BLEU points of difference between FTE and Verbatim and 0.96 between FTE and ASR. Since the ROVER combination gets the best results, it is obvious that the rovering is most of useful for the TC-STAR system.

About the metrics, BLEU and IBM metrics give here similar scores, contrary to the English-to-Spanish direction. The results for mPER are approximately 12% higher than for mWER which is higher than last year results (15% higher), which confirm the improvement of the word reordering.

Table 22 presents the results of secondary systems for Spanish-to-English EPPS task.

Task	Site	NIST	BLEU	IBM	mWER	mPER	WNM	AS-WER
FTE	IBM	10.65	48.80	48.80	41.53	29.01	45.50	-
	UPC	10.38	47.00	47.02	42.51	30.29	44.01	-
	JHU	9.80	43.49	43.49	46.13	32.70	41.52	-
Verbatim	IBM	10.78	50.07	50.07	39.44	27.48	46.37	-
	UPC	10.28	47.14	47.14	41.82	30.01	43.33	-
ASR	IBM	9.49	39.08	39.08	49.22	33.25	43.38	47.98
	UPC	9.24	37.55	37.55	50.58	35.39	40.29	49.51

Table 22: Evaluation results of secondary systems for the Spanish-to-English task

Two TC-STAR consortium systems and one external participant participated in this track. Here, IBM and UPC systems get higher results for FTE which seems more coherent, but results are lower for JHU and Verbatim and ASR tasks.

Table 23 shows the scoring results of primary systems for Spanish-to-English EPPS.

Task	Site	NIST	BLEU	IBM	mWER	mPER	WNM	AS-WER
FTE	IBM	10.49	51.84	51.84	38.60	27.41	50.94	-
	IRST	10.35	51.08	51.08	39.39	28.09	50.29	-
	RWTH	10.40	51.20	51.23	38.93	27.73	48.86	-
	UKA	10.57	52.49	52.37	37.83	27.10	51.17	-
	UPC	10.34	50.88	50.91	39.55	28.46	48.23	-
	JHU	9.98	47.88	47.88	41.92	29.96	46.25	-
	NICT-ATR	10.26	49.79	49.79	40.21	28.41	47.18	-
	Translendium	9.40	43.67	43.69	45.37	32.94	45.06	-
	UDS	8.66	36.72	36.09	52.51	35.01	40.87	-
	ROVER	***10.61***	***53.07***	***53.08***	***37.63***	***26.93***	***52.10***	-
	Reverso	*8.19*	*36.37*	*36.39*	*58.04*	*45.66*	*41.19*	-
	Systran	*9.12*	*41.26*	*41.28*	*46.72*	*33.81*	*41.65*	-
Verbatim	IBM	10.60	52.93	52.93	36.23	25.86	52.96	-
	IRST	10.23	50.46	50.46	38.59	27.66	51.43	-
	LIMSI	10.52	52.14	52.14	36.67	25.84	51.98	-
	RWTH	10.33	50.74	50.53	37.62	27.09	51.48	-
	UKA	10.60	52.60	51.63	36.06	26.21	51.31	-
	UPC	10.41	51.31	50.96	37.63	27.24	48.48	-
	JHU	9.86	47.32	47.22	41.21	29.58	48.94	-
	ROVER	***10.66***	***53.48***	***53.18***	***35.47***	***25.77***	***53.69***	-
	Reverso	*8.07*	*36.12*	*36.12*	*57.65*	*46.11*	*42.82*	-
	Systran	*9.23*	*41.96*	*41.96*	*44.68*	*31.91*	*43.10*	-
ASR	IBM	9.53	42.87	42.87	44.93	31.37	49.96	43.77
	IRST	9.54	42.87	42.80	45.23	31.88	48.99	43.85
	LIMSI	9.61	43.04	42.52	44.25	31.10	49.43	43.27
	RWTH	9.30	41.30	41.30	46.51	33.20	48.93	45.27
	UKA	9.59	42.07	41.33	44.10	31.66	48.49	43.12
	UPC	9.48	42.23	41.33	45.46	32.66	45.92	44.88
	JHU	9.11	39.42	38.75	48.04	33.99	47.24	46.98
	ROVER	***9.79***	***44.80***	***44.42***	***43.07***	***30.41***	***51.32***	***42.15***
	Reverso	*7.11*	*27.90*	*27.90*	*68.29*	*52.23*	*39.96*	*58.59*
	Systran	*8.31*	*33.40*	*33.40*	*53.86*	*37.70*	*40.47*	*57.55*

Table 23: Evaluation results of primary systems for the Spanish-to-English EPPS task

Table 24 shows the ranking of primary systems for Spanish-to-English EPPS.

Task	Site	NIST	BLEU	IBM	mWER	mPER	WNM	AS-WER
FTE	IBM	3	3	3	3	3	3	-
	IRST	5	5	5	5	5	4	-
	RWTH	4	4	4	4	4	5	-
	UKA	2	2	2	2	2	2	-
	UPC	6	6	6	6	7	6	-
	JHU	8	8	8	8	8	8	-
	NICT-ATR	7	7	7	7	6	7	-

	Translendium	9	9	9	9	9	9	-
	UDS	11	11	12	11	11	12	-
	ROVER	***1***	***1***	***1***	***1***	***1***	***1***	-
	Reverso	*12*	*12*	*11*	*12*	*12*	*11*	-
	Systran	*10*	*10*	*10*	*10*	*10*	*10*	-
Verbatim	IBM	2	2	2	3	3	2	-
	IRST	7	7	7	7	7	5	-
	LIMSI	4	4	3	4	2	3	-
	RWTH	6	6	6	5	5	4	-
	UKA	2	3	4	2	4	6	-
	UPC	5	5	5	6	6	8	-
	JHU	8	8	8	8	8	7	-
	ROVER	***1***	***1***	***1***	***1***	***1***	***1***	-
	Reverso	*10*	*10*	*10*	*10*	*10*	*10*	-
	Systran	*9*	*9*	*9*	*9*	*9*	*9*	-
ASR	IBM	5	3	2	4	3	2	4
	IRST	4	3	3	5	5	4	5
	LIMSI	2	2	4	3	2	3	3
	RWTH	7	7	7	7	7	5	7
	UKA	3	6	5	2	4	6	2
	UPC	6	5	5	6	6	8	6
	JHU	8	8	8	8	8	7	8
	ROVER	***1***	***1***	***1***	***1***	***1***	***1***	***1***
	Reverso	*10*	*10*	*10*	*10*	*10*	*10*	*10*
	Systran	*9*	*9*	*9*	*9*	*9*	*9*	*9*

Table 24: Ranking of primary systems for the Spanish-to-English EPPS task

Table 22 presents the results of secondary systems for Spanish-to-English EPPS task:

Task	**Site**	**NIST**	**BLEU**	**IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**
FTE	IBM	10.53	52.08	52.08	38.34	27.26	50.61	-
	UPC	10.23	50.14	50.17	39.62	28.71	47.83	-
	JHU	9.85	47.20	47.20	42.58	30.47	46.23	-
Verbatim	IBM	10.60	52.88	52.88	36.36	25.96	52.94	-
	UPC	9.98	48.58	48.58	39.73	29.08	47.55	-
ASR	IBM	9.61	43.34	43.34	44.59	31.03	50.17	43.58
	UPC	9.35	41.40	40.84	46.04	33.08	45.43	45.38

Table 25: Evaluation results of secondary systems for the Spanish-to-English EPPS task

Table 26 shows the scoring results of primary systems for Spanish-to-English CORTES.

Task	Site	NIST	BLEU	IBM	mWER	mPER	WNM	AS-WER
FTE	IBM	9.70	44.79	44.79	44.90	31.05	42.79	-
	IRST	9.39	43.39	43.39	46.14	32.72	43.38	-
	RWTH	9.58	44.18	44.18	45.36	31.74	43.12	-
	UKA	9.78	45.22	45.22	44.15	30.62	43.64	-
	UPC	9.48	43.79	43.79	46.18	32.70	42.80	-
	JHU	8.94	39.98	39.98	49.35	34.58	40.41	-
	NICT-ATR	9.51	43.08	43.08	45.88	31.73	41.34	-
	Translendium	9.05	39.89	39.89	50.72	35.25	39.33	-
	UDS	7.86	30.88	30.27	59.55	39.05	35.58	
	ROVER	***9.61***	***44.91***	***44.91***	***44.45***	***31.58***	***44.20***	-
	Reverso	*7.77*	*32.46*	*32.46*	*62.93*	*48.82*	*36.81*	-
	Systran	*8.78*	*37.76*	*37.76*	*50.72*	*36.14*	*37.68*	-
Verbatim	IBM	9.86	46.64	46.64	42.92	29.43	43.34	-
	IRST	9.48	44.84	44.84	44.81	31.66	44.40	-
	LIMSI	9.91	46.57	46.57	42.60	28.88	44.11	-
	RWTH	9.64	45.86	45.86	43.43	31.01	44.55	-
	UKA	10.08	47.45	47.22	41.64	28.93	43.93	-
	UPC	9.74	46.04	46.04	43.81	30.67	43.46	-
	JHU	8.97	39.36	39.06	48.27	33.69	39.34	-
	ROVER	***9.80***	***46.90***	***46.90***	***42.59***	***30.22***	***45.72***	-
	Reverso	*7.61*	*32.31*	*32.34*	*64.09*	*50.75*	*35.89*	-
	Systran	*8.94*	*39.33*	*39.33*	*48.93*	*34.59*	*37.37*	-
ASR	IBM	8.65	35.21	35.21	53.31	35.24	40.94	51.97
	IRST	8.58	35.42	35.42	53.85	36.52	41.45	51.71
	LIMSI	8.66	35.39	35.39	52.79	35.61	41.72	51.26
	RWTH	8.45	34.66	34.66	55.13	37.99	41.61	53.17
	UKA	8.74	35.42	34.50	52.17	35.57	41.60	50.84
	UPC	8.43	34.04	34.04	54.90	37.38	40.53	53.35
	JHU	8.05	29.98	29.35	55.95	38.56	37.58	54.81
	ROVER	***8.75***	***36.40***	***36.40***	***52.57***	***35.68***	***42.35***	***50.76***
	Reverso	*6.43*	*22.75*	*22.75*	*76.22*	*57.26*	*33.84*	*63.21*
	Systran	*7.70*	*28.32*	*28.32*	*60.45*	*41.38*	*34.97*	*61.94*

Table 26: Evaluation results of primary systems for the Spanish-to-English CORTES task

Table 27 shows the ranking of primary systems for Spanish-to-English CORTES.

Task	Site	NIST	BLEU	IBM	mWER	mPER	WNM	AS-WER
FTE	IBM	2	3	3	3	2	6	-
	IRST	7	6	6	6	7	3	-
	RWTH	4	4	4	4	5	4	-
	UKA	1	1	1	1	1	2	-
	UPC	6	5	5	7	6	5	-
	JHU	9	8	8	8	8	8	
	NICT-ATR	5	7	7	5	4	7	-
	Translendium	8	9	9	9	9	9	-

	UDS	11	12	12	11	11	12	-
	ROVER	***3***	***2***	***2***	***2***	***3***	***1***	-
	Reverso	*12*	*11*	*11*	*12*	*12*	*11*	-
	Systran	*10*	*10*	*10*	*9*	*10*	*10*	-
Verbatim	IBM	3	3	3	4	3	7	-
	IRST	7	7	7	7	7	3	-
	LIMSI	2	4	4	3	1	4	-
	RWTH	6	6	6	5	6	2	-
	UKA	1	1	1	1	2	5	-
	UPC	5	5	5	6	5	6	-
	JHU	8	8	9	8	8	8	-
	ROVER	***4***	***2***	***2***	***2***	***4***	***1***	-
	Reverso	*10*	*10*	*10*	*10*	*10*	*10*	-
	Systran	*9*	*9*	*8*	*9*	*9*	*9*	-
ASR	IBM	4	5	4	4	1	6	5
	IRST	5	2	2	5	5	5	4
	LIMSI	3	4	3	3	3	2	3
	RWTH	6	6	5	7	7	3	6
	UKA	2	2	6	1	2	4	2
	UPC	7	7	7	6	6	7	7
	JHU	8	8	8	8	8	8	8
	ROVER	***1***	***1***	***1***	***2***	***4***	***1***	***1***
	Reverso	*10*	*10*	*10*	*10*	*10*	*10*	*10*
	Systran	*9*	*9*	*9*	*9*	*9*	*9*	*9*

Table 27: Ranking of primary systems for the Spanish-to-English CORTES task

Table 28 presents the results of secondary systems for Spanish-to-English CORTES task.

Task	**Site**	**NIST**	**BLEU**	**IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**
FTE	IBM	9.79	45.42	45.42	44.81	30.81	43.66	-
	UPC	9.57	43.74	43.75	45.48	31.92	43.13	-
	JHU	8.88	39.69	39.69	49.77	34.99	40.50	-
Verbatim	IBM	10.02	47.57	47.57	42.22	28.86	44.11	-
	UPC	9.69	45.85	45.85	43.71	30.85	43.44	-
ASR	IBM	8.64	35.19	35.19	53.41	35.28	41.32	51.93
	UPC	8.42	34.18	34.18	54.70	37.48	40.12	53.22

Table 28: Evaluation results of secondary systems for the Spanish-to-English CORTES task

CORTES-EPPS comparison. For all the systems, the results from EPPS inputs are strongly higher than those from CORTES inputs. Moreover, the ranking does not vary with very few exceptions.

Even if scores are strongly different, correlation between EPPS and CORTES BLEU scores and ranks are very high, as shown in Table 29.

Task	**Scoring**	**Ranking**
FTE	97.78	97.90

Verbatim	97.03	96.36
ASR	97.55	82.84

Table 29: Pearson correlation between EPPS and CORTES scores and ranks

Thus, systems behaviour is similar for the both EPPS and CORTES track, but the domain of the data is different, which can explain the higher scores for EPPS. But all the systems get the same behaviour, even the systems which have not received any training, so the difference would be due to the intrinsic quality (in terms of vocabulary, grammar, etc.) of the test corpus, instead the lack of CORTES training data compared to EPPS data.

Automatic results for Chinese-to-English. Data statistics for Chinese-to-English source documents are the following:

- Verbatim: 21 274 words, for 917 sentences,
- ASR: 19 898 words.

Table 30 presents the scoring results of primary systems for Chinese-to-English VOA.

Task	**Site**	**NIST**	**BLEU**	**IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**
Verbatim	IRST	7.09	21.78	20.94	71.89	48.91	30.35	-
	RWTH	7.35	24.52	23.56	68.14	47.34	33.31	-
	UKA	6.47	18.56	17.95	72.08	51.74	30.07	-
	ICT	6.67	20.05	19.31	74.94	51.54	29.32	-
	NICT-ATR	6.51	18.39	17.72	73.85	52.13	29.08	-
	UDS	4.89	10.43	10.43	83.42	62.82	22.95	-
	XMU	5.61	12.39	11.97	78.04	57.80	25.42	-
	Systran	*4.30*	*6.74*	*6.74*	*91.25*	*70.63*	*23.51*	-
ASR	IRST	6.45	19.70	19.00	71.65	52.26	29.22	72.45
	RWTH	6.80	22.50	21.76	68.68	50.73	32.00	69.10
	UKA	5.82	16.49	16.01	71.39	54.61	28.39	72.50
	ICT	6.01	18.25	17.60	74.06	55.71	28.33	75.06
	XMU	5.09	11.42	11.07	76.71	59.94	24.35	78.69
	Systran	*4.08*	*6.65*	*6.65*	*87.80*	*70.72*	*23.12*	*88.83*

Table 30: Evaluation results of primary systems for the Chinese-to-English EPPS task

Table 31 presents the ranking of primary systems for Chinese-to-English VOA.

Task	**Site**	**NIST**	**BLEU**	**IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**
Verbatim	IRST	2	2	2	2	2	2	-
	RWTH	1	1	1	1	1	1	-
	UKA	5	4	4	3	4	3	-
	ICT	3	3	3	5	3	4	-

	NICT-ATR	4	5	5	4	5	5	-
	UDS	7	7	7	7	7	8	-
	XMU	6	6	6	6	6	6	-
	Systran	8	8	8	8	8	7	-
ASR	IRST	2	2	2	3	2	2	2
	RWTH	1	1	1	1	1	1	1
	UKA	4	4	4	2	3	3	3
	ICT	3	3	3	4	4	4	4
	XMU	5	5	5	5	5	5	5
	Systran	6	6	6	6	6	6	6

Table 31: Ranking of primary systems for the Chinese-to-English EPPS task

RWTH gets the best results for both Verbatim and ASR tasks. The external participants (ICT, NICT-ATR, UDS, XMU and Systran products) have various results: ICT is in the third position for Verbatim and ASR tasks, with good results and before one TC-STAR participant, UKA. NICT-ATR gets also good results and is before UKA for the Verbatim task. UDS, XMU and Systran get much lower scores.

Within the TC-STAR consortium, ranking is the same for both Verbatim and ASR task, with RWTH in the lead, very close IRST, and UKA further.

Differences between ASR and Verbatim scores have been reduced: 3.68 BLEU points between the last year best Verbatim and ASR systems instead of 2.02 BLEU points for this year.

mPER scores are approximately 17-21% higher than for mWER which is similar to last year results (16-20% higher).

Table 32 presents the scoring results of secondary systems for Chinese-to-English task:

Task	**Site**	**NIST**	**BLEU**	**IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**
Verbatim	UDS	4.37	7.98	7.97	98.41	74.33	25.17	-
	XMU	5.74	12.16	11.94	80.00	57.92	26.16	-
ASR	UDS	3.75	6.52	6.52	94.20	76.53	24.90	92.82
	XMU	5.27	11.56	11.26	78.69	60.33	25.16	79.96

Table 32: Evaluation results of secondary systems for the Chinese-to-English task

UDS only participated in secondary track, but XMU participated in both primary and secondary tracks and get better results for the secondary.

