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.: D16
Title: Evaluation Report

Due date of the deliverable:	31 September 2006
Actual submission date:	*June 2006 (version 1)* *September 2006 (version 2)*
Start date of the project:	1st of April 2004
Duration:	36 months
Lead contractor for this deliverable:	ELDA
Authors:	D. Mostefa (ELDA), M.-N. Garcia (ELDA), O. Hamon (ELDA), N. Moreau (ELDA)

Revision: Final 1.19

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

Revision history:

Version	Date	Changes	Editor
0.5	2006-05-05	Version circulated to PMC	Djamel Mostefa
0..6	2006-05-17	Integration of ITC-irst feedback	Djamel Mostefa
1.0	2006-05-30	Final document	Djamel Mostefa
1.1	2006-05-30	Human evaluation results for SLT added	Djamel Mostefa
1.2	2006-06-07	Formal correction and TTS results updated	Djamel Mostefa
1.8	2006-09-29	TTS updated	Djamel Mostefa
1.11	2006-10-13	TTS updated	Djamel Mostefa
1.12	2006-10-25	ASR updated	Djamel Mostefa
1.13	2006-10-31	TTS and SLT updated	Djamel Mostefa
1.15	2006-12-06	SLT/TTS updated with comments from ITC, RWTH and UPC	Djamel Mostefa
1.16	2006-12-07	End-to-end section update with comments from UPC	Djamel Mostefa
1.17	2006-12-14	TTS updated with comments from UPC	Nicolas Moreau
1.18	2006-12-15	TTS updated with comments from UPC; global review	Nicolas Moreau
1.19	2006-12-15	Final revision	Djamel Mostefa

1 Introduction

This document reports on the evaluation activities carried out in TC-STAR as part of the second TC-STAR Evaluation Campaign. This campaign took place during Month 25 of the project, more precisely from 1 February 2006 to 15 March 2006. The results of the evaluation campaign were presented at the Second TC-STAR evaluation Workshop[1] which was held in Barcelona in June 2006

The aim of the evaluation campaign was to measure the progress made during the second year of the project in:

- Automatic Speech Recognition (ASR),
- Spoken Language Translation (SLT),
- Text To Speech (TTS) processing.
- Integration of components (ASR+SLT, ASR+STL+TTS)

1.1 Evaluation Tasks

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

- **European Parliament Plenary Sessions (EPPS)**: the evaluation data consisted of audio recordings of the EPPS original channel[2] 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 was exclusively on the Parliament Members speaking in English and in Spanish, therefore the interpreters speeches were not used this year. These resources were 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, we decided to use audio recordings of the Spanish Parliament (Congreso de Los Diputados). The data were used in addition to the EPPS Spanish data to evaluate ASR in Spanish and SLT from Spanish into 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 were used to evaluate speech recognition systems in Mandarin Chinese and translation from Mandarin into English (Zh→En).

1.2 Participants

This year in addition to the TC-STAR partners, many external participants joined the evaluations.

The list of participants in the Second TC-STAR Evaluation Campaign is given below.

- Internal participants:

[1] http://www.elda.org/tcstar-workshop

[2] this channel includes speeches of Member of the Parliament in their original language

  - IBM, Germany
  - Istituto Trentino di Cultura - Il Centro per la ricerca scientifica e tecnologica (ITC-irst), Italy
  - Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), France
  - Nokia, Finland
  - Rheinisch-westfälische Technische Hochschule (RWTH) Aachen, Germany
  - Siemens, Germany
  - Sony, Germany
  - Universität Karlsruhe (UKA), Germany
  - Universitat Politècnica de Catalunya (UPC), Spain
- External participants :
  - DFKI: Deutsches Forschungszentrum für Künstliche Intelligenz, Germany
  - ICT: Institute of Computing Technology, China
  - NLPR: National Laboratory of Pattern Recognition, China
  - NRC: National Research Council, Canada
  - UED: University of Edinburgh
  - UW: University of Washington, United States
  - CAS : Chinese Academy of Science, China
  - IBM China
  - UD : University of Dresden, Germany
  - UM : University of Munich, Germany
  - UV : University of Vigo, Spain

Table 1 gives an overview of participation for Automatic Speech Recognition, Spoken Language Translation and Text To Speech. External participants are in bold.
Moreover, in order to compare SLT results with a commercial product, we have computed the SLT scores of a commercial off-the-shelve SYSTRAN product.

	Automatic Speech Recognition			Spoken Language Translation			Text To Speech		
	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				
NOKIA	X						X		X
RWTH	X	X		X	X	X			
SIEMENS							X	X	
SONY									
UKA	X		X	X	X	X			
UPC				X	X		X	X	
ATT							X	X	
CAS									X
DFKI				X	X				
ICT						X			
NLPR						X			
NRC						X			
U. Edinburgh				X	X				
Univ. Dresden							X		
Univ. Munich							X		
U. Vigo		X							
U. Washington				X	X				

Table 1 Participant in the Second TC-STAR Evaluation Campaign

2 ASR Evaluation

2.1 Tasks and conditions

There were three tasks and three different training conditions for each task:

1. For the EPPS task, automatic speech recognition systems were evaluated on recordings of the Parliament's sessions in English and Spanish recorded in September-December 2005.
2. For the CORTES task, recordings from November 2005 were used for the evaluation.
3. For the Mandarin language, VOA task, broadcast news recordings of December 1998 of the radio "Mandarin Voice of America" were used.

For each task, three training conditions were 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 was 31st of May 2005 for EPPS and CORTES. Systems were not allowed to use any training data (audio recordings, text data, etc) produced after the 31st of May 2005.

For VOA, a black-out period covering December 1998 was defined, rather than a cut-off date.

Segmentation

Unlike the first evaluation, no manual segmentation in sentences was provided for the Spanish Parliament data and for the VOA broadcast news recordings.

A manual segmentation was 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 were used:

- Word Error Rate (WER) for the EPPS task,
- Character Error Rate (CER) for the VOA task

For Spanish and English, the scoring was 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 minimizes the misalignment of two strings of words [1].

Three kinds of errors are taken into account in computing the word error rate, i.e. substitution, deletion and insertion errors:

- Substitution: a reference word is replaced by another word in the best alignment between the reference and the system hypothesis.
- Deletion: a reference word is not present in the system hypothesis in the best alignment.
- Insertion: Some extra words are present in the system hypothesis in the best alignment between the reference and the hypothesis.

The SCTK software toolkit developed by NIST was used for scoring.

2.2 Language resources for ASR

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

2.2.1 ASR Training Data Sets

- **Restricted condition**

For the *restricted* condition, only data produced within TC-STAR could be used for training purposes. This data was produced on recordings of the European Parliament from 3 May 2004 to 26 May 2005. The audio files were recorded and provided by RWTH. The manual transcriptions of the English recordings were done and provided by RWTH, while those of the Spanish recordings were done and provided by UPC.
In addition, for the EPPS tasks, the Final Text Edition (FTE) of the documents published by the EC, from April 1996 to May 2005, were downloaded and provided by RWTH.
In addition to the EPPS data, 40 hours of the CORTES Spanish parliament were recorded and transcribed by UPC.

	Transcribed		Non transcribed	TOTAL
	Politicians	Interpreters		
EPPS English	21h	70h	75h	**166h**
EPPS Spanish	10h	51h	90h	**151h**
CORTES Spanish	40h			**40h**

Table 2 Training resources for the restricted condition

Public condition

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

Table 2 summarizes the data that could have been used for training in the public condition by the participants.

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 3 Public condition training resources

Open condition

For the *open* condition, any data before the cut-off date could be used.