Automatic results for 2005 and 2006 systems. In order to measure the improvement of the TC-STAR systems throughout the three years of evaluation, participants are asked to provide the output of their 2006 and 2005 systems on the 2007 evaluation data. For the English-to-Spanish direction, 11 outputs have been submitted for the year 2006. For the Spanish-to-English direction, 9 outputs have been submitted for the year 2006, 3 for the year 2005. For the Chinese-to-English direction, 2 outputs have been submitted for the year 2006, 2 for the year 2005.

Table 33 presents the results of 2006 systems for English-to-Spanish task.

Task	**Site**	**NIST**	**BLEU**	**IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**

FTE	IBM-2006	9.98	49.82	49.78	39.08	30.36	50.86	-
	IRST-2006	10.22	51.36	50.63	37.80	29.70	51.49	-
	RWTH-2006	10.03	49.75	49.15	38.99	30.62	49.26	-
	UPC-2006	10.06	50.30	49.73	39.42	30.87	48.85	-
Verbatim	IBM-2006	9.79	47.35	47.21	41.11	30.88	48.71	-
	IRST-2006	9.51	45.75	45.02	43.20	33.49	49.12	-
	RWTH-2006	9.93	48.18	47.71	40.21	30.59	49.22	-
	UPC-2006	9.84	47.50	46.30	41.31	31.42	47.49	-
ASR	IBM-2006	8.55	36.23	35.63	50.82	38.54	45.77	50.07
	IRST-2006	7.40	31.70	33.20	55.22	46.36	45.42	53.43
	UPC-2006	8.41	34.72	34.20	50.54	39.85	44.42	49.74

Table 33: Evaluation results of 2006 systems for the English-to-Spanish task

Table 34 presents the results of 2005 and 2006 systems for Spanish-to-English task.

Task	**Site**	**NIST**	**BLEU**	**IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**
FTE	IRST-2005	8.96	38.00	38.01	50.95	37.91	38.75	-
	IRST-2006	10.23	46.09	46.20	44.42	31.95	42.05	-
	RWTH-2006	9.87	43.78	43.78	45.64	32.60	40.74	-
	UPC-2006	10.37	46.77	46.78	42.69	30.04	42.88	-
Verbatim	IRST-2005	8.86	38.12	38.27	52.39	40.40	38.45	-
	IRST-2006	9.91	44.45	45.75	46.22	34.85	41.71	-
	LIMSI-2006	9.47	39.54	38.93	45.65	32.66	38.80	-
	RWTH-2006	10.38	46.92	46.92	41.46	28.83	44.23	-
	UPC-2006	10.37	47.37	47.37	41.74	29.55	41.75	-
ASR	IRST-2005	8.12	30.65	30.65	56.63	40.05	37.83	55.21
	IRST-2006	8.55	33.88	33.88	56.87	41.38	40.86	53.78
	UPC-2006	9.24	37.18	36.91	50.57	35.19	40.05	49.39

Table 34: Evaluation results of 2005-2006 systems for the Spanish-to-English task (EPPS&CORTES)

Table 35 presents the results of 2005 and 2006 systems for Spanish-to-English EPPS task.

Task	**Site**	**NIST**	**BLEU**	**IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**
FTE	IRST-2005	8.97	40.84	40.86	46.86	34.65	43.20	-
	IRST-2006	10.23	49.60	49.61	40.20	28.43	47.04	-
	RWTH-2006	9.84	47.15	47.15	42.59	30.78	45.61	-
	UPC-2006	10.32	50.26	50.30	39.53	28.14	47.15	-
Verbatim	IRST-2005	8.88	41.06	41.09	46.84	35.63	44.12	-
	IRST-2006	10.01	48.68	48.49	40.52	29.78	48.55	-

	LIMSI-2006	9.32	41.36	40.79	43.10	31.77	44.30	-
	RWTH-2006	10.23	49.55	49.55	38.34	27.15	49.73	-
	UPC-2006	10.18	49.76	49.76	39.36	28.30	47.44	-
ASR	IRST-2005	8.30	34.32	34.32	52.04	37.76	42.73	50.96
	IRST-2006	9.13	39.66	39.06	47.74	34.16	46.72	44.80
	UPC-2006	9.31	40.88	40.45	46.58	33.18	45.41	45.87

Table 35: Evaluation results of 2005-2006 systems for the Spanish-to-English task (EPPS)

Table 36 presents the results of 2005 and 2006 systems for Spanish-to-English CORTES task:

Task	**Site**	**NIST**	**BLEU**	**IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**
FTE	IRST-2005	8.19	34.95	34.41	55.16	41.26	37.49	-
	IRST-2006	9.24	41.71	42.41	48.74	35.56	40.66	-
	RWTH-2006	9.02	40.34	40.34	48.76	34.47	39.77	-
	UPC-2006	9.48	43.19	43.19	45.93	31.99	42.09	-
Verbatim	IRST-2005	8.13	35.29	35.42	57.42	44.73	36.87	-
	IRST-2006	8.86	39.99	42.73	51.39	39.43	39.42	-
	LIMSI-2006	8.87	37.87	37.23	47.95	33.47	37.04	-
	RWTH-2006	9.66	44.58	44.58	44.28	30.36	42.32	-
	UPC-2006	9.68	45.23	45.23	43.89	30.68	41.08	-
ASR	IRST-2005	7.39	27.35	27.36	60.79	42.12	37.11	59.01
	IRST-2006	8.20	32.30	31.96	55.61	38.34	40.46	52.40
	UPC-2006	8.46	33.81	33.67	54.20	37.05	39.58	52.54

Table 36: Evaluation results of 2005-2006 systems for the Spanish-to-English task (CORTES)

Table 37 presents the results of 2005 and 2006 systems for Chinese-to-English task:

Task	**Site**	**NIST**	**BLEU**	**IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**
Verbatim	IRST-2005	5.56	13.27	12.94	79.87	59.33	27.39	-
	IRST-2006	6.36	16.98	16.33	75.33	53.60	29.00	-
ASR	IRST-2005	5.08	12.18	11.88	78.85	61.85	26.59	78.03
	IRST-2006	5.87	15.68	15.09	74.82	56.03	28.35	75.87

Table 37: Evaluation results of 2005-2006 systems for the Chinese-to-English task

Figure 7 reflects the improvement of the systems for the English-to-Spanish direction (as well as the general trend of the BLEU scores). Verbatim and FTE results are improved very slightly (between 1.75% and 5.94% of improvement), except for the Verbatim results from IRST (10.49% of improvement). For FTE task, best improvements are from RWTH and UPC (resp. 5.51% and 5.94%). For Verbatim task,

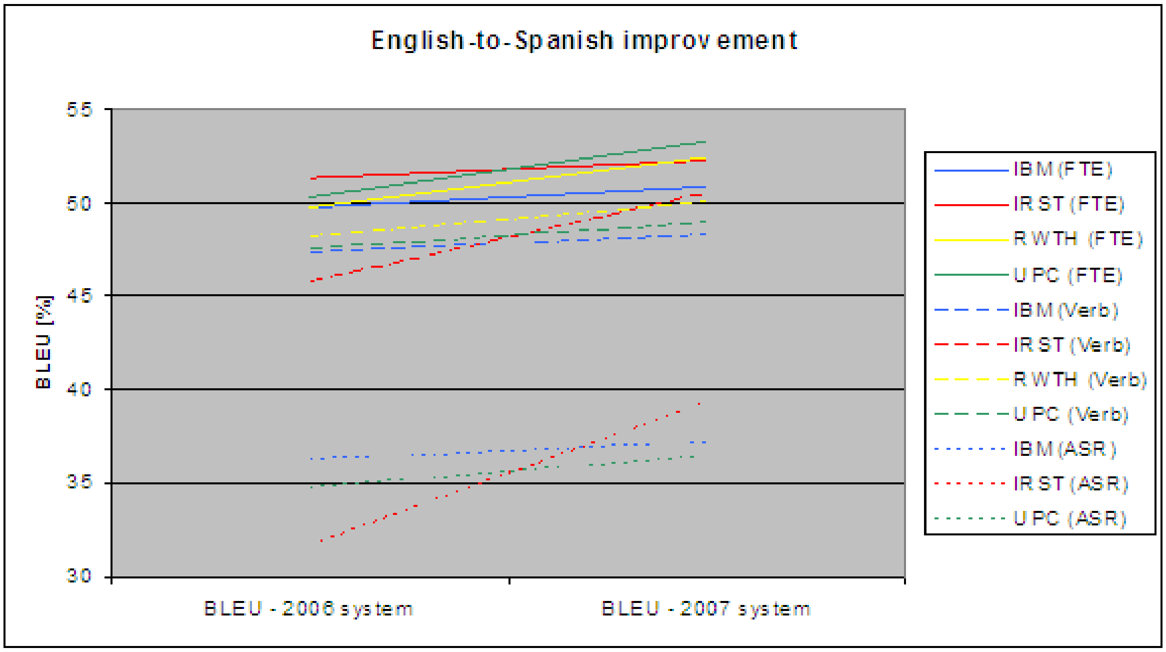


Figure 7: English-to-Spanish improvement.

IRST clearly get the best improvement while each of the three other systems has the same kind of enhancement. For the ASR task, UPC and IBM improve slightly (resp. 2.54% and 4.93%) while IRST still get a strong improvement (24.04% of improvement).

Figure 8 shows the improvement for Spanish-to-English systems, including EPPS and CORTES data. Only IRST submitted the 2005 system outputs, so we can observe the three years improvement only for that site. For that system, curves follow different trends according to the task. The improvement is rather good in 2006 for the ASR task (10.54%), but still increases in 2007 (for a whole of 27.08% of improvement). For the FTE and Verbatim tasks, improvement is high in 2006 and goes flat in 2007 (21.29% then 2.56%). RWTH system gets high improvement from 2006 to 2007 (9%) within the FTE task, and LIMSI gets a large improvement within Verbatim task (24.51%). Other systems get slightly improvements (between 1.24 and 2.56% of improvement).

In a general trend, improvements are slightly higher for the EPPS data than for the CORTES data, but the slope of the curves are quite similar and there is no noticeable differences between both Figure 9 and Figure 10.

Figure 11 presents the improvements for the 2005 and 2006 IRST systems which are the only one submitted for the Verbatim and ASR tasks in addition to the 2007 system. As we can observe, improvements are large and more significant than for the other language directions. From 2005 to 2007, system improves of 64.13% for the Verbatim task and 61.74% for the ASR task. Moreover, Verbatim and ASR results have similar improvements.

Notice that for all the directions and tasks no systems get lower or identical results throughout the evaluation campaigns.

3.6 Data analysis

3.6.1 Statistical analysis of the evaluation metrics

In Table 38 we present the metrics correlations. The used metrics to compute the Pearson correlation scores are BLEU, IBM, WNM and mPER (as we see in the first evaluation [8] that mWER and mPER