2.2.2 ASR Development Data Sets

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

For EPPS tasks, the development data consisted of audio recordings (in English and Spanish) of Parliament's sessions from the 6th of June to the 7th of July 2005, manually transcribed by ELDA. In each language, 3 hours of recordings were selected and transcribed, corresponding to approximately 35,000 running words in English and 33,000 running words in Spanish. Only English- (resp. Spanish-) speaking politicians were transcribed (e.g. no interpreters data). ELDA also provided the corresponding Final Text Editions, which are the official transcriptions of the parliamentary debates, published by the EC in English and Spanish.

For the CORTES task, audio recordings of the Spanish Parliament's sessions from the 1st and 2nd of December 2004 were used.

For the VOA task, the development data consisted of 3 hours of audio recordings from the broadcasted news of Mandarin Voice of America between 1 and 11 December 1998. It corresponded to approximately 42,000 Chinese characters. ELDA produced the manual transcriptions.

Table 4 gives some statistics about the development set. Usually it is considered that a development or evaluation set is speaker independent if the speaker perplexity is higher than 20, which is the case for Spanish and English.
The speaker perplexity Px is given by:

$$Px = \exp(-\sum_{i=1}^{n} p_i * \log(p_i))$$

where: p_i is the proportion of speaker i and n is the number of speakers.
If each speaker is equally represented in the data set, then $Px = n$

Data		Total	Male speakers	Female speakers
Spanish (EPPS + CORTES)	#Speakers	61	44	17
	Duration	5.8h	79.69%	20.31%
	Perplexity	34.12		
English (EPPS)	#Speakers	41	32	9
	Duration	3h	75.26%	24.74%
	Perplexity	20.51		

Table 4: Development set statistics

2.2.3 ASR Test Data Set

The same general procedure was followed to produce the test data as the one used to produce the development data.

Here we only highlight the differences with the development data set:

- For EPPS tasks, the Parliamentary sessions from which the audio recordings were selected ran from September to November 2005.
- For the CORTES task, recordings from November 24, 2005 were used.
- For VOA tasks, the data was selected from news broadcasted between 23 and 25 December 1998.

2.2.4 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 were successfully validated by SPEX. More details can be found in [11]

2.3 Schedule

The development phase took place from 10 November 2005 to 1 February 2006.
The ASR Run took place from 1 February to 12 February 2006.

2.4 Participants and Submissions

There were 8 participating sites in the ASR evaluation, 7 from the TC-STAR consortium and one external participating site. Each participant had to submit for evaluation the output of at least one system trained under one of the specified conditions (i.e. open, public, or restricted). There were 33 different submissions: 22 for English, 10 for Spanish and 1 for Mandarin.
Table 5 reports the number of submissions (i.e. ASR system outputs submitted for evaluations) for the English EPPS task by different sites and under the different training conditions.

	Open	Public	Restricted
IBM		*1*	*3*
ITC-irst		*1*	
LIMSI		*3*	
NOKIA	*5*		
RWTH			*2*
SONY	*2*		
UKA		*1*	*1*
TC-Star[3]		*3*	
Total	*7*	***9***	***6***

Table 5: Number of submissions for the English EPPS task

In Spanish, there were 10 submissions all in restricted condition:

- 1 by IBM
- 1 by ITC-irst
- 2 by LIMSI
- 1 by RWTH
- 4 by University of Vigo
- 1 by TC-STAR

In Chinese, there was one joint submission by LIMSI/UKA in the restricted training condition.

2.5 Evaluation Results

For a particular condition (i.e. open/public/restricted) sites could submit results for different systems' configurations, but one of them had to be specified as the site's primary system.

Other submissions are considered as contrasts to the primary system. Each site estimated the processing time required by its system on a single processor platform.

2.5.1 English ASR Results

We received 22 different submissions from 7 participating sites.

Table 6 presents the results obtained by the primary systems.

The best results were obtained by the TC-STAR system combination with a Word Error Rate of 6.9% in public condition. The TC-STAR combination uses the Recognizer Output Voting Error Reduction (ROVER) method [3]. The ROVER system is able to reduce error rates by exploiting differences in the nature of the errors made by multiple ASR systems.