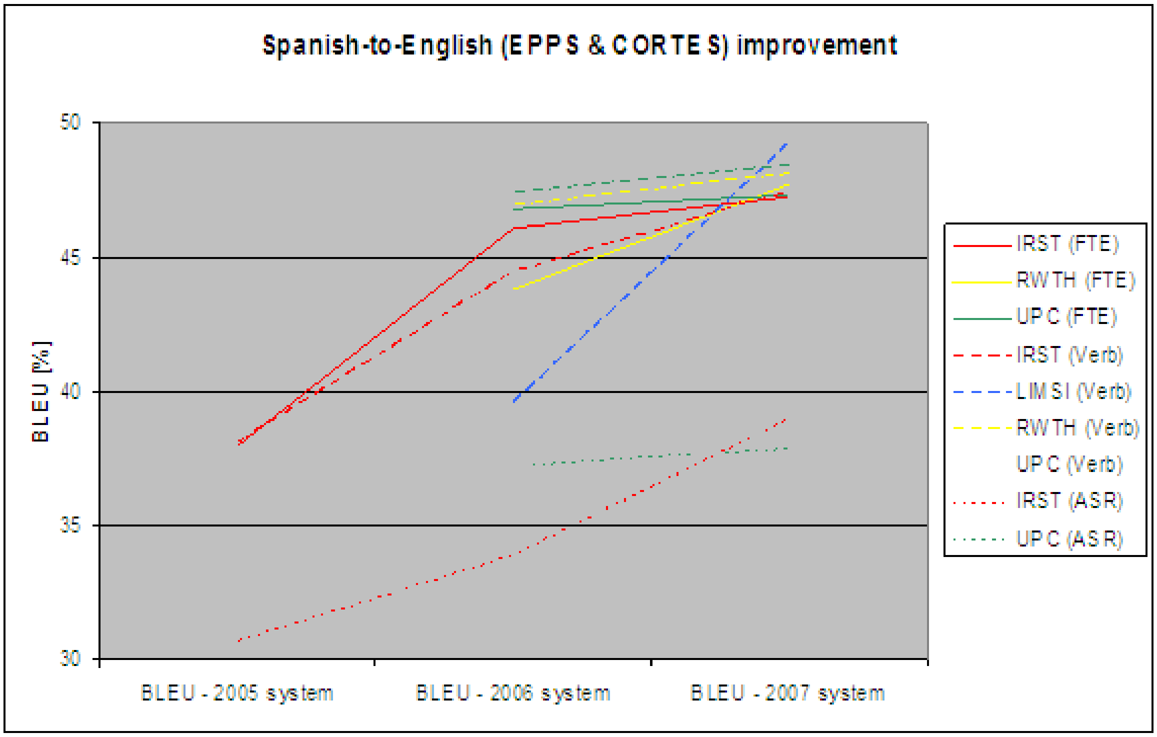


Figure 8: Spanish-to-English (EPPS&CORTES) improvement

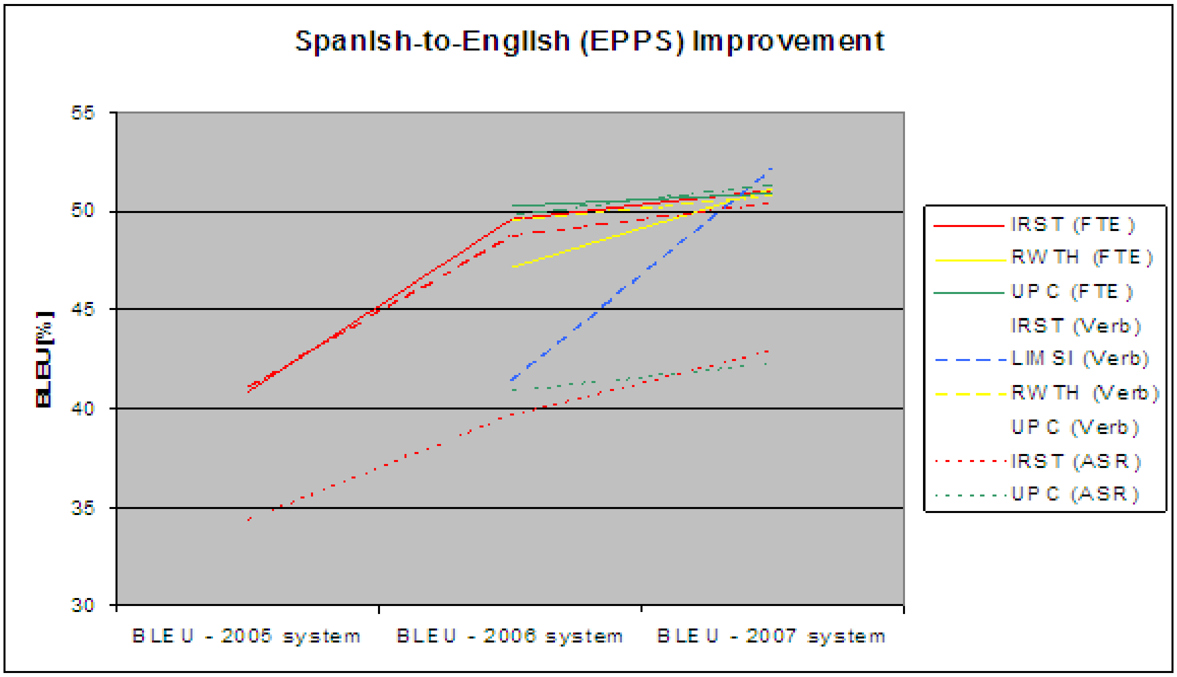


Figure 9: Spanish-to-English (EPPS) improvement

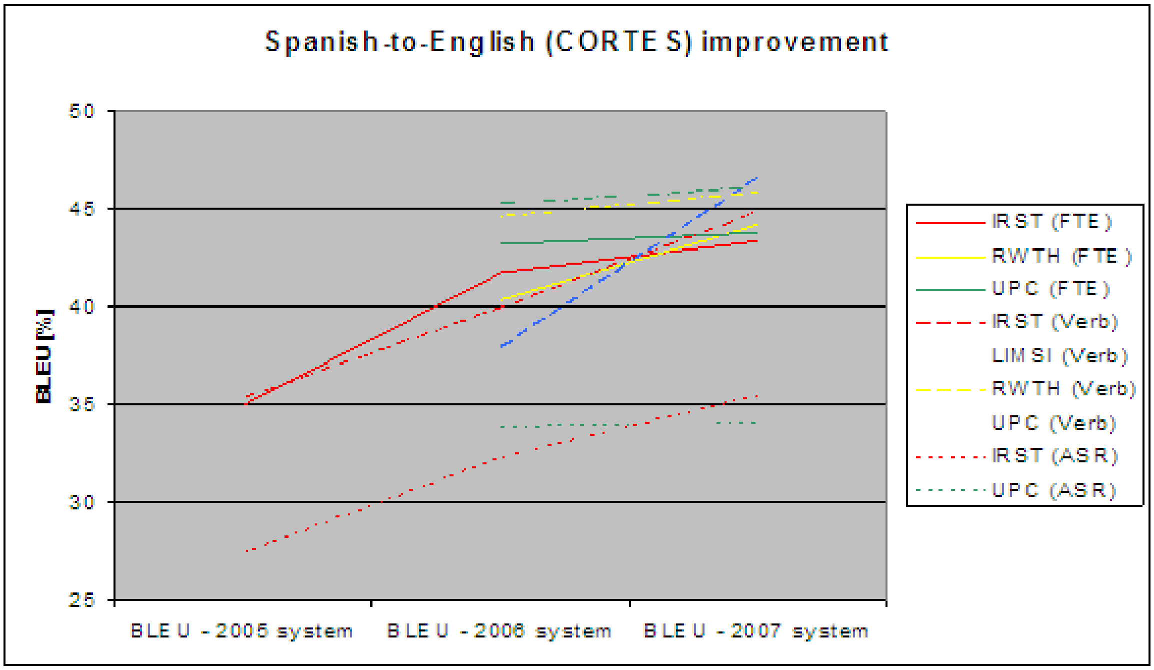


Figure 10: Spanish-to-English (CORTES) improvement

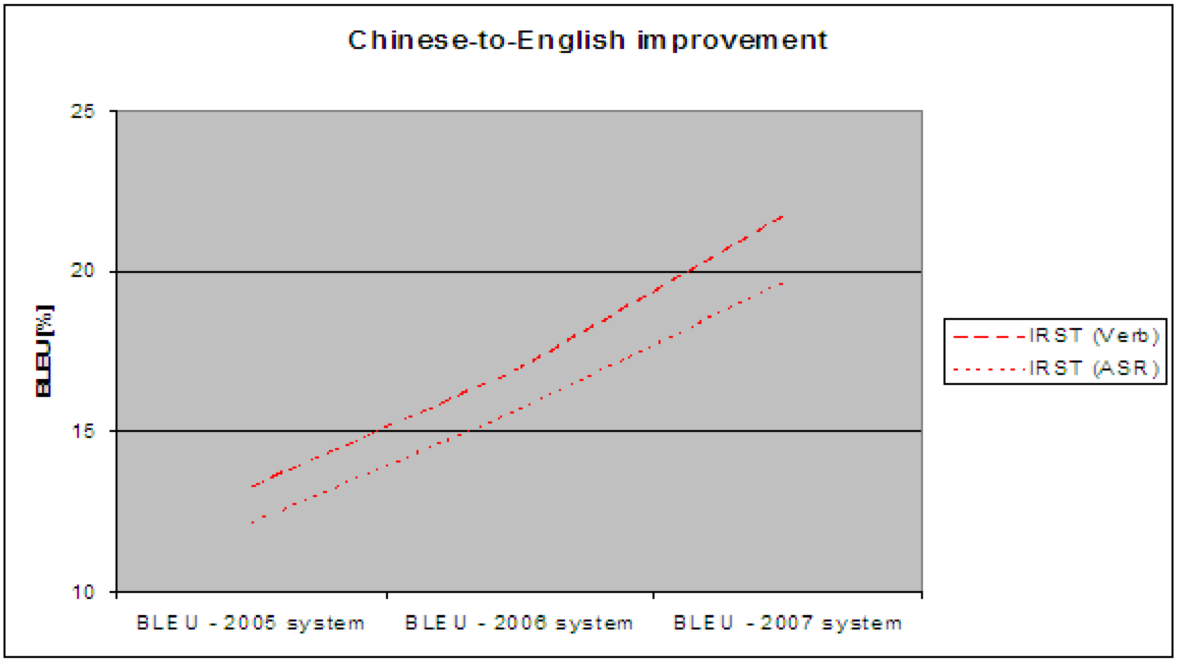


Figure 11: Figure 9: Chinese-to-English improvement

metrics are strongly correlate).

Metric	En->Es			Es->En			Zh->En	
	ASR	Text	Verb	ASR	Text	Verb	ASR	Verb
BLEU vs IBM	99.94	99.90	99.90	99.74	99.96	99.94	99.99	99.99
BLEU vs mPER	97.93	97.72	98.19	94.30	88.77	94.10	97.37	96.09
BLEU vs WNM	99.25	98.93	97.01	96.33	96.18	95.03	96.75	97.81
IBM vs mPER	97.56	97.50	98.06	93.34	87.88	93.96	97.28	95.95
IBM vs WNM	99.55	99.24	96.87	95.96	95.88	95.22	96.58	98.02
mPER vs WNM	96.57	96.56	94.41	84.42	81.87	83.49	92.10	91.29

Table 38: Pearson correlation between metrics scoring

In Table 39 shows how many systems get a different rank if the performance measure is exchanged.

Metric	En->Es			Es->En			Zh->En	
	ASR	Text	Verb	ASR	Text	Verb	ASR	Verb
BLEU vs IBM	2	2	0	6	0	2	0	0
BLEU vs mPER	3	4	5	4	7	2	2	0
BLEU vs WNM	4	6	4	4	3	6	2	4
IBM vs mPER	3	6	5	6	7	4	2	0
IBM vs WNM	2	4	4	4	3	6	2	4
mPER vs WNM	5	7	8	5	9	7	0	4

Table 39: Number of systems with a different rank when comparing two metrics

All the metrics are strongly correlated, more than for the second evaluation. For the second year, the average correlation is 94.09 instead of 95.85 for this year. The average difference of rank is 4.58 instead of 3.73 for this year.

3.6.2 Meta-evaluation of the metrics

The automatic metrics are compared to the human evaluation results. The meta-evaluation considers only the English to Spanish direction since the human evaluation is done on this direction only. For that we compute the correlations between the automatic metrics' scores and fluency/adequacy scores.

Metrics	ASR scoring	Text scoring	Verb scoring	2006 scoring
BLEU vs. Fluency	98.16	86.68	92.93	96.7
IBM vs. Fluency	98.47	86.71	92.51	97.03
mPER vs. Fluency	94.87	78.1	85.62	92.65
WNM vs. Fluency	98.97	87.85	94.34	89.32
BLEU vs. Adequacy	97.26	84.23	93.83	99.14
IBM vs. Adequacy	97.46	84.13	93.46	98.78
mPER vs. Adequacy	96.57	73.74	87.14	96.31
WNM vs. Adequacy	98.48	81.19	89.36	91.69

Table 40: Meta-evaluation of the automatic metrics

The correlations and distances are quite good except for FTE for which correlations are around 85%. As already observed in the second year evaluation, correlations are higher for the ASR and the Verbatim than for the FTE somehow correlation seems higher when translations have lower quality. The FTE scores are better and the correlations are lower than for the Verbatim which has lower scores but better correlations, etc.

3.6.3 Automatic evaluation of the English-to-Spanish human subset

We do an automatic scoring of the English-to-Spanish evaluation subset used for the human evaluation (see 3.5.1, in order to check whether the subset is representative. Pearson correlations have been computed for each metric involved.

Task	**Site**	**NIST**	**BLEU**	**IBM**	**mWER**	**mPER**	**WNM**
FTE	IBM	9.13	50.59	50.29	38.55	30.54	48.36
	IRST	9.35	51.36	50.65	36.77	29.37	49.31
	RWTH	9.42	52.10	51.58	36.71	29.02	49.30
	UKA	9.52	53.57	52.47	35.53	28.21	49.41
	UPC	9.40	52.49	51.79	36.43	29.39	49.11
	UDS	8.19	41.77	41.61	45.38	35.09	44.08
	ROVER	9.47	53.39	53.39	35.83	28.45	50.90
	Reverso	7.64	35.82	35.82	50.00	40.44	40.50
	Systran	7.76	35.99	36.00	46.32	37.17	40.19
Correlation		**99.88**	**99.77**	**99.76**	**99.89**	**99.93**	**99.72**
Verbatim	IBM	8.95	47.87	47.87	41.20	31.47	46.35
	IRST	9.39	49.94	49.04	38.54	29.64	48.14
	LIMSI	9.35	50.60	50.10	38.85	29.53	47.10
	RWTH	9.27	49.92	49.23	39.30	30.40	48.04
	UKA	9.40	50.31	49.54	37.85	29.35	47.62
	UPC	9.05	47.55	47.31	40.90	31.35	46.69
	UDS	7.63	36.41	36.43	52.77	39.25	42.54
	ROVER	9.47	51.79	50.83	37.44	29.24	48.81
	Reverso	7.55	34.61	34.63	52.64	41.63	39.03
	Systran	7.55	34.01	34.02	49.10	38.45	38.32
Correlation		**99.92**	**99.89**	**99.84**	**99.82**	**99.90**	**99.84**
ASR	IBM	7.83	35.57	35.05	51.49	39.03	42.20
	IRST	8.24	38.36	37.96	47.56	37.59	45.79
	LIMSI	8.13	37.30	36.66	48.23	38.20	43.84
	RWTH	8.13	38.81	37.92	49.39	38.03	43.95
	UKA	7.91	35.53	35.28	47.95	39.11	44.45
	UPC	7.89	35.29	34.88	50.30	39.34	42.52
	ROVER	8.37	39.50	39.18	46.02	36.88	45.57
	Reverso	6.50	23.84	23.87	63.93	50.84	35.14
	Systran	6.51	23.89	23.48	60.59	47.33	34.62
Correlation		**99.96**	**99.91**	**99.93**	**99.94**	**99.93**	**99.42**

Table 41: Automatic evaluation of the human evaluation subset

All the correlations are up to 99% and so we can conclude the subset is well representative of the whole corpus.

3.6.4 Impact of ASR errors

In this section we try to estimate the impact of speech recognition errors on the SLT results for the ROVER combination system.

To obtain Figure 12 and 13, we have computed the SLT-mWER as a function of the ASR-WER (curves with triangles) for the systems which participate to the English-to-Spanish and to the Spanish-to-English evaluation. For each system it shows the result obtained on the same data but by using the Verbatim input which can be considered as a perfect automatic transcription (i.e. the ASR-WER is equal to zero). It shows the trend curves for the both kind of data too.

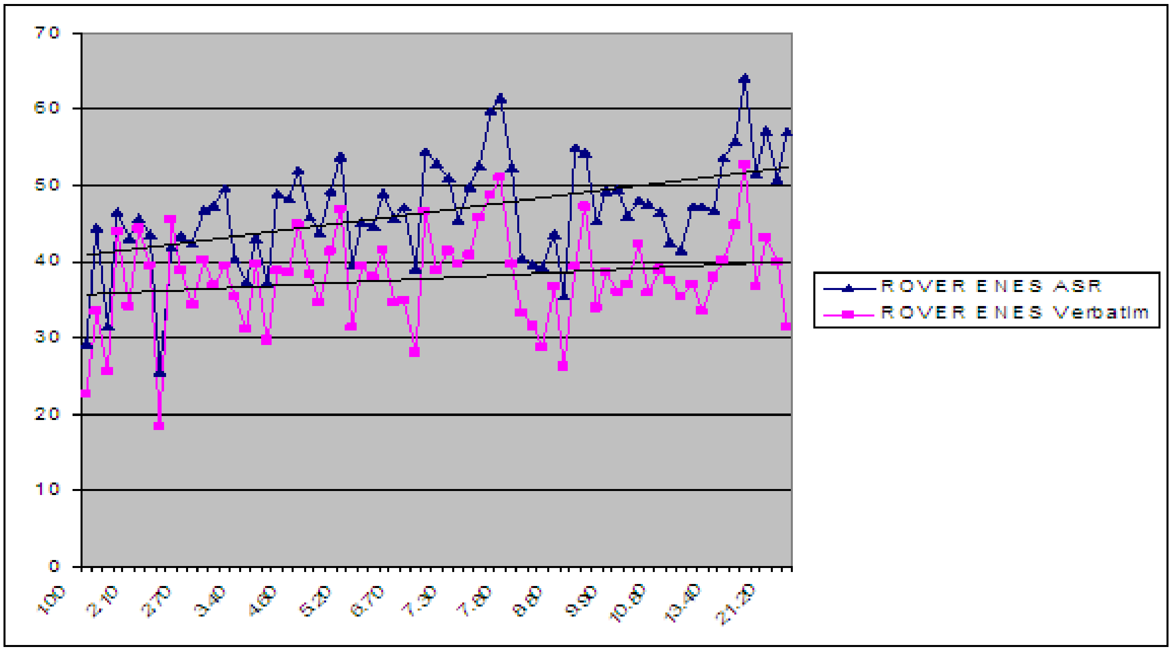


Figure 12: mWER-SLT as a function of WER-ASR for the English-to-Spanish EPPS task

Both ASR and Verbatim curves behave in a very similar manner and mWER results are worst taking into account the translation of the ASR output. The trend curves underline a slightly improvement for the SLT systems according to the improvement of the ASR output. That trend is less important for the Verbatim translation, but is present nevertheless, which shows up some sentences not so easily intrinsically to translate.

4 TTS evaluation

For the 3^{rd} text-to-speech (TTS) evaluation, TC-STAR partners decided to focus on the evaluation of the global TTS systems (no TTS component evaluation as last year) and on the voice conversion tasks. A new voice conversion task was defined, based on found data (target voices are extracted from the EPPS recordings).
For more information, you can refer to the TC-STAR Deliverable D8 [1].

4.1 Tasks and languages

The TTS evaluation comprises 8 different tasks and involves 3 languages: English, Spanish and Chinese (Mandarin). The TCSTAR-TTS 2007 evaluation tasks were:

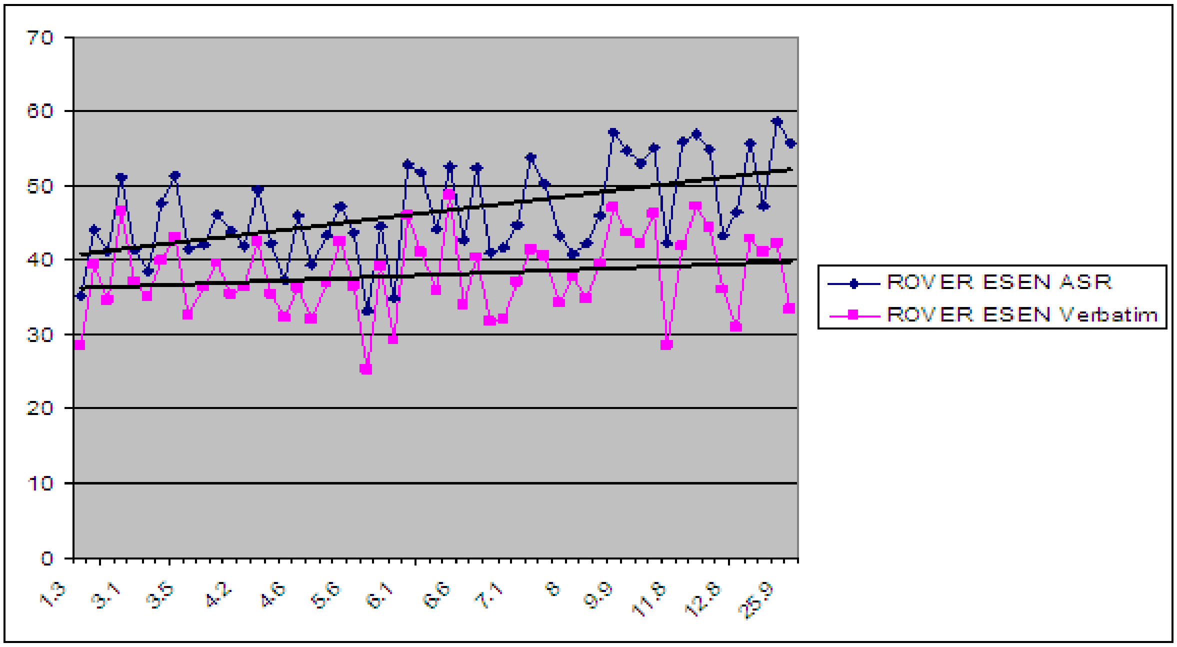


Figure 13: mWER-SLT as a function of WER-ASR for the Spanish-to-English EPPS task

- **Task S1** (TTS System): MOS tests with 10 evaluation criteria
  Languages: English, Spanish and Chinese.

- **Task S2** (TTS System): Evaluation of intelligibility in the translation scenario.
  Languages: English and Spanish.

- **Task IVC1** (Intra-lingual Voice Conversion): Comparison of speaker identities in the context of Intra-lingual Voice Conversion (IVC).
  Languages: English and Spanish.

- **Task IVC2** (Intra-lingual Voice Conversion): Evaluation of overall speech quality (MOS test) in the context of Intra-lingual Voice Conversion (IVC).
  Languages: English and Spanish.

- **Task CVC1** (Cross-lingual Voice Conversion): Comparison of speaker identities in the context of Intra-lingual Voice Conversion (CVC).
  Language: Spanish (conversion direction: Spanish-to-English).

- **Task CVC2** (Cross-lingual Voice Conversion): Evaluation of overall speech quality (MOS test) in the context of Cross-lingual Voice Conversion (CVC).
  Language: Spanish (conversion direction: Spanish-to-English).

- **Task fCVC1** (Found Data-based CVC): Comparison of speaker identities in the context of Cross-lingual Voice Conversion using found data (fCVC).
  Language: Spanish (conversion direction: Spanish-to-English).

- **Task fCVC2** (Found Data-based CVC): Evaluation of overall speech quality (MOS test) in the context of Cross-lingual Voice Conversion using found data (fCVC).
  Language: Spanish (conversion direction: Spanish-to-English).

The 8 TTS evaluation tasks and the corresponding evaluation methods and metrics are described in the following sections. This year, all tasks are evaluated through subjective tests. Some information on how the subjective tests are carried out is given in section 4.4

4.1.1 Evaluation of the whole TTS System (S1)

Each subject listened to N synthesised sentences. Subjects are asked to rate a sentence according to the following categories, proposed by the ITU.P85 recommendations (see [7]).

For each sentence they listen to, the evaluators are asked a series of 10 questions, to which they have to answer using 5-point scales.

The questions are detailed in Annex C.

The average score in each category is computed for each system.

4.1.2 Evaluation of intelligibility of the TTS System in the translation scenario (S2)

Participants have to synthesise sentences extracted from the output of the ASR+SLT system. These sentences contain some recognition and/or translation errors.

During the evaluation, subjects are asked to listen to N synthesised sentences. After having listened to each sentence, they have to write down word-by-word what they have just heard.

The WER (Word Error Rate) is computed for each system. It is the percentage of words that the subject did not correctly transcribe. This percentage is computed using the original text as a reference.

4.1.3 Voice conversion: comparison of speaker identities (IVC1 and CVC1)

Since TC-STAR aims at translating speech from one language to another, it is important to assess how good the translated voice is, i.e. how “close” it is to the original voice.

Voice Conversion (VC) consists in converting a sentence pronounced by a natural voice *A* (source voice) to the same sentence pronounced by a synthesised voice *B* (target voice).

In the case of intra-lingual voice conversion (IVC) voices A and B use the same language. In the case of cross-lingual voice conversion (CVC) voices A and B use different languages. The final goal of CVC is to convert the voice generated by the TTS, so that it is close to the voice of the person who speaks in the original language.

The conversion evaluation consists in comparing a sentence pronounced by the natural target voice *B* with the same sentence pronounced by the synthesised voice *B*. For different pairs of voices, subjects are asked to judge if the 2 voices come from the same person.

The evaluators are asked whether the two speakers are identical or not. 3 kinds of comparison are made:

- target voice versus transformed (converted) voice,
- target voice versus source voice (baseline result),
- target voice versus the same target voice (baseline result).

Of course, the evaluators always ignore the origin of the spoken sentences they listen to.

In the CVC case, the language of training data for speaker B (target) is different from the language of speaker A (source). However, the evaluation data for speaker B (target) happens to be bilingual.

The listeners compare the transformed data (modification of source A) with the voice of speaker B (target) in the same language. So, for the judges, the IVC and CVC tests are exactly the same (comparison of pairs of sentences spoken in the same language). Only the training data is different.

This year, the CVC task is only done in the Spanish-to-English direction.

Example: the Spanish data for speaker A is modified to sound like speaker B (target). In the case of IVC, we have training data for speaker B in Spanish. In the case of CVC, we can only use English data for speaker B. However in both cases, the judges listen to the transformed voice (in Spanish) and to the target voice B, also in Spanish.

The evaluators received the following instructions.

“We are analysing differences of voices. For this reason, you are asked to identify if two samples come from the same person or not. Please, do not pay attention to the recording conditions or quality of each sample, only the identity of the person. So, for each pair of voices, do you think they are”:

1. Definitely different
2. Probably different
3. Not sure
4. Probably identical
5. Definitely identical

The average comparison score is computed for each voice conversion system in each conversion direction.

4.1.4 Voice conversion: evaluation of overall speech quality (IVC2 and CVC2)

Subjects are asked to evaluate the overall quality of the converted voices. In this task, the conversion is not evaluated, only the quality of the resulting synthesised voices.

The evaluators are asked to rate the sentences they listen to as:

1. Bad
2. Poor
3. Fair
4. Good
5. Excellent

The average voice quality score is computed for each voice conversion system.

4.1.5 Global voice conversion score

This year, a new metric is introduced. It is the average between the VC1 and VC2 scores. Its aim is to reflect the better compromise between voice conversion precision and voice quality.

The global voice conversion score of each system is the average between its VC2 score and its “mean VC1 score”.

The “mean VC1 score” is the average VC1 score in all conversion directions for that system.

4.1.6 Voice conversion based on found data (fCVC1 and fCVC2)

This is a new task introduced in this year's evaluation. The goal is to change the identity of a TTS output in Spanish to be ”similar” to the voice of a real English politician.

Target voices
Audio excerpts of 2 different male English politicians are chosen by ELDA from the EPPS recordings. The 2 politicians are native English speakers. The voices are chosen to be clearly different.

Source voices
ELDA provided the Spanish translations of the excerpts transcriptions (verbatim) to IBM, who synthesised them using a Spanish TTS male voice.

Evaluation fCVC1
Participants converted the source voice synthesised by IBM to make it similar to one of the target voices.

Evaluators are native Spanish speakers, they have to compare voice pairs:

- one converted voice (synthesised voice, in Spanish)
- and one target voice (real politician voice, in English).

They are asked if both voices sound like they are coming from the same speaker (although not using the same language).
The evaluators received the following instructions:

“We are analysing differences of voices. For this reason, you are asked to identify if two samples come from the same person or not. Please, do not pay attention to the language, the recording conditions or the quality of each sample. Just focus on the identity of the person. So, for each pair of voices, do you think they are”:

1. Definitely different
2. Probably different
3. Not sure
4. Probably identical
5. Definitely identical

Evaluation fCVC2

Similar to the IVC2 and CVC2 evaluations previously described.
The same metrics as above are computed: VC1, VC2 and global voice conversion scores.

4.2 Language resources

Data sets in English and Spanish are produced using EPPS material (Final Text Edition (FTE), verbatim transcriptions, and audio recordings), ASR and SLT outputs.
Data sets in Chinese consist in “863 program data” material: TTS evaluation corpus for National High-Tech program 863 TTS evaluation in 2003. (ref 2003-863-002. Copyright ChineseLDC [2]).

4.2.1 Training data

The training data is developed by the TC-STAR partners as described in D8 (see [1]). This data is used for voice conversion and complete TTS system.
For VC, only the C33 corpus is used (see [1]). For CVC, the English-Spanish data is used.
For the complete TTS system, external partners (and also IBM for Mandarin) used their own training data.

4.2.2 Development and evaluation data

The development set is used for tuning and preparing the system to the evaluation task. Therefore, development data is required to be of the same nature and format as data to be used for the evaluation.

Test data are of the same nature and format as development data.
Development and test data sets are detailed in section 9 The evaluation corpora are subsets of the whole data sets received by the participants (the "Inputs"). Each participant processed the whole data and sent their results back to ELDA. ELDA performed the evaluations using the evaluation subsets only.

4.3 Schedule

The following schedule was respected. The TTS run took place from the 16th to the 22nd of February, 2007. Subjective tests were conducted from the 9th to 16th of March, 2007. Scorings and evaluation results were released on the 22nd of March, 2007.

4.3.1 Participants and submissions

There are 6 participating sites to the TTS evaluation:

- 4 from the TC-STAR consortium:
  - IBM (IBM)
  - Nokia (NOK)
  - Siemens (SIE)
  - Universitat Politècnica de Catalunya (UPC)
- 2 external participants:
  - Chinese Academy of Science (CAS)
  - Verbio (VER)

There are 46 submissions in total (18 in English, 26 in Spanish and 2 in Chinese)
Participants and submissions are reported in 4.3.1.
For each evaluation task, ELDA selected a portion of the submitted audio files to form the different evaluation data subsets.

4.4 Subjective test settings

Subjective tests are carried out via the web. An access to high-speed Internet connection and good listening material are required. The duration of the tests for each language is about 1 hour (20 minutes for Chinese).

	TTS system		Voice conversion			TOTAL
	S1	S2	IVC	CVC	Found Data CVC	
IBM	2 En 2 Es	2 En 2 Es	2 En 2 Es	1 Es		13
NOK	1 Zh		1 En			2
SIE	2 En	2 En	2 En 2 Es	2 Es	2 Es	12
UPC	2 Es	2 Es	1 En 1 Es	1 Es	1 Es	12
CAS	1 Zh					1
VER	3 Es	3 Es				6
Total	6 En 7 Es 2 Zh	6 En 7 Es	6 En 5 Es	4 Es	3 Es	46

Table 42: TTS participants and submissions

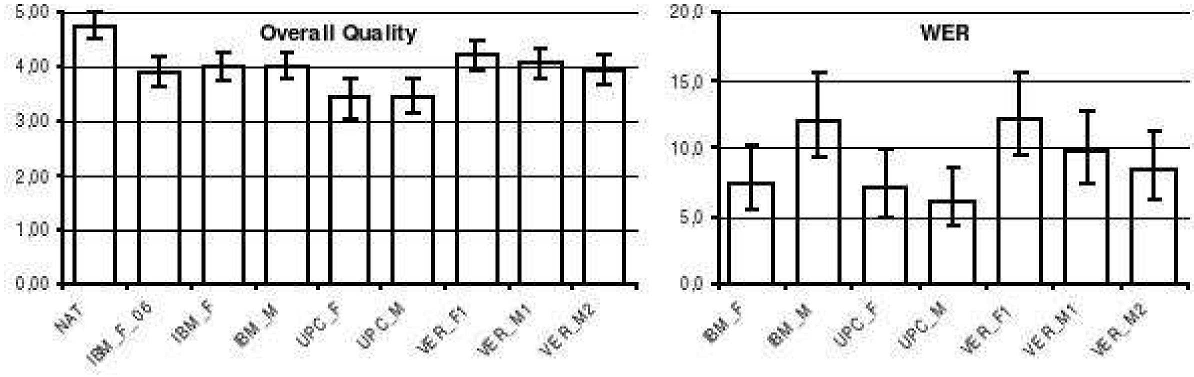


Figure 14: Overall Quality (S1) and WER (S2) results with intervals of confidence (English).

The following sections provide some details about the TTS human evaluations for English, Spanish and Chinese.

A total number of 20 judges were recruited and paid to perform the English subjective tests. They are 18 to 40 years old native English speakers with no known hearing problem. No one is a speech synthesis expert. More details are given in Table 81 in Annex C.

A total number of 20 judges were recruited and paid to perform the Spanish subjective tests. They are 18 to 40 years old native Spanish speakers with no known hearing problem. No one is a speech synthesis expert. More details are given in Table 82.

A total number of 11 judges were recruited and paid. They are 18 to 40 years old native Mandarin Chinese speakers with no known hearing problem. No one is a speech synthesis expert. More details are given in Table 83.

4.5 Evaluation results

4.5.1 Results for English

TTS component (S1, S2). The results are reported in Table 43 and Figure 14. Only the overall quality test results are reported here. The intervals of confidence are also reported: the interval of confidence (at 95%) for S1, and the Wilson score interval (at 95%) for S2.

Section 9 provides a more detailed presentation of these results, including the 10 judgement categories of task S1.

TTS Component Evaluation

System	**S1 (Overall Quality)**			**S2**						
	Score	**IC**	**Rank**	**WER(%)**	**WSI**					**Rank**
NAT	4.59	±0.20	*1*	-			-			-
IBM_F_06	3.00	±0.41	*4*	-			-			-
IBM_F	3.42	±0.37	*3*	12.8	[	10,1	-	16,1	]	*3*
IBM_M	3.49	±0.30	***2***	12.4	[	9,7	-	15,7	]	*2*
SIE_F	2.31	±0.35	*7*	14.8	[	11,9	-	18,2	]	*5*
SIE_M	1.58	±0.26	*8*	22.2	[	18,7	-	26,1	]	*6*
UPC_F	2.86	±0.36	*5*	8.7	[	6,5	-	11,6	]	***1***
UPC_M	2.74	±0.29	*6*	14.5	[	11,6	-	18,0	]	*4*

Table 43: Results of the TTS component evaluation tasks S1 and S2 (English)

Legend:

- NAT Natural voice, used as top-line in subjective tests.
- IBM_F_06 Female voice submission made by IBM last year (re-evaluated this year),
- IBM_F/M IBM submission using female / male voices,
- SIE_F/M Siemens submission using female / male voices,
- UPC_F/M UPC submission using female / male voices,
- WER Word Error Rate,
- IC Interval of Confidence (at 95%),
- WSI Wilson Score Interval (at 95%).

The best Overall Quality score is obtained by the male and female IBM voices, which also yielded the 2^{nd} and 3^{rd} lowest word error rates in test S2, after the UPC female voice.
In terms of Overall Quality score, the IBM female voice (IBM_F) perform slightly better as last year (IBM_F_06).

Voice conversion (VC1, VC2). Results of the comparative tests (VC1) and the overall quality judgement tests (VC2) are reported in Table 44 and 45. There is no cross-lingual voice conversion (English to Spanish) for English this year.

Intra-lingual Voice Conversion: Comparison test VC1 (*Scoring: 5>1*)								
Conversion System	**Conversion F(75)→F(76)**		**Conversion F(75)→M(79)**		**Conversion M(80)→F(76)**		**Conversion M(80)→M(79)**	
	Score	**Rank**	**Score**	**Rank**	**Score**	**Rank**	**Score**	**Rank**
IVC_IBM1	2.10	*5*	2.56	*6*	1.92	*6*	2.71	*3*
IVC_IBM2	3.20	***2***	3.00	*4*	2.57	***2***	2.25	*7*
IVC_NOK	2.67	*3*	2.50	*7*	1.60	*7*	1.89	*8*
IVC_SIE1	1.64	*9*	1.50	*8*	1.44	*8*	2.40	*5*
IVC_SIE1_06	2.00	*7*	2.80	*5*	2.56	*3*	2.40	*5*
IVC_SIE2	2.62	*4*	3.67	***2***	2.33	*4*	2.60	*4*
IVC_UPC	2.10	*5*	3.67	***2***	2.17	*5*	3.57	***2***
SRC-TGT	1.90	*8*	1.00	*9*	1.00	*9*	1.63	*9*
TGT-TGT	4.42	*1*	4.21	*1*	4.42	*1*	4.21	*1*

Table 44: Results of the intra-lingual voice conversion comparison tests VC1 for English

Legend:

- IVC Intra-lingual voice conversion (English to English) There are 6 IVC submissions: 2 from IBM (IVC_IBM1+2), 1 from NOK IVC_NOK), 2 from Siemens (IVC_SIE1+2) and 1 from UPC (IVC_UPC).
- IVC_SIE1_06 The submission made by Siemens last year is re-evaluated.
- F(*n*) Female voice number *n*
- M(*n*) Male voice number *n*
- *A*→*B* Conversion from voice *A* (source) to voice *B* (target). Target voice *B* and source voice *A* are English voices. The *A*→*B* conversion consists in synthesising voice *B* from the natural voice *A*. The conversion evaluation score results from comparing the natural voice *B* with the synthesised voice *B*.
- SRC-TGT This result corresponds to the comparison between the natural source voice and the natural target voice (no conversion).
- TGT-TGT This result corresponds to the comparison between two sentences uttered by the same natural target voice (used as baseline result).

System	**IVC2 (Overall Quality)**	
	Score (1<5)	**Rank**
TARGET	4.32	*1*
IVC_IBM1	3.63	***2***
IVC_IBM2	2.71	*4*
IVC_NOK	1.45	*8*
IVC_SIE1	3.11	*3*
IVC_SIE1_06	2.63	*5*
IVC_SIE2	2.00	*7*
IVC_UPC	2.50	*6*

Table 45: Results of the intra-lingual voice conversion quality judgement tests VC2 for English

Table 46 gives the global VC scores for each system. The global score is computed as the mean between the VC2 score and the mean VC1 score (i.e. the average VC1 score in all conversion directions).

System	**Global VC Score**			
	mean VC1 score	**VC2 score**	**mean (VC1,VC2)**	**Rank**
IVC_IBM1	2.32	3.63	2.98	***1***
IVC_IBM2	2.76	2.71	2.73	*2*
IVC_NOK	2.17	1.45	1.81	*7*
IVC_SIE1	1.75	3.11	2.43	*5*
IVC_SIE1_06	2.44	2.63	2.54	*4*
IVC_SIE2	2.81	2.00	2.40	*6*
IVC_UPC	2.88	2.50	2.69	*3*

Table 46: Global VC scores for English

The results in table 46 reflect the trade-off between the conversion accuracy and the voice quality. The best trade-off is obtained by IVC_IBM1.
In comparison to last year (IVC_SIE_06), the IVC_SIE1 system has improved the voice quality, losing some voice conversion accuracy at the same time.

4.5.2 Results for Spanish

TTS component (S1, S2). The results are reported in Table 47 and Figure 15. Only the Overall Quality test results are reported here. The intervals of confidence are also reported: the interval of confidence (at 95%) for S1, and the Wilson score interval (at 95%) for S2.
Section 9 provides a more detailed presentation of these results, including the 10 judgement categories of task S1.

TTS Component Evaluation											
System	**S1 (Overall Quality)**				**S2**						
	Score	**IC**		**Rank**	**WER(%)**	**WSI**					**Rank**
NAT	4.75	±	0.24	*1*	-			-			-
IBM_F_06	3.89	±	0.29	*7*	-			-			-
IBM_F	4.00	±	0.26	*4*	7.5	[	5.4	-	10.3	]	*3*
IBM_M	4.00	±	0.24	*4*	12.1	[	9.4	-	15.5	]	*6*
UPC_F	3.42	±	0.37	*9*	7.1	[	5.0	-	9.9	]	*2*
UPC_M	3.47	±	0.32	*8*	6.0	[	4.2	-	8.6	]	***1***
VER_F1	4.22	±	0.26	***2***	12.2	[	9.5	-	15.5	]	*7*
VER_M1	4.06	±	0.27	*3*	9.7	[	7.3	-	12.8	]	*5*
VER_M2	3.94	±	0.27	*6*	8.4	[	6.2	-	11.3	]	*4*

Table 47: Results of the TTS component evaluation tasks S1 and S2 (Spanish)

Legend:

- NAT Natural voice, used as top-line in subjective tests.
- IBM_F_06 Female voice submission made by IBM last year (re-evaluated this year)
- IBM_F/M IBM submission using female / male voices

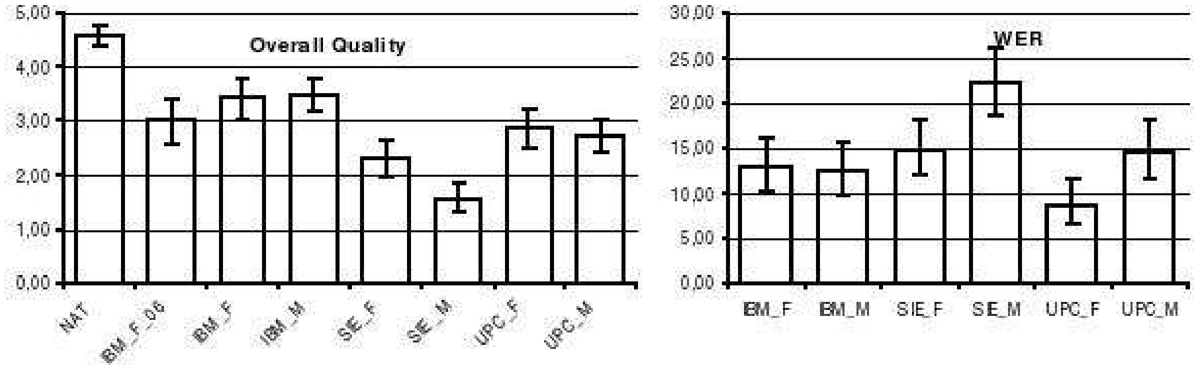


Figure 15: Overall Quality (S1) and WER (S2) results with intervals of confidence (Spanish).