[3] The TC-STAR submissions are combination of multiple ASR outputs

Site	Open	Public	Restricted
IBM*		8.8%	
ITC-irst		11.0%	
LIMSI*		8.2%	
NOKIA	18.3%		
RWTH			10.2%
SONY	37.1%		
UKA		14.0%	
*TC-STAR**		6.9%	

Table 6: Primary system results, in terms of WER for the English EPPS task

2.5.2 Spanish ASR Results

All submissions were in the *restricted* training conditions. This might be explained by the small number of publicly available resources usable for this task.
We received 10 different submissions from 5 participants: IBM, ITC-irst, LIMSI, RWTH and University of Vigo.

The performance of primary systems is between 10.2% and 28.4%. Again a ROVER combination of all hypotheses was performed by LIMSI.
The ROVER gave the best result with a WER of 8.1%, (6.2% on the EPPS data).
A summary of these results from the 10 submissions can be seen in Table 7

	EPPS+CORTES	EPPS	CORTES
IBM*	10.6%	8.3%	12.5%
ITC-irst	13.5%	9.7%	16.8%
LIMSI*	10.7%	7.8%	13.3%
RWTH*	10.2%	8.0%	12.1%
TC-STAR*	8.1%	6.2%	9.8%
U. VIGO	28.4%	20.1%	35.7%

* Late submission (the system output was sent to ELDA after the official deadline of Feb 12th, 2006)

Table 7 Spanish ASR results

2.5.3 Chinese ASR Results

There was 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 CER is 9.8%.

2.6 Error analysis

Here we focus on an error analysis of the ROVER combinations for Spanish and English. This is done for two reasons. One the one hand, the ROVER gave the best results (6.9% for English and 9.8% for Spanish). On the other hand, the ROVER combinations were used as input for the Spoken Language Translation systems.

2.6.1 Female versus male speakers

Systems performances are better on male speakers than on female ones for Spanish and English. This is mainly due to the fact that there are more male speakers in the European Parliament. For example, only 25% of the data are spoken by women in the English evaluation set. We can expect the same ratio in the training sets, so male speakers' models are better trained and then perform better than the female ones.

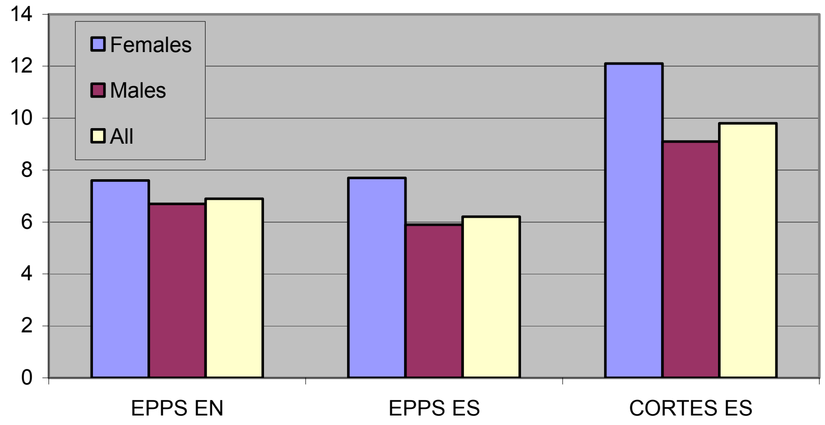


Figure 1 TC-STAR system performance for male and female speakers

2.6.2 Short words

Most errors are substitution ones for which the recognizer supplied an incorrect word for a reference word. This is especially true for short words composed by only one or two phonemes (*a, and, has, is, its, his,* for English, or *al, el, en, y, lo* for Spanish).
For example, the ten most common substitutions for the ROVER combinations for English, Spanish and Chinese are given in the Table 8. The first word is the reference word and the second is the wrongly substituted one.