- UPC_F/M UPC submission using female / male voices
- VER_F1/M1 Verbio submission using female / male voices
- VER_M2 2^{nd} Verbio submission using male voice
- WER Word Error Rate.
- IC Interval of Confidence (at 95%)
- WSI Wilson Score Interval (at 95%)

The best Overall Quality score is obtained by the male and female Verbio voices, but the difference with the IBM scores is not statistically significant.
On the other hand, UPC male and female voices give the 2 lowest word error rates in test S2.
In terms of Overall Quality score, the IBM female voice (IBM_F) performs slightly better as last year (IBM_F_06), but the difference is not statistically significant (cf. the confidence interval).

Voice conversion (VC1, VC2). Results of the comparative tests (VC1) and the overall quality judgement tests (VC2) are reported in Tables below. Table 48 and Table 49 refer to the intra-lingual voice conversion task, Table 51 and Table 52 to the cross-lingual voice conversion task.

Intra-lingual Voice Conversion: Comparison test VC1 (*Scoring: 5>1*)								
Conversion System	**Conversion F(75)→F(76)**		**Conversion F(75)→M(79)**		**Conversion M(80)→F(76)**		**Conversion M(80)→M(79)**	
	Score	**Rank**	**Score**	**Rank**	**Score**	**Rank**	**Score**	**Rank**
IVC_IBM1	2.10	*5*	2.30	*5*	2.50	***3***	1.90	*4*
IVC_IBM2	2.40	*4*	3.10	*4*	2.00	*5*	1.90	*4*
IVC_SIE1	1.10	*8*	2.00	*7*	1.10	*7*	1.30	*8*
IVC_SIE2	1.90	*6*	2.20	*6*	2.00	*5*	1.80	*6*
IVC_UPC	2.90	***3***	2.90	***3***	2.20	*4*	3.00	***3***
IVC_UPC1_06	3.80	***2***	3.80	***2***	3.70	***2***	3.50	***2***
SRC-TGT	1.75	*7*	1.00	*8*	1.00	*8*	1.43	*7*
TGT-TGT	4.74	*1*	4.56	*1*	4.74	*1*	4.56	*1*

Table 48: Results of the intra-lingual voice conversion comparison tests VC1 for Spanish

Legend:

- IVC Intra-lingual voice conversion (Spanish to Spanish) There are 5 IVC submissions: 2 from IBM (IVC_IBM1+2), 2 from Siemens (IVC_SIE1+2) and 1 from UPC (IVC_UPC1).
- IVC_UPC1_06 The submission made by UPC last year is re-evaluated.

- CVC Cross-lingual voice conversion (Spanish to English) There are 4 CVC submissions: 1 from IBM (CVC_IBM1), 2 from Siemens (CVC_SIE1+2) and 1 from UPC (CVC_UPC1).
- F(*n* Female voice number *n*
- M(*n*) Male voice number *n*
- *A*→*B* Conversion from voice *A* (source) to voice *B* (target). Source voice *A* is a Spanish voice. Target voice *B* is a Spanish voice (in the case of IVC) or an English voice (in the case of CVC). The *A*→*B* conversion consists in synthesising voice *B* from the natural voice *A*. The conversion evaluation score results from comparing the natural voice *B* with the synthesised voice *B*.
- SRC-TGT This result corresponds to the comparison between the natural source voice and the natural target voice (no conversion).

- TGT-TGT This result corresponds to the comparison between two sentences uttered by the same natural target voice (used as baseline result).

System	**IVC2 (Overall Quality)**	
	Score (1<5)	**Rank**
TGT	4.72	*1*
IVC_IBM1	3.48	***2***
IVC_IBM2	2.92	*4*
IVC_SIE1	3.30	*3*
IVC_SIE2	2.35	*7*
IVC_UPC1	2.85	*5*
IVC_UPC1_06	2.55	*6*

Table 49: Results of the intra-lingual voice conversion quality judgement tests VC2 for Spanish

Table 50 gives the global IVC scores for each system. The global score is computed as the mean between the VC2 score and the mean VC1 score (i.e. the average VC1 score in all conversion directions).

System	**Global IVC Score**			
	mean VC1 score	**VC2 score**	**mean (VC1,VC2)**	**Rank**
IVC_IBM1	2.20	3.48	2.84	***2***
IVC_IBM2	2.35	2.92	2.64	*4*
IVC_SIE1	1.38	3.30	2.34	*5*
IVC_SIE2	1.98	2.35	2.16	*6*
IVC_UPC1	2.75	2.85	2.80	*3*
IVC_UPC1_06	3.70	2.55	3.13	***1***

Table 50: Global IVC scores for Spanish

The results in Table 50 reflect the trade-off between the conversion accuracy and the voice quality. Among this year's systems, the best trade-off is obtained by IVC_IBM1, which also yields the best quality result (VC2).
Last year's system IVC_UPC1_06 yields the best overall results. It has a lower quality score as this year's system (IVC_UPC1), but compensates this with a much better conversion accuracy, resulting in a better trade-off (Global IVC Score).

Cross-lingual Voice Conversion: Comparison test VC1 (*Scoring: 5>1*)								
Conversion System	**Conversion F(75)→F(76)**		**Conversion F(75)→M(79)**		**Conversion M(80)→F(76)**		**Conversion M(80)→M(79)**	
	Score	**Rank**	**Score**	**Rank**	**Score**	**Rank**	**Score**	**Rank**
CVC_IBM1	2.10	*4*	2.00	*3*	1.40	*5*	1.60	*4*
CVC_SIE1	1.40	*6*	1.20	*5*	1.50	*4*	1.40	*6*
CVC_SIE2	2.60	*3*	1.40	*4*	2.00	***2***	1.70	*3*
CVC_UPC	2.70	***2***	2.30	***2***	1.70	*3*	3.80	***2***
SRC-TGT	1.75	*5*	1.00	*6*	1.00	*6*	1.43	*5*
TGT-TGT	4.74	*1*	4.56	*1*	4.74	*1*	4.56	*1*

Table 51: Results of the cross-lingual voice conversion comparison tests VC1 for Spanish

System	**CVC2 (Overall Quality)**	
	Score (1<5)	**Rank**
TGT	4.72	*1*
CVC_IBM1	3.52	***2***
CVC_SIE1	3.23	*3*
CVC_SIE2	2.02	*5*
CVC_UPC1	2.80	*4*

Table 52: Results of the cross-lingual voice conversion quality judgement tests VC2 for Spanish

Table 53 gives the global CVC scores for each system. The global score is computed as the mean between the VC2 score and the mean VC1 score (i.e. the average VC1 score in all conversion directions).

System	Global CVC Score			
	mean VC1 score	VC2 score	mean (VC1,VC2)	Rank
CVC_IBM1	1.78	3.52	2.65	*2*
CVC_SIE1	1.38	3.23	2.30	*3*
CVC_SIE2	1.93	2.02	1.97	*4*
CVC_UPC1	2.63	2.80	2.71	***1***

Table 53: Global CVC scores for Spanish

Regarding CVC, the best voice quality is obtained by CVC_IBM1, but the best "*conversion accuracy vs. quality*" trade-off is obtained by UPC.

Voice conversion based on found data (fCVC1,fCVC2). Results of the comparative tests (fCVC1) and the overall quality judgement tests (fCVC2) are reported in Table 54 and Table 55.

Legend:

- fCVC Cross-lingual voice conversion based on found data (Spanish to English) 3 submissions: 2 from Siemens (fCVC_SIE1+2) and 1 from UPC (fCVC_UPC).
- M_ES(73) Source male voice, in Spanish (synthesised by IBM)
- M_EN(*n*) Target male voice, in English (found data, European Parliament)
- *A*→*B* Conversion from voice *A* (source) to voice *B* (target).
- SRC-TGT This result corresponds to the comparison between the natural source voice and the natural target voice (no conversion).
- TGT-TGT This result corresponds to the comparison between two sentences uttered by the same natural target voice (used as baseline result).

Cross-lingual Voice Conversion based on Found Data (fCVC1)				
Conversion System	Conversion: M_ES(73)→M_EN(01)		Conversion: M_ES(73)→M_EN(02)	
	Score	Rank	Score	Rank
SRC-TGT	1.50	*3*	1.89	*4*
TGT-TGT	4.85	*1*	4.80	*1*
fCVC_SIE1	1.00	*5*	1.90	*3*
fCVC_SIE2	1.40	*4*	1.40	*5*
fCVC_UPC	1.70	***2***	2.10	***2***

Table 54: Results of the fCVC1 comparison tests (Spanish)

System	fCVC2 (Overall Quality)	
	Score (1<5)	Rank
TARGET	4.60	*1*
fCVC_SIE1	3.00	*3*

fCVC_SIE2	1.79	*4*
fCVC_UPC	3.19	***2***

Table 55: Results of the fCVC2 comparison tests (Spanish)

Table 56 gives the global fCVC scores for each system. The global score is computed as the mean between the VC2 score and the mean VC1 score (i.e. the average VC1 score in all conversion directions).

System	**Global fCVC Score**			
	mean VC1 score	**VC2 score**	**mean (VC1,VC2)**	**Rank**
fCVC_SIE1	1.45	3.00	2.23	*2*
fCVC_SIE2	1.40	1.79	1.60	*3*
fCVC_UPC	1.90	3.19	2.55	***1***

Table 56: Global fCVC scores (Spanish)

The results in 56 reflect the trade-off between the conversion accuracy and the voice quality. For this new task, the best trade-off is obtained by UPC.

4.5.3 Results for Chinese

TTS component (S1) The results are reported in Table 57 and Figure 16. Only the Overall Quality test results are reported here. The intervals of confidence (at 95%) are also reported.
Annex C provides a more detailed presentation of these results, including the 10 judgement categories of task S1.

Legend:

- NAT Natural voice, used as top-line in subjective tests,
- NOK_06 Submission made by Nokia last year (re-evaluated this year),
- IC Interval of Confidence (at 95%).

TTS Component Evaluation			
System	**S1 (Overall Quality)**		
	Score	**IC**	**Rank**
NAT	4.19	±0.24	*1*
CAS	3.86	±0.26	***2***
NOK	2.85	±0.33	***3***
NOK_06	2.61	±0.34	*4*

Table 57: Results of the TTS component evaluation tasks S1 (Chinese)

The best Overall Quality score is obtained by the CAS voice.
Nokia's TTS voice performs slightly better this year as last year (NOK_06).

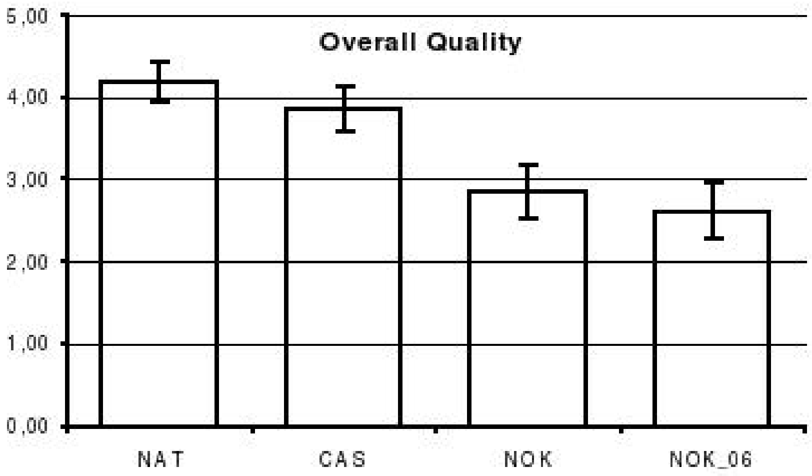


Figure 16: Overall Quality (S1) and WER (S2) results with intervals of confidence (Chinese).

5 End-to-end evaluation

5.1 Tasks and conditions

As for the second evaluation campaign of TC-STAR, an end-to-end evaluation has been carried out in the third evaluation campaign. This evaluation includes speech recognition, spoken translation and speech synthesis.

In translation, the two basic concepts to take into account are *adequacy* and *fluency*. However, we think that in *speech-to-speech* translation, rather than asking these questions to translation experts, it is preferable to use *adequacy* and *fluency* questionnaires, to be filled in by human judges acting as potential users. In particular, we believe it is very difficult for an expert to make a *judgement* about the adequacy, based on the listening of the human source speech and the synthetic speech in the target language. Instead, we use a *functional test* where the understanding is rated.

- Adequacy: comprehension test on potential users allows measuring the intelligibility rate.
- Fluency: judgement test with several questions related to fluency and also usability of the system

The end-to-end evaluation is carried out only for the English-to-Spanish translation direction.

5.2 Language resources

Although three different directions are performed in TC-STAR (English-to-Spanish, Spanish-to-English, Chinese-to-English) we only consider the English-to-Spanish direction for time and cost constraints. The evaluation data consist of same audio recordings in English of the European Parliament Plenary Sessions (EPPS) used in ASR and SLT. The evaluation data is made of 20 segments of around 3 minutes each. So in total the evaluation set is composed of one hour of speech and around 8,000 running English words. The European Parliament is translating and broadcasting in real time, each Plenary Session in many languages, including Spanish. Therefore, the corresponding Spanish audio translation made by professional interpreters was recorded. This human translation audio data is evaluated in the same way as the automatic translation. The TC-STAR system includes the following modules. The ASR module is the combination of several ASR engines. The SLT component is provided by RWTH. The TTS module is the system provided by UPC. These three components are trained on data including training corpora built from the EPPS recordings. For each audio sample in English an ASR output is produced, then the ASR output is automatically translated into Spanish and finally, the SLT output is synthesised in Spanish by the TTS module using the alignment between SLT and ASR to get the prosodic features from the source language. The transit from one component to another is done manually but no modifications on the different outputs were done and so the system can be considered as fully automatic.

5.3 Schedule

The end-to-end run and the evaluation took place in February-March 2007.

5.4 Participants and submissions

One joint submission from the TC-STAR consortium is evaluated and the corresponding interpreters speeches as well. The speech from the interpreters is collected as a top-line. The table below summarises the participants for each component.

Component	**Input**
ASR	ROVER (ASR)
SLT	ROVER (SLT)
TTS	UPC

Table 58: Test data

5.5 Protocol

End-to-end samples are common with the SLT evaluation data selected for the SLT human evaluation. This year, one third of the SLT data has been evaluated with human judges (approx. 1 hour of audio). The selection procedure is detailed here.

1. ELDA selected some *semantically interesting* audio samples, taken from English (source) politicians, approximately 20 speeches x 2 or 3minutes.

2. The corresponding data is extracted in each module (ASR, SLT, TTS) and the 20 evaluations samples were evaluated

3. Corresponding speeches from interpreters (in Spanish, as target language) were collected. These are the *20 reference samples*, which are the top-line.

The ASR and SLT outputs were produced during the respective evaluations. After the two evaluations were done, SLT output and ASR output were sent to UPC, who produced the synthesised audio.

Both interpreter and TC-STAR samples are used for the evaluation.

The evaluation is done by human judges without any specific experience on speech technology. For processing the subjective evaluation, ELDA has recruited 20 subjects who are native Spanish speaker, 18-40 years old and with no hearing problem. They are not experts in speech synthesis and they are paid for the task. Subjects are required to have access to high-speed/ADSL Internet connection and good listening material. Subjective tests are carried out via the web. A specific interface has been developed, similar to the interface used for the SLT human evaluation.

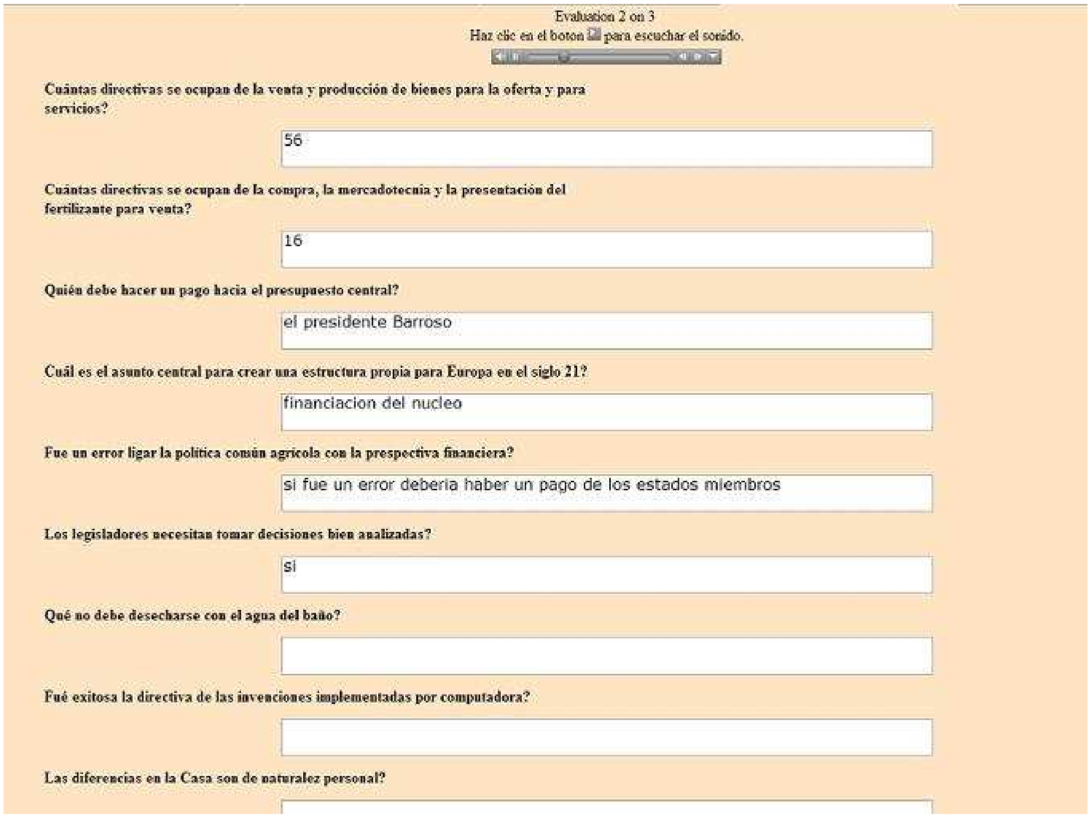
Four evaluations by each evaluator are done. Each sample is presented to judge with the adequacy and fluency questionnaires, and each judge assesses two TC-STAR samples and two interpreter samples. As there are a total of 40 audio, each sample is evaluated twice, by two different judges. In that way, we are able to observe the inter-judges evaluation agreement.

Evaluators are explained the TC-STAR system and the evaluation procedure. Within the interface, the evaluator can play the sound corresponding to either TC-STAR speech or interpreter speech, during the evaluation session. Each evaluator assesses at least one TC-STAR audio and one interpreter audio.

They are instructed to:

- read the questionnaire,
- listen the whole sample,
- listen a second time. They are allowed to interrupt the listening and to write down the answers in the adequacy questionnaire.

At the end of the evaluation session, they are asked to fill the fluency questionnaire.



Evaluation 2 on 3

Haz clic en el boton para escuchar el sonido.

Cuántas directivas se ocupan de la venta y producción de bienes para la oferta y para servicios?

56

Cuántas directivas se ocupan de la compra, la mercadotecnia y la presentación del fertilizante para venta?

16

Quién debe hacer un pago hacia el presupuesto central?

el presidente Barroso

Cuál es el asunto central para crear una estructura propia para Europa en el siglo 21?

financiacion del nucleo

Fue un error ligar la politica común agricola con la prespectiva financiera?

si fue un error deberia haber un pago de los estados miembros

Los legisladores necesitan tomar decisiones bien analizadas?

si

Qué no debe desecharse con el agua del baño?

Fué exitosa la directiva de las invenciones implementadas por computadora?

Las diferencias en la Casa son de naturalez personal?

Figure 17: Interface for the end-to-end evaluation.

Adequacy questionnaire. For each sample, 20 comprehension questionnaires have been prepared, based on the English speeches, by a native English speaker. For each sample, 10 questions are asked about the sample the listener has just heard. To prepare the questionnaire, the whole 200 questions have been created from the manual transcriptions of speeches, and preserved with the answers to the questions, which account for the "reference answers". Then the answers and questions have been translated into Spanish to be inserted into the evaluation interface and used to check and score the evaluations.

Questions are asked taking into account criteria from question-answering domain. Three types of question are asked: "simple Factual" (70%), "Yes/No" (20%) and "List" (10%), without reformulation.

For the 2^{nd} year evaluation, two criteria were used: "correct" (the answer is good) or "inexact" (the answer is not good) [9]. This year, we have introduced the "incomplete/wrong" criteria from information retrieval domain, when the answer is not complete or not formulated correctly. To measure the performance, the cumulative precision measure is used, computing the percentage of correct answers for each criteria and cumulate them.

The two examples depicted in Table 59 illustrate the relevance of using 3 criteria.

Question	**Correct answer**	**Incomplete answer**	**Inexact answer**
When is the ministerial meeting on the Northern Dimension?	The 21st November	In November	The 12th May
Where did fascist or military dictatorships exist 35 years ago?	Greece, Spain and Portugal	Greece and Spain	Switzerland

Table 59: Example of questions asked to the evaluators

After all the evaluations are done, a native Spanish speaker compares the answers of the evaluators to the reference answers. It has been asked to this person to be “flexible”, as the reference answers are not exactly the same than the evaluator answers. As example, the references answer to the question “Por qué publicación está concernido el vocero del grupo?” (“Which publication is the speaker’s group concerned about?” in English) was “La publicación del código de conducta para las organizaciones no lucrativas” (resp. “The publication of the code of conduct for not-for-profit organisations”), while the evaluator answer “del código de conducta sobre las organizaciones sin ánimo de lucro” (resp. “The code of conduct for organizations without profit objectives”), which is correct. Then it is obvious the evaluation could only be done by a human, and not automatically: each evaluator answers differently (with a sentence, or just the completion of the question, or a single word, etc.) even if the answer submitted is good. Furthermore synonyms could be used, or paraphrases, etc.

An objective verification has also been done to check the presence of the answers in each component of the end-to-end process (ASR and SLT), in order to determine in which component of the TC-STAR system the information is lost. This objective verification is done by a native speaker who checks each answer given by a judge and compares it with the reference answer. The same identification is quite easier for the TC-STAR system, as we already know where the evaluation could be lost, namely when the information past trough one of the two components (ASR or SLT). For that, we study the whole end-to-end chain in order to see where the information is lost. A native Spanish read each question, and look at whether the answers are present within the SLT text or within the ASR text, in case the answer is not found before (actually we consider that if information is found within a component -including subjective evaluation- information is also in the component upstream). Of course, for an objective comparison the person who checks the files has the reference answers in plain view.

Fluency questionnaire. Fluency questions are done at the end of the evaluation of each sample (since the two “systems” are evaluated, it is difficult to assess all the audio samples). Then, the mean of each system is computed for the interpreter and the TC-STAR system.

Test	**Fluency questionnaire**
Understanding	Do you think that you have understood the message? 1: Not at all 5: Yes, absolutely
Fluent Speech	Is the speech in good Spanish? 1: No, it is very bad 5: Yes, it is perfect

Effort	Rate the listening effort 1: Very high 5: Low, as natural speech
Overall Quality	Rate the overall quality of this audio sample 1: Very bad, unusable 5: It is very useful

Table 60: Fluency questionnaire

Each answer is a choice within a five-point scale, from the worst level to the best. After all the evaluations are done, the means for the interpreter speeches and the TC-STAR speeches has been computed.

5.6 Results

5.6.1 Fluency evaluation (subjective evaluation)

Speech	**Samples**	**Understanding** 1: low quality 5: better quality	**Fluent Speech** 1: low quality 5: better quality	**Effort** 1: low quality 5: better quality	**Overall Quality** 1: low quality 5: better quality
Interpret	1	3.5	3.5	3	3.5
	2	4	5	3.5	5
	3	4	3.5	3	4
	4	4.5	5	4	4.5
	5	3.5	4	3.5	4
	6	5	5	5	5
	7	3.5	3	2	3
	8	5	4	4	4
	9	4.5	4	4	4
	10	4.5	4.5	4.5	4.5
	11	4	4	4	4.5
	12	3	3.5	3	3
	13	4.5	4.5	3.5	3.5
	14	3	4	3	4
	15	3.5	5	3.5	3.5
	16	3.5	3.5	3	4
	17	2.5	3	2.5	4
	18	3.5	4	3	3.5
	19	4.5	4.5	2.5	4.5
	20	3	4	3	4.5
	mean	**3.85**	**4.08**	**3.38**	**4.03**

Table 61: Fluency evaluation results for the interpreter

The scores of the interpreter samples are rather high. Only the “Effort” is lower, but it can be explain by the difficulties for the interpreter to translate in a same time what the speaker says. So the speech can be more disrupted than a conversational speech.

Speech	**Samples**	**Understanding** 1: low quality 5: better quality	**Fluent Speech** 1: low quality 5: better quality	**Effort** 1: low quality 5: better quality	**Overall Quality** 1: low quality 5: better quality
TCSTAR	1	3.5	2	1.5	2.5
	2	2	1.5	1.5	2
	3	3.5	3	3	2.5
	4	1.5	3	1.5	1.5
	5	3.5	3	2.5	2.5
	6	3	2	2	2
	7	3	2.5	1.5	2
	8	4	2	1.5	2
	9	2	2	2	1.5
	10	2	2	1	2
	11	1.5	1.5	1	2.5
	12	2	2	1.5	1.5
	13	2.5	2	2	2
	14	3	1	2	1.5
	15	2	2	1	2
	16	2	1.5	2	2
	17	1.5	1.5	1	1
	18	1.5	2	1	3.5
	19	2.5	2	1.5	2
	20	2	2	1.5	2.5
	mean	**2.43**	**2.03**	**1.63**	**2.05**

Table 62: Fluency evaluation results for the TC-STAR system

The results for the TC-STAR system are quite low, except for some samples. There is no sample which gets higher score for TC-STAR than for the interpreter. As for last year results, the difference with the interpreter sample is still very large.

	Understanding 1: low quality 5: better quality	**Fluent Speech** 1: low quality 5: better quality	**Effort** 1: low quality 5: better quality	**Overall Quality** 1: low quality 5: better quality
ITP-2006	**3.45**	**3.48**	**3.19**	**3.52**
ITP-2007	**3.85**	**4.08**	**3.38**	**4.03**
TC-STAR-2006	**2.34**	**1.93**	**1.55**	**1.93**
TC-STAR-2007	**2.43**	**2.03**	**1.63**	**2.05**

Table 63: Fluency comparison between 2nd **year and 3rd year evaluations**

In general terms, the trend of the scores are the same than last year, although the scores are slightly higher.

5.6.2 Adequacy evaluation (comprehension evaluation)

The table below presents the results of the adequacy evaluation. It shows:

- the two evaluated systems: the interpreter (ITP) and the TC-STAR automatic speech-to-speech translation system;
- identifiers of the audio file. Source data are the same for interpreter and TC-STAR, namely the English speech;
- subj. E2E: the subjective results of the end-to-end evaluation were done by the same assessors who did the fluency evaluation. It shows the percentage of good answers;
- fair E2E: objective verification of the question answers presence: the audio files have been validated to check whether they contained the answers to the questions or not (as the question were created from the English source). It shows the percentage of answer presence or the maximum answers that can be found in the Spanish translations. For example information in English could have been not translated by the interpreter because he/she feels that this information is meaningless and can be discarded. We consider those results as an objective evaluation. For the interpreter it corresponds to the speaker audio, for the TC-STAR system, this is the TTS audio output.
- SLT, ASR: verification of the answers presence in each component of the end-to-end process: in order to determine where the information for the TC-STAR system is lost, files from each component (recognised files for ASR, translated files for SLT, and synthesised files for TTS in the "fair E2E" column) have been checked.

Speech	**Audio**	**subj. E2E** 0 : low 1 : better	**fair E2E** 0 : low 1 : better
ITP	1	0.40	0.50
	2	0.60	1.00
	3	0.60	0.70
	4	0.85	1.00
	5	0.85	1.00
	6	0.75	1.00
	7	0.85	1.00
	8	1.00	1.00
	9	0.95	1.00
	10	0.65	0.80
	11	0.70	0.90
	12	0.70	1.00
	13	0.85	1.00
	14	0.75	1.00
	15	0.80	1.00
	16	0.45	0.80
	17	0.60	0.80
	18	0.45	0.70
	19	1.00	1.00
	20	0.90	1.00
	mean	**0.74**	**0.91**

Table 64: Adequacy evaluation results for the interpreters

Results are surprisingly no so perfect than it could be imagined. For seven samples, all the questions could not be answered. The first sample is especially worst than all the other samples. The objective validation concludes there is an overall loss of information of 9%, which means evaluators could not answer to 9% of the questions. Actually, effective overall loss is 26% with the subjective evaluation, which means evaluators did not answer to a quarter of the questions. For three of the samples, evaluation is more difficult, since scores are less than 0.50. 17 audio samples contain more than 75% of correct answers (considering the fair evaluation) but evaluators found 75% of correct answers for only 11 audio samples.

There are several explanations to that quality decrease.

- Interpreters have difficulties to follow the speaker flood, most of the time due to the assigned time to translate speaker discourse. Thereby they give fewer details and filter information. For instance, a question for the sample 1 is (in English): *Which main German newspaper published a report denying the link between the World Cup and an increase in trafficking and forced prostitution ?* The correct answer is the German newspaper ”*Der Spiegel*” but in fact the words *Der Spiegel* are never said by the interpreter, so the evaluators can not answer this kind of questions.

- As a consequence, interpreters flood is not continuous. They are often forced to concentrate the information and reduce the number of sentences. It is possible to have five English sentences reduced to two Spanish sentences.

- The difficulties to be tackled by the interpreters can also been explained: speaker flood, speaker hesitations, time needed for translation, but also the grammatical construction of sentences. Indeed, direction of sentences in English is not necessarily the same than in Spanish, and so the interpreters have to wait for the end of the speaker sentence to start the Spanish translated sentence (and the speaker can hesitate, or take back himself, etc.).

- Interpreters reformulate speaker sentences, and so increase the ambiguity of some questions.

Opposite to these points, it is possible that interpreters take back themselves. It allows evaluators better understand the information (repetition of a same information often paraphrased).

Speech	**Audio**	**subj. E2E** 0 : low 1 : better	**fair E2E** 0 : low 1 : better	**SLT** 0 : low 1 : better	**ASR** 0 : low 1 : better
TCSTAR	1	0.65	0.80	0.80	0.80
	2	0.75	0.90	0.90	1.00
	3	0.70	1.00	1.00	1.00
	4	0.45	0.70	0.90	1.00
	5	0.85	1.00	1.00	1.00
	6	0.80	1.00	1.00	1.00
	7	0.65	1.00	1.00	1.00
	8	0.90	1.00	1.00	1.00
	9	0.40	1.00	1.00	1.00
	10	0.55	0.80	0.90	0.90
	11	0.40	0.60	0.70	0.90
	12	0.80	1.00	1.00	1.00
	13	0.70	0.80	0.80	1.00
	14	0.80	0.90	0.90	0.90
	15	0.85	0.90	1.00	1.00
	16	0.30	0.80	0.80	1.00
	17	0.10	0.70	0.70	0.90

	18	0.70	1.00	1.00	1.00
	19	0.85	1.00	1.00	1.00
	20	0.60	0.90	0.90	0.90
	mean	**0.64**	**0.89**	**0.92**	**0.97**

Table 65: Adequacy evaluation results for the TC-STAR system

TC-STAR results are quite lower than the interpreter ones. Evaluators found 64% of the answers while they could answer 89% of the questions. Here again, there is a strong difference between the subjective evaluation and the objective evaluation and evaluators did not find 71% of the answers they can. No audio sample has all the answers correct and eight of them have 75% of the answers correct. Then the overall of the TC-STAR system is 36% while the overall loss of the interpreter is 26%.

The validation of each TC-STAR components allows us to get the overall loss at each step of the process. The ASR component gets an overall loss of 3%, 5% in addition for the SLT component and finally 3% in addition for the TTS component, making a total overall loss of 11%. The ASR component quality does not decrease too much (3% of the answers could not be found) and all audio samples contain at least 80% of the information (6 files do not contain all the answers). Information lost concerns typical recognition errors which could affect the meaning of sentences. The SLT component quality decreases a little bit more (8% of the answers could not be found), only two audio samples are under 75% and half of the audio samples contain the information to answer all the questions. Information lost is due to typical translation errors. Finally, TTS component quality decreases also (11% of the answers could not be found) and 9 audio samples contain the information to answer all the questions. Information lost is due to synthesis errors. This point is particularly interesting since it is quite difficult to understand why the synthesis of sentences could affect their meaning, all the more the fluency evaluation is not really disastrous. Two typical cases are described below:

- Synthesis issues with the named entities. Named entities badly synthesised can affect details of a sentence. For the sample 10, the name *Sophie Veld* is correctly translated and the answer of the question *Who said that we need action from the Commission and from the Finnish Presidency?* can be easily found. But the TTS component synthesises the name as something not understandable, even listening many times.

- Prosody issues. Bad prosody can affect the meaning of sentences. For the sample 15, the following sentence is a good example: *[...]mientras que si esa empresa estaba fuera de la Unin Europea cada Estado miembro comprobar concienzudamente y que es un problema* which is a translation of the recognized sentence: *whereas if that company was outside the European Union every Member state would check thoroughly and that's a problem* even the translation quality is low, the sentence is understandable, and evaluators (and the person who made the validation) can easily answer to the question *In which condition would Member States examine thoroughly a financial services company?, Only when the company was outside the E.U.*. The synthesis accentuates the prosody of the syllable *bar* of the word *comprobar*, by letting imagine another sentence begins. Then there is one sentence in the SLT output, but the prosody splits into two sentences in the TTS output, making the question hard to answer. This issue is also increased by the quality of the sentence itself. As an example, the infinitive *comprobar* makes understandable the meaning of the sentence, but without punctuation or vocally it does not reflect its position in a sentence (or two).

Interpreters filter and reformulate the information while the TC-STAR system can not: for the automatic speech-to-speech translation all the information is pass through the chain, without selection. The table below summarises the comparison between the two systems about the information loss.

	ITP	TC-STAR	
		SLT	ASR
Objective loss	9%	8%	3%
Subjective loss	26%	36%	-
Audios > 80%	9	19	20

Table 66: Information loss for the two systems

To objectively compare interpreter and TC-STAR, we have selected only the questions whose answers were included in the interpreter files. The goal is to compare the overall quality of the speech-to-speech translation to interpreters' quality, without the noise factor of the information missing. So we get a new subset of the TC-STAR results, on the information kept by the interpreter. The same study as before has been done for the three components.

Speech	**Audio**	**subj. E2E** 0 : low 1 : better	**TTS** 0 : low 1 : better	**SLT** 0 : low 1 : better	**ASR** 0 : low 1 : better
TCSTAR (ITP 1.00 only)	1	0.80	1.00	1.00	1.00
	2	0.75	0.90	0.90	1.00
	3	0.71	1.00	1.00	1.00
	4	0.45	0.70	0.90	1.00
	5	0.85	1.00	1.00	1.00
	6	0.80	1.00	1.00	1.00
	7	0.65	1.00	1.00	1.00
	8	0.90	1.00	1.00	1.00
	9	0.40	1.00	1.00	1.00
	10	0.69	0.88	0.88	0.88
	11	0.44	0.67	0.78	0.89
	12	0.80	1.00	1.00	1.00
	13	0.70	0.80	0.80	1.00
	14	0.80	0.90	0.90	0.90
	15	0.85	0.90	1.00	1.00
	16	0.38	0.88	0.88	1.00
	17	0.13	0.63	0.63	0.88
	18	0.71	1.00	1.00	1.00
	19	0.85	1.00	1.00	1.00
	20	0.60	0.90	0.90	0.90
	mean	**0.66**	**0.91**	**0.93**	**0.97**

Table 67: Limited evaluation results for TC-STAR

Results of 8 audio samples increase but only three increase significantly. Overall scores are quite better with at the most 2% in addition.