Confusion pairs for English	Confusion pairs for Spanish	Confusion pairs for Mandarin
a / the	las / la	她 / 他
and / in	Del / el	了 / 的
the / a	el / del	它 / 他
(%hesitation) / and	(%hesitation) / de	的 / 地
that / the	(%hesitation) / que	利 / 力
the / that	del / de	是 / 时
or / all	el / al	作 / 做
too / to	(%hesitation) / en	地 / 的
been / being	al / el	呢 / 的
had / have	de / del	是 / 使

Table 8: Top ten substitution errors for the English and Spanish EPPS task and the Mandarin VOA task

(%hesitation) represents small words that are commonly used to fill in the sentence in spontaneous speech while the speaker is thinking about what he will say next. For English (%hesitation) represents words like *uhm, oh, eh, ah*, .etc.

Here is an example of a substitution error that occurred in the EPPS English task. The word *"our"* was recognized as *"a"*. We can notice that the two words are quite close phonetically and that the system's hypothesis is semantically and syntaxically correct.

In this example and the following ones, words in capital letters denote words that have not been well recognized by the system.

Example:

```
REF:  is that in OUR globalised world no single country can tackle these
problems alone even in their own country

HYP:  is that in A   globalised world no single country can tackle these
problems alone even in their own country
```

2.6.3 Accents and speaking style

If we look to the performance of the TC-STAR system for English speaker by speaker we can see that the worse performance are obtained on non-native speakers or speakers with a strong accent.

For example the worse performances are obtained on Mr. Charlie McCreevy speeches with a WER of 19.2%. Mr. McCreevy speaks with a strong Irish accent and moreover quite fast. The second worse performances are obtained on the Hungarian Member of the Parliament Mrs Zita Gurmai with a WER of 17.7%.

Another factor on the performances is the fluency of the speech. Usually speeches are well prepared, quite close to read speech and the speaker speaks fluently. But in some cases the speaker hesitates a lot, makes corrections and false starts which make the recognition task difficult. For example here is the recognition output of the TC-STAR system on a speech of Mrs María del Pilar Ayuso González.

```
REF:  EN   LOS   TRES SUBTIPOS de de virus A QUE HEMOS TENIDO EN EL H
UNO    EN EL     dieciocho en el H DOS EN     EL cincuenta y siete Y EN
EL H TRES EN    EL    sesenta y ocho
```

```
HYP:  PERO USTED SUS  TIPOS    de de virus * *** ***** ****** ** **
SAQUEMOS TENÍA LA SEGUNDA dieciocho en el * *** ACCESO DE cincuenta y
siete * ** ** * **** TIENE SIETE sesenta y ocho
```

2.7 Final remarks on the Second ASR Evaluation Campaign

All sites that are involved in ASR (WP2) participated in this evaluation.
One external participant joined the evaluation for Spanish. Thirty-three system outputs were submitted (twenty-two for English, ten for Spanish and one for Chinese).
Rover combinations were performed for English and Spanish. The best word error rate is 6.9% for English, 8.1% for Spanish (6.2% on EPPS) and 9.8% for Chinese.

There has obviously been impressive progress during Year 2 of TC-STAR. For English and Spanish, there was a 40% decrease in WER of the TC-STAR system (from 9.9% to 6.9% for English and 10.7% to 6.2% for Spanish EPPS). Figure 2 shows the best results obtained by the TC-STAR partners in ASR since the beginning of the project. Please note that the results are obtained on:

- EPPS data for English
- EPPS data for Spanish for Year 1 and Year 2 and TC_STAR_P data for the Baseline (see [5] for details)
- VOA data for Year 1 and Year 2 and other broadcast news recordings for the baseline (see [5] for details)

To be totally complete, we must say that the EPPS evaluation sets for this year are a little bit easier than the previous one since no interpreter data was considered this year and that the WER is lower on politicians' speeches [8].