Speech	**Audio**	**subj. E2E** 0 : low 1 : better
	1	0.80
	2	0.60
	3	0.86

	4	0.85
	5	0.85
	6	0.75
	7	0.85
	8	1.00
	9	0.95
	10	0.81
	11	0.78
	12	0.70
	13	0.85
	14	0.75
	15	0.80
	16	0.56
	17	0.75
	18	0.64
	19	1.00
	20	0.90
	mean	**0.80**

Table 68: Limited evaluation results for interpreter

As for the TC-STAR system the results increase, but differences are more important. Overall loss is lower of 6% and 7 audio samples increase scores significantly. However, overall loss for the subjective evaluation is still 20% regarding the objective evaluation.

	ITP	TC-STAR	
		SLT	ASR
Objective evaluation	100%	92%	97%
Subjective evaluation	80%	66%	-
Audios > 80%	12	18	20

Table 69: sum up of the evaluation

TC-STAR system needs to improve, but we get promising results, while it recovers 91% of the information that the interpreter could give on these samples (with specific data and questions).

About the evaluation itself, protocol needs to improve again. Questions were often to difficult and detailed, or on the contrary to unspecific, allowing sometimes many answers. It seems evaluators interpolate information and deduct answers when it is possible. For instance when a question begins by ”How many...” it is easily to know that a number is wanted, and so focus attention on the number given by the interpreter.

6 Conclusion

Although it is hard to summarize all the tests carried out by a few scores, let us try to illustrate these to give a rough idea. For ASR, the best results obtained, given the test conditions and test data, by an individual site is as good as 7.1% error rate for English (respectively 6.9% for Spanish) for open training conditions and about 9% for public training (resp. 8.9% for Spanish). The TC-STAR System, based on the ROVER approach, achieved a word error rate of 6.9% for English and 7.4 for Spanish.

Progress from previous years (campaign 1 of 2005 and Campaign 2 of 2006) have been measured and reported on for sites that have kept their annual releases. This assessment shows substantial improvements from all sites e.g. for Spanish RWTH went from 11.5% to 9%. For SLT we have carried out both human evaluations (of adequacy and fluency) and automatic assessment using a set of automatic metrics (e.g. BLEU, WER). As by the past campaigns three conditions of system input have been exploited: FTE, Verbatim and ASR outputs.

The best BLEU scores obtained by a single system for English to Spanish is about 54.11 (FTE), 51.53 (verbatim) and 39.66 (ASR); the ROVER system performed little bit less than the best single system for FTE (53.85) but better for Verbatim (52.63) and ASR (40.61).

Compared to 2006 systems, 2007 systems achieved an important improvement of over 4% in absolute for the BLEU score and English-Spanish pair. Similar results are reported on for Spanish to English and Mandarin to English.

For TTS, most of the evaluations are based on subjective tests and are hard to summarize. Let us just give some for the global quality score for English which is about 3.63 to compare with 4.32 (human voice) for the interlingual voice conversion.

For the global quality we achieved a score of 3.49 to compare with the score of natural voice of 4.59 (out of 5).

End to end evaluation has been also performed and TC-STAR was compared to the human interpreters to assess adequacy, fluency, and information preservation for both. Details are given with the protocol used to conduct such evaluation.

References

[1] Bonafonte A. et al "TC-STAR Deliverable D8: TTS Baselines and Specifications."

[2] Chinese language data consortium http://www.chineseldc.org

[3] ELRA's catalog of Language Resources http://catalog.elra.info

[4] Fiscus J. G. "A Post-Processing System To Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)" in *1997 Proc. IEEE ASRU Workshop*, Santa Barbara, 1997, pp.347–352

[5] Hamon O., Popescu-Belis A., Choukri K., Dabbadie M., Hartley A., Mustafa El Hadi W., Rajman M., and Timimi I., "CESTA: First Conclusions of the Technolanguage MT Evaluation Campaign" in *Proc. of LREC'06*, Genoa, Italy, 2006, pp.179–184

[6] Hunt M. "Figures of merit for assessing connected word recognizers" in *Speech Communication 9*, pp. 329–336

[7] ITU-T Recommendations "A Method for Subjective Performance Assessment of the Quality of Speech Output Devices" in *International Telecommunication Union publication*, 1994.

[8] Mostefa D. et al "TC-STAR Deliverable D12: First campaign evaluation report.", 2005

[9] Mostefa D. et al "TC-STAR Deliverable D16: Second campaign evaluation report.", 2006

[10] Salton G. and Buckley C. "Term-weighting approaches in automatic text retrieval" in *Information Processing & Management 24*, 1988, pp.513–523

[11] Softissimo web site http://www.softissimo.com

[12] Systran web site http://www.systran.com

[13] Van den Heuvel H. and Senders E. "Validation of Language Resources in TC-STAR" in *Proc. of the TC-STAR Evaluation Workshop* pp.165-170 http://www.elda.org/tcstar-workshop/

7 Annex A: ASR additional information

7.1 Public training data table

The table 70 gives a list of publicly available data used for training.

Language	Reference	Amount
Chinese	Mandarin 1997 BN (Hub4-NE) LDC98S73 (audio) & LDC98T24 (transcr)	∼30h
	Mandarin 2001 Call (Hub5) LDC98S69, LDC98T26 (transcr)	∼40h
	Mandarin TDT2 LDC2001S93 & LDC2001T57 (transcr)	
	Mandarin TDT3 LDC2001S95 & LDC2001T58	
	Mandarin Chinese News Text LDC95T13	250M words
	Mandarin CALLHOME LDC96S34, LDC96T16 (transcr)	
	Chinese Gigaword LDC2003T09	1.1G words
	Hong Kong News Parallel Text LDC2000T46 (Zh/En)	18147 articles
Spanish	EPPS_SP (text): Apr 1996 - May 2005	>36M words
	TC-STAR_P Spanish BN	10h transcribed
	Spanish LDC 1997, BN speech (Hub4-NE), LDC98S74	
	Spanish LDC CallHome, LDC96S35	
English	EPPS_EN (text): Apr 1996 - May 2005	>36M words
	TC-STAR_P English BN	10h transcribed
	English LDC 1995 (CSR-IV Hub 4 Marketplace LDC96S31), 1996, 1997, official NIST Hub4 training sets, LDC97S44 and LDC98S71, USC Marketplace Broadcast News Speech (LDC99S82)	
	English LDC TDT2 and TDT3 data with closed-captions, about 2000h, LDC99S84 and LDC2001S94	
	English LDC Switchboard 1, 2-I, 2-II, 2-III, LDC97S62, LDC98S75, LDC99S79	
	English LDC Callhome, LDC97S42, LDC2004S05, LDC2004S09	
	English LDC Meeting corpora, ICSI LDC2004S02, ISL LDC2004S05, NIST LDC2004S09	

Table 70: Public condition training resources

Language		Female speakers	Female speakers	Total
Chinese	Number	10	15	25
	Speech duration	0.9h	1.5h	2.4h
	Perplexity	20.1		
English	Number	13	37	50
	Speech duration	0.7h	2.1h	2.8h
	Perplexity	36.3		
Spanish	Number	16	36	52
	Speech duration	1.5h	4.5h	6h
	Perplexity	33.2		

Table 71: Evaluation sets statistics

	Chinese	English	Spanish
Daedalus		2P	
IBM		1O+1R+1P	1O+1R
ITC-irst		4P+1R	1O+1R
LIMSI	1O	1P*	1R*
LIUM		1P+1R+1P	+1R
RWTH		2P+2R	2R
UKA	1O	6P	
UPC			1R
TC-STAR		1P*	2P*

Table 72: Submission table for Chinese, English and Spanish for each training condition (P=Public, R=Restricted, O=Open.).

7.2 Evaluation data statistics table

Table 71 gives an overview of the evaluation data in terms of duration, number of female and male speakers and perplexity.

7.3 Submission table

Submissions marked with a star are late submissions, e.g. submissions received after the official deadline of Jan 28th.

8 Annex C: SLT additional information

8.1 Training data table

The following table gives the training resources used in public training condition.

Direction	Data
Zh->En	FBIS Multilanguage Texts
	UN Chinese English Parallel Text Version 2
	Hong Kong Parallel Text
	English Translation of Chinese Treebank
	Xinhua Chinese-English Parallel News Text Version 1.0 beta 2
	Chinese English Translation Lexicon version 3.0
	Chinese-English Name Entity Lists version 1.0 beta
	Chinese English News Magazine Parallel Text
	Multiple-Translation Chinese (MTC) Corpus
	Multiple Translation Chinese (MTC) Part 2
	Multiple Translation Chinese (MTC) Part 3
	Chinese News Translation Text Part 1
	Chinese Treebank 5.0
	Chinese Treebank English Parallel Corpus
Es->En	EPPS Spanish verbatim transcriptions May 2004 - Jan 2005
	EPPS Spanish Final Text Edition April 1996 to Jan 2005
En->Es	EPPS English verbatim transcriptions May 2004- Jan 2005
	EPPS English Final Text Edition April 1996 to Jan 2005
En<->Es	EU Bulletin Corpus
	JRC-Acquis Multilingual Parallel Corpus
	UN Parallel Corpus

Table 73: Training data for SLT

8.2 SLT development set table

Direction	Data	Epoch
Zh->En	VOA Verbatim transcriptions with 2 references translations	From December 1, 1998 to December 11, 1998
	VOA ASR transcriptions	
	VOA Verbatim transcriptions with 2 references translations	From December 14, 1998 to December 16, 1998
	VOA ASR transcriptions	
	VOA Verbatim transcriptions with 2 references translations	From December 23, 1998 to December 25, 1998
	VOA ASR transcriptions	
Es->En	EPPS verbatim transcriptions with 2 reference translations	From October 25, 2004 to October 28, 2004
	EPPS FTE documents with 2 reference translations	
	EPPS verbatim transcriptions with 2 reference translations	From June 6, 2005 to July 7, 2005
	EPPS ASR transcriptions	

	EPPS FTE documents with 2 reference translations	
	CORTES verbatim transcriptions with 2 reference translations	December 1 & 2, 2004
	CORTES ASR transcriptions	
	CORTES FTE documents with 2 reference translations	
	EPPS verbatim transcriptions with 2 reference translations	From September 5, 2005 to November 17, 2005
	EPPS ASR transcriptions	
	EPPS FTE documents with 2 reference translations	
	CORTES verbatim transcriptions with 2 reference translations	November 24, 2005
	CORTES ASR transcriptions	
	CORTES FTE documents with 2 reference translations	
En->Es	EPPS verbatim transcriptions with 2 reference translations	From October 25, 2004 to October 28, 2004
	EPPS FTE documents with 2 reference translations	
	EPPS verbatim transcriptions with 2 reference translations	From June 6, 2005 to June 9, 2005
	EPPS ASR transcriptions	
	EPPS FTE documents with 2 reference translations	
	EPPS verbatim transcriptions with 2 reference translations	From September 7, 2005 to September 26, 2005
	EPPS ASR transcriptions	
	EPPS FTE documents with 2 reference translations	

Table 74: Development data sets

8.3 SLT evaluation data table

Direction	Data	Epoch
Zh->En	VOA Verbatim transcriptions with 2 references translations	From December 26, 1998 to December 27, 1998
	VOA ASR transcriptions	
Es->En	EPPS verbatim transcriptions with 2 reference translations	From June 12, 2006 to September 28, 2006
	EPPS ASR transcriptions	
	EPPS FTE documents with 2 reference translations	
	CORTES verbatim transcriptions with 2 reference translations	From June 14, 2006 to June 20, 2006
	CORTES ASR transcriptions	
	CORTES FTE documents with 2 reference translations	

En->Es	EPPS verbatim transcriptions with 2 reference translations	From June 12, 2006 to July 4, 2006
	EPPS ASR transcriptions	
	EPPS FTE documents with 2 reference translations	

Table 75: Evaluation data sets

8.4 Participation table

The first table gives the participation of 2007 systems while the second one depicts the submissions of 2005/2006 systems.

Site	En◊Es			Es->En			Zh->En	
	ASR	**FTE**	**Verbatim**	**ASR**	**FTE**	**Verbatim**	**ASR**	**Verbatim**
IBM	2P	2P	2P	1P + 1S	1P + 1S	1P + 1S		
IRST	4P	1P	1P	4P	1P	1P	4P	4P
LIMSI	1P		2P	1P		1P		
RWTH	2P	3P	4P	5P	4P	4P	5P	5P
UKA	3P	3P	3P	3P	3P	3P	2P	4P
UPC	1P + 1S	2P + 1S	1P + 1S	1P + 1S	1P + 1S	1P + 1S		
ICT							7P	9P
JHU				1P	1P + 1S	1P		
NICT-ATR					1P			2P
Tranlendium					1P			
UDS		1P	1P		1P		1S	1P + 1S
XMU							1P + 2S	1P + 2S
ROVER	*2P*	*2P*	*2P*	*2P*	*2P*	*2P*		
Systran	*2P*	*2P*	*2P*	*2P*	*2P*	*2P*	*2P*	*2P*
Reverso	*1P*	*1P*	*1P*	*1P*	*1P*	*1P*		

Table 76: Submissions by condition type (P=Primary; S=Secondary)

Site	En◊Es			Es->En			Zh->En	
	ASR	**FTE**	**Verbatim**	**ASR**	**FTE**	**Verbatim**	**ASR**	**Verbatim**
IBM – 2006	1P	1P	1P					
IRST – 2006	1P	1P	1P	1P	1P	1P	1P	1P
LIMSI – 2006						1P		
RWTH – 2006		1P	1P		1P	1P		
UPC – 2006	1P	1P	1P	1P	1P	1P		

IRST – 2005				1P	1P	1P	1P	1P

Table 77: Submissions of 2005/2006 systems by condition type (P=Primary; S=Secondary)

8.5 Systems outputs statistic table (En→Es)

Table 78 shows some statistics in terms of number of words for the submitted translations of the primary systems and for one reference translation. *"Ref (mean)"* is the mean of words for the references.

Input	**Site**	**number of words**	**words per sentence**	**words src / words trans**
ASR	IBM	27 174	23.29	0.99
	IRST	25 424	21.79	1.06
	LIMSI	26 586	22.79	1.01
	RWTH	27 119	23.24	0.99
	UKA	25 435	21.8	1.06
	UPC	26 487	22.7	1.01
	ROVER	*25 828*	*22.14*	*1.04*
	Systran	*27 502*	*23.57*	*0.98*
	Reverso	*26 178*	*22.44*	*1.03*
	Ref (mean)	*27 869*	*23.89*	*0.96*
Verbatim	IBM	27 616	23.67	0.98
	IRST	26 267	22.51	1.04
	LIMSI	27 227	23.34	1.00
	RWTH	27 025	23.16	1.01
	UKA	26 211	22.47	1.04
	UPC	27 334	23.43	0.99
	UDS	26 804	22.97	1.01
	ROVER	*26 562*	*22.77*	*1.02*
	Systran	*26 971*	*23.12*	*1.01*
	Reverso	*26 805*	*22.97*	*1.01*
	Ref (mean)	*27 869*	*23.89*	*0.98*
Text	IBM	26478	23.44	0.94
	IRST	25 182	22.29	0.99
	RWTH	26 256	23.24	0.95
	UKA	24 631	21.8	1.01
	UPC	26 230	23.22	0.95
	UDS	25 439	22.52	0.98
	ROVER	*26 933*	*23.84*	*0.92*
	Systran	*25688*	*22.74*	*0.97*
	Reverso	*25454*	*22.53*	*0.98*
	Ref (mean)	*27 032*	*23.93*	*0.92*

Table 78: LRs statistics for English-to-Spanish EPPS task

8.6 Systems outputs statistic table (Es→En)

As with English-to-Spanish, we computed some statistics about the average number of words per sentence that are shown in Table 79, for the whole CORTES and EPPS data.

Input	**Site**	**number of words**	**words per sentence**	**words src / words trans**
ASR	IBM	57 018	42.49	1.05
	IRST	57 906	43.15	1.04
	LIMSI	57 106	42.56	1.05
	RWTH	59 286	44.18	1.01
	UKA	55 131	41.09	1.09
	UPC	57 586	42.92	1.04
	JHU	55 156	41.1	1.09
	ROVER	57 064	42.53	1.05
	Systran	58 110	43.31	1.03
	Reverso	59 961	44.69	1.00
	Ref (mean)	*56 017*	41.75	1.07
Verbatim	IBM	57 102	42.55	1.00 00
	IRST	57 974	43.2	0.99
	LIMSI	56 396	42.03	1.01
	RWTH	56 956	42.45	1.00
	UKA	54 829	40.86	1.04
	UPC	57 206	42.63	1.00
	JHU	55 654	41.48	1.03
	ROVER	*56 507*	*42.11*	*1.01*
	Systran	*58 014*	*43.23*	*0.99*
	Reverso	*59 817*	*44.58*	*0.96*
	Ref (mean)	*56 017*	*41.75*	*1.02*
Text	IBM	52 113	35.46	0.97
	IRST	52 961	36.03	0.95
	RWTH	52 939	36.02	0.96
	UKA	49 966	34	1.01
	UPC	53 089	36.12	0.95
	JHU	52 229	35.53	0.97
	NICT-ATR	49 795	33.88	1.02
	Translendium	53 466	36.38	0.95
	UDS	47 824	32.54	1.06
	ROVER	*52 195*	*35.51*	*0.97*
	Systran	*52 997*	*36.06*	*0.95*
	Reverso	*54 569*	*37.13*	*0.93*
	Ref (mean)	*49 907*	*33.96*	*1.01*

Table 79: LRs statistics for the Spanish-to-English task

8.7 Systems outputs statistic table (Zh→En)

Some statistics about the average number of words per sentence are shown in Table 80.

Input	**Site**	**number of words**	**words per sentence**	**words src / words trans**
ASR	IRST	20 754	22.64	0.96
	RWTH	20 466	22.32	0.98
	UKA	19 408	21.17	1.03
	ICT	20 631	22.5	0.97
	UDS	24 482	26.7	0.82
	XMU	20 072	21.89	1.00
	Systran	*23 408*	*25.53*	*0.86*
	Ref (mean)	*22 426*	*24.46*	*0.89*
Verbatim	IRST	20 962	22.64	1.03
	RWTH	20 602	22.86	1.02
	UKA	20 049	22.47	1.04
	ICT	20 750	21.87	1.07
	NICT-ATR	20 692	22.63	1.03
	UDS	22 102	22.57	1.03
	XMU	20 249	24.11	0.97
	Systran	*23 930*	*22.09*	*1.06*
	Ref (mean)	*22 426*	*26.1*	*0.89*

Table 80: LRs statistics for the Chinese-to-English VOA task

9 Annex C: TTS additional information

9.1 Questionnaire for S1

Overall Speech Quality:
"How do you rate the quality of the sound of what you have just heard?"

1. Bad
2. Poor
3. Fair
4. Good
5. Excellent

Listening Effort:
"How would you describe the effort you were required to make in order to understand the message?"

1. No meaning understood with any feasible effort
2. Considerable effort required
3. Moderate effort required
4. Attention necessary; no appreciable effort required
5. Complete relaxation possible; no effort required

Comprehension:
"Did you find certain words hard to understand?"

1. All of the time
2. Often
3. Occasionally
4. Rarely
5. Never

Pronunciation:
"Did you notice any anomalies in pronunciation?"

1. Yes, very annoying
2. Yes, annoying
3. Yes, slightly
4. Yes, but not annoying
5. No

Articulation:
“Were the sounds distinguishable?”

1. No, not at all
2. No, not very clear
3. Fairly clear
4. Yes, clear enough
5. Yes, very clear

Speaking Rate:
“The average speed of delivery was:”

1. Extremely fast or extremely slow
2. Very fast or very slow
3. Fairly fast or fairly slow
4. Slightly fast or slightly slow
5. Just right

Naturalness:
“How do you rate the naturalness of the sound of what you have just heard?”

1. Very unnatural (very odd)
2. Unnatural (odd)
3. Neutral
4. Natural
5. Very natural

Ease of Listening:
“Would it be easy or difficult to listen to this voice for long periods of time?”

1. Very difficult
2. Difficult
3. Neutral
4. Easy
5. Very easy

Pleasantness:
“How would you describe the pleasantness of the voice?”

1. Very unpleasant
2. Unpleasant
3. Neutral
4. Pleasant
5. Very pleasant

Audio Flow:
"How would you describe the continuity or flow of the audio?"

1. Very discontinuous
2. Discontinuous
3. Neutral
4. Smooth
5. Very smooth

9.2 Subjective tests tables

- *E*valuation Task is the identity of the evaluation task (see 4.1).
- *N*umber of subjects gives the number of evaluators who took part to the evaluation task. Not all evaluators were used for each task.
- *N*umber of Evaluation Data gives the total number of audio files used for the evaluation task. The number of submissions per evaluated system is also given (the natural voices are considered as a system here).
- *A*verage number of Tests / Subject is the average number of subjective tests performed by each evaluator who took part to the evaluation task.
- *T*otal number of tests is the total number of subjective tests performed for the evaluation task.

9.3 Data Sets

9.3.1 Development Data Sets

The development set is used for tuning and preparing the system to the evaluation task. Therefore, development data is required to be of the same nature and format as data to be used for the evaluation. ELDA was in charge of the production of the voice conversion development data. Development data are listed in Table 84 .

Eval tasks	***Input/Reference***	***Amount of dev data :***
ENGLISH		

VC	English voice conversion dataset, with 4 different voices (2 male, 2 female). ELDA selected 75% of the data set for evaluation. There are 4 conversion directions: 75 (F) -> 76 (F) 75 (F) -> 79 (M) 80 (M) -> 76 (F) 80 (M) -> 79 (M) 75,76,79,80 voices have been produced by Siemens and UPC. Input data: For each source voice, the participants get: - audio files (channel 1, 96kHz, 24 bits) - xxL files: laringograph output (text files with the time of epoch closure) corresponding to the audio files - xxP files: phoneme segmentation corresponding to the audio files - xxS files: SAM files (text, prosodic information, etc.) corresponding to the audio files	Dev: 126 sentences for each voice
SPANISH		
VC	Same as for the English voice conversion set. ELDA selected 75% of the Spanish VC data set for evaluation.	Dev: 154 sentences for each voice
fCVC	Target voices: ELDA selected audio excerpts of 2 English male speakers in the EPPS 2006 data set: Speaker 01: 249 audio segments (total: 27 min) Speaker 02: 161 audio segments (total: 19 min) Source voice: The source is the IBM TTS voice, based on the TC-STAR baseline voice: spk 73, male. Hence 2 conversion directions: 01->73 02->73 The Spanish translation (Verbatim) of the English voice excerpts were synthesized by IBM. .	Dev: - English audio excerpts - Spanish translations synthesized by IBM

Table 84: TTS development data

Evaluation Task	Number of evaluated systems[1]	Number of subjects	Number of evaluation data	Average number of Tests / Subject	Total number of Tests
S1	8	20	144 (18 sentences / system)	14,4	288
S2	6	20	120 (20 sentences / system)	12,0	240
VC1	9	20	180 (20 sentences / system)	18,0	360
VC2	8	20	160 (20 sentences / system)	16,0	320

Table 81: Information about subjective tests for English

9.3.2 TTS Test Data Sets

The test data sets (the “Inputs”) were sent to the participants. The evaluation corpora are subsets of the whole data sets.
For each task, each participant processed the whole test data set and sent its results back to ELDA. ELDA performed the evaluations using the evaluation subsets only.
ELDA was in charge of the test data production. Test data sets are reported in Table 85.

Eval tasks	***Input/Reference***	***Amount of data : Input / Evaluation corpus***
ENGLISH		
S1	40 paragraphs selected by ELDA from the English EPPS FTE, year 2006. Format: SSML / Unicode UTF-8.	Input: 40 paragraphs Eval: 18 paragraphs
S2	160 sentence selected by ELDA from the output of the ASR + SLT ROVER system. Input data English EPPS, year 2006. Format: SSML / Unicode UTF-8.	Input: 160 sentences Eval: 20 sentences

VC	English voice conversion dataset, with 4 different voices (2 male, 2 female). ELDA selected 25% of the data set for evaluation. There are 4 conversion directions: 75 (F) -> 76 (F) 75 (F) -> 79 (M) 80 (M) -> 76 (F) 80 (M) -> 79 (M) 75,76,79,80 voices have been produced by Siemens and UPC. Input data: For each source voice, the participants get: - audio files (channel 1, 96kHz, 24 bits) - xxL files: laringograph output (text files with the time of epoch closure) corresponding to the audio files - xxP files: phoneme segmentation corresponding to the audio files - xxS files: SAM files (text, prosodic information, etc.) corresponding to the audio files	Input: 42 sentences for each voice Eval: 5 sentences per conversion direction
SPANISH		
S1	40 paragraphs selected by ELDA from the Spanish EPPS FTE, year 2006. Format: SSML / Unicode UTF-8.	Input: 40 paragraphs Eval: 18 paragraphs
S2	160 sentence selected by ELDA from the output of the ASR + SLT ROVER system. Input data Spanish EPPS, year 2006. Format: SSML / Unicode UTF-8.	Input: 160 sentences Eval: 20 sentences
VC	Same as for the English voice conversion set. ELDA selected 25% of the Spanish VC data set for evaluation.	Input: 52 sentences for each voice Eval: 5 sentences per conversion direction
fCVC	ELDA selected 50 sentences from the Spanish EPPS-FTE 2006 corpus. These data were sent to IBM who synthesized them (Spanish voice 73(M)). .	Input: 50 Spanish sentences synthetized by IBM Eval: 5 sentences per conversion direction
CHINESE		
S1	37 paragraphs selected by ELDA from the "863 program" data set. Format: UTF-8 encoding, SSML format	Input: 37 paragraphs Eval: 12 paragraphs

Table 85: TTS test data

Evaluation Task	Number of Evaluated Systems	Number of Subjects	Number of Evaluation Data	Average Number of Tests / Subjec	Total Number of Tests
S1	9	20	162 (18 sentences / system)	16,2	324
S2	7	20	140 (20 sentences / system)	14,0	280
VC1	12	20	240 (20 sentences / system)	24,0	480
VC2	11	20	220 (20 sentences / system)	22,0	440
fCVC1	5	20	50 (10 sentences / system)	5,0	100
fCVC2	4	20	40 (10 sentences / system)	4,0	80

Table 82: Information about subjective tests for Spanish

Evaluation Task	**Number of Evaluated Systems**	**Number of Subjects**	**Number of Evaluation Data**	**Average Number of Tests / Subject**	**Total Number of Tests**
S1	4	11	48 (12 sentences / system)	12,0	132

Table 83: Information about subjective tests for Chinese

9.4 Detailed Results of the TTS component evaluation (S1, S2)

This section is a more detailed presentation of the TTS component evaluation results. For the 3 languages, we give the results obtained in the 10 judgment categories of the S1 evaluation, and the sentence error rate (SER) obtained in the S2 evaluation.

9.4.1 Detailed Results for English

Table 86 gives the results of judgment tests S1 carried out on TTS systems taken as a whole. Judges had to rate the synthesized voices according to the 10 categories mentioned below, using 5 point-scales (in all cases: ‘5’ represents the best score and ‘1’ the worse).
The results for S2 are reported in and Table 87.

Legend:

Judgment categories:

OQ: Overall Quality, **LE**: Listening Effort, **Pr**: Pronunciation; **C**: Comprehension, **A**: Articulation, **SR**: Speaking Rate, **N**: Naturalness, **EL**: Easy of Listening, **Pl**: Pleasantness, **AF**: Audio Flow.

S1										
System	**OQ**	**LE**	**Pr**	**C**	**A**	**SR**	**N**	**EL**	**Pl**	**A**
Scoring (1<5)										
NAT	4.59	4.73	4.89	4.89	4.86	4.65	4.43	4.41	4.24	4.46
IBM_F_06	3.00	3.42	3.95	4.21	3.42	4.32	2.42	2.68	3.11	2.47
IBM_F	3.42	3.89	4.14	4.17	3.64	4.61	3.22	3.25	3.50	3.03
IBM_M	3.49	3.78	3.76	4.19	3.54	4.41	2.92	3.00	3.24	2.73
SIE_F	2.31	2.91	3.46	3.34	2.71	3.89	2.23	2.06	2.91	2.26
SIE_M	1.58	2.26	3.05	2.71	2.13	3.42	1.63	1.50	2.08	1.47
UPC_F	2.86	3.31	3.61	3.72	3.11	4.14	2.67	2.61	3.06	2.22
UPC_M	2.74	3.15	3.21	3.44	2.94	4.00	2.18	2.24	2.62	2.12
Ranking										
NAT	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
IBM_F_06	*4*	*4*	*3*	**2**	*4*	*4*	*5*	*4*	*4*	*4*
IBM_F	*3*	**2**	**2**	*4*	**2**	**2**	**2**	**2**	**2**	**2**
IBM_M	**2**	*3*	*4*	*3*	*3*	*3*	*3*	*3*	*3*	*3*
SIE_F	*7*	*7*	*6*	*7*	*7*	*7*	*6*	*7*	*6*	*5*
SIE_M	*8*	*8*	*8*	*8*	*8*	*8*	*7*	*8*	*8*	*8*
UPC_F	*5*	*5*	*5*	*5*	*5*	*5*	*4*	*5*	*5*	*6*
UPC_M	*6*	*6*	*7*	*6*	*6*	*6*	*8*	*6*	*7*	*7*

Table 86: Results of the TTS component evaluation S1 (English)

The results for S2 are reported in and Table 87.

Legend:

WER Word Error Rate.
SER Sentence Error Rate.

S2				
System	**WER**		**SER**	
	Score	**Rank**	**Score**	**Rank**
IBM_F	12.8	*3*	71.1	4
IBM_M	12.4	*2*	57.9	2
SIE_F	14.8	*5*	76.3	5

SIE_M	22.2	*6*	78.9	6
UPC_F	8.7	***1***	52.6	***1***
UPC_M	14.5	*4*	69.4	3

Table 87: Results of the TTS component evaluation S2 (English)

9.4.2 Detailed Results for Spanish

Table 88 gives the results of judgment tests S1 carried out on TTS systems taken as a whole. Judges had to rate the synthesized voices according to the 10 categories mentioned below, using 5 point-scales (in all cases: '5' represents the best score and '1' the worse).
The results for S2 are reported in and Table 89.

Legend:

NAT Natural voice, used as top-line in subjective tests.
IBM_F_06 Female voice submission made by IBM last year (re-evaluated this year)
IBM_F/M IBM submission using female / male voices
UPC_F/M UPC submission using female / male voices
VER_F1/M1 Verbio submission using female / male voices
VER_M2 2^{nd} Verbio submission using male voice

Judgment categories:

OQ: Overall Quality, **LE**: Listening Effort, **Pr**: Pronunciation; **C**: Comprehension, **A**: Articulation, **SR**: Speaking Rate, **N**: Naturalness, **EL**: Easy of Listening, **Pl**: Pleasantness, **AF**: Audio Flow.

S1										
System	**OQ**	**LE**	**Pr**	**C**	**A**	**SR**	**N**	**EL**	**Pl**	**A**
Scoring (1<5)										
NAT	4.75	4.67	4.92	4.86	4.86	4.83	4.61	4.28	4.39	4.36
IBM_F_06	3.89	4.00	4.11	4.44	3.97	4.08	2.50	2.86	3.31	2.31
IBM_F	4.00	4.19	4.11	4.67	4.19	4.36	2.97	3.28	3.39	2.75
IBM_M	4.00	4.11	4.37	4.49	4.26	4.49	3.26	3.43	3.63	3.09
UPC_F	3.42	3.86	3.92	4.44	3.94	4.03	2.50	2.89	3.17	2.39
UPC_M	3.47	3.94	3.83	4.44	4.08	4.50	2.81	3.11	3.25	2.81
VER_F1	4.22	4.22	4.36	4.61	4.25	4.53	3.25	3.50	3.75	3.19
VER_M1	4.06	4.22	4.25	4.44	4.14	4.31	3.17	3.47	3.47	3.28
VER_M2	3.94	4.11	4.22	4.67	4.17	4.64	3.11	3.22	3.42	3.14
Ranking										
NAT	1	1	1	1	1	1	1	1	1	1
IBM_F_06	7	7	6	6	8	8	8	9	7	9
IBM_F	4	4	7	**2**	4	6	6	5	6	7
IBM_M	5	5	2	5	**2**	5	2	4	3	5
UPC_F	9	9	8	7	9	9	9	8	9	8
UPC_M	8	8	9	8	7	4	7	7	8	6
VER_F1	**2**	**2**	3	4	3	3	3	**2**	**2**	3
VER_M1	3	3	4	9	6	7	4	3	4	**2**
VER_M2	6	6	5	3	5	**2**	5	6	5	4

Table 88: Results of the TTS component evaluation S1 (Spanish)

The results for S2 are reported in and Table 89.

Legend:

WER Word Error Rate.

SER Sentence Error Rate.

S2				
System	**WER**		**SER**	
	Score	**Rank**	**Score**	**Rank**
IBM_F	7.5	*3*	37.5	*3*
IBM_M	12.1	*6*	53.8	*7*
UPC_F	7.1	*2*	33.3	*2*
UPC_M	6.0	***1***	37.5	*3*
VER_F1	12.2	*7*	50.0	*6*
VER_M1	9.7	*5*	42.5	*5*
VER_M2	8.4	*4*	30.0	***1***

Table 89: Results of the TTS component evaluation S2 (Spanish)

9.4.3 Detailed Results for Chinese

Table 90 gives the results of judgment tests S1 carried out on TTS systems taken as a whole. Judges had to rate the synthesized voices according to the 10 categories mentioned below, using 5 point-scales (in all cases: ‘5’ represents the best score and ‘1’ the worse).

Legend:

NAT Natural voice, used as top-line in subjective tests.

NOK_06 Submission made by Nokia last year (re-evaluated this year).

Judgment categories:

OQ: Overall Quality, **LE**: Listening Effort, **Pr**: Pronunciation; **C**: Comprehension, **A**: Articulation, **SR**: Speaking Rate, **N**: Naturalness, **EL**: Easy of Listening, **Pl**: Pleasantness, **AF**: Audio Flow.

S1										
System	**OQ**	**LE**	**Pr**	**C**	**A**	**SR**	**N**	**EL**	**Pl**	**A**
Scoring (1<5)										
NAT	4.19	4.62	4.73	4.92	4.77	4.77	4.00	3.85	3.69	4.23
CAS	3.86	3.75	3.57	4.29	3.64	4.54	2.86	2.89	3.04	2.96
NOK	2.85	2.31	2.50	3.38	2.77	3.69	2.08	2.12	2.19	2.42
NOK_06	2.61	2.74	2.57	3.70	2.91	3.70	2.04	2.17	2.04	1.96
Ranking										
NAT	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
CAS	***2***	***2***	***2***	***2***	***2***	***2***	***2***	***2***	***2***	***2***
NOK	*3*	*4*	*4*	*4*	*4*	*4*	*3*	*4*	*3*	*3*
NOK_06	*4*	*3*	*3*	*3*	*3*	*3*	*4*	*3*	*4*	*4*

Table 90: Results of the TTS component evaluation S1 (Chinese)