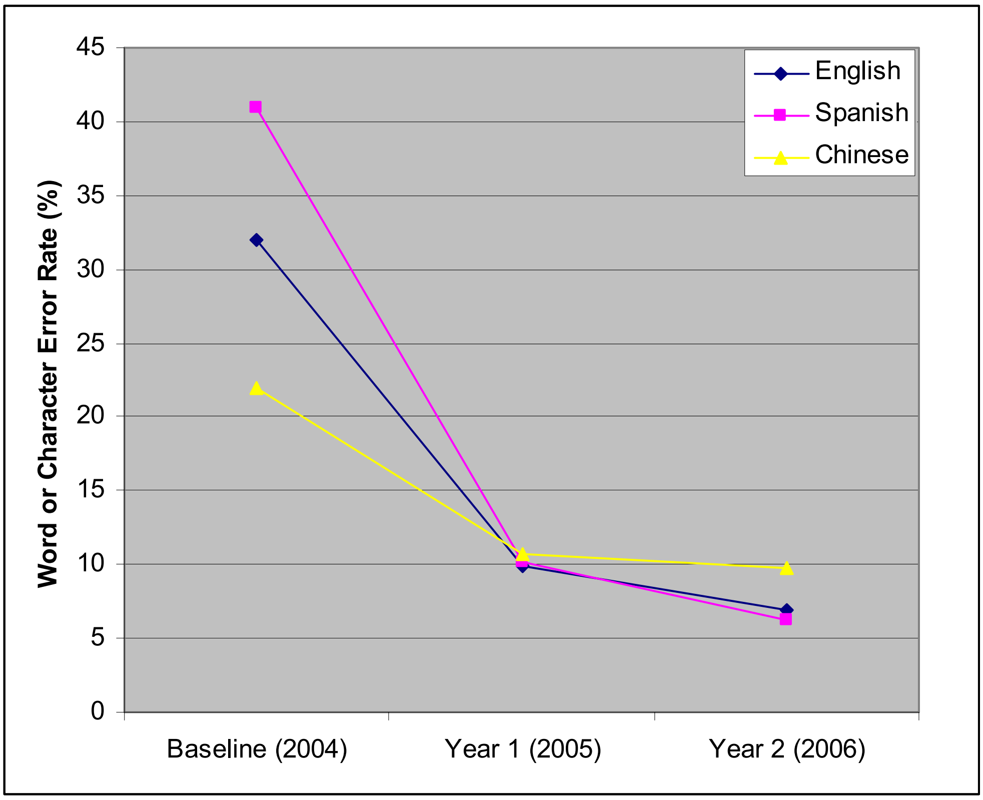


Figure 2 ASR progress

Nevertheless, the WER must still be further reduced, as SLT systems need even lower error rates, especially as machine translation models get better.

There is still plenty of room for improvement of ASR systems at all levels. As a general direction for Year 3, these improvements can come from:

- The availability of more training data, especially large amount of non-transcribed data.
- More collaboration amongst partners, not only within WP2 but also between WP2 and WP1, especially by using not only the best hypothesis but also word graph outputs from ASR.
- Better combination of systems, e.g. by using system cascade and ROVER with combination.

The next evaluations will take place on December 2006 and will focus on the same tasks and languages. More training data will be available and external participants will be attracted to the ASR evaluation.

2.8 ASR Evaluation packages

An evaluation package which includes resources, protocols, scoring tools, results of the ASR official campaign, etc., those were 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 the campaign itself. Three evaluation packages (one per language) are available on ELRA's catalog of language resources [12]

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 5th September and the 17th of November 2005 were 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 the 24th November 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 were 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 were used as input:

1. The first one is the output of a combination of some automatic speech recognition systems. The ASR ROVER combination, which gave the lowest error rate, was used. The text was in true case and punctuation marks were provided. This year no manual segmentation in sentences was provided and the SLT systems have to segment the ASR output automatically. Then the SLT output data was 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.
2. 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 were produced for English and Spanish. As for the ASR output, the text data was provided with punctuation and in true case. This type of data is called “Verbatim” in the results parts.
3. The last one is the text data input. Final Text Editions (FTE), provided by the European Parliament and the Spanish Parliament, were 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 punctuations, uppercase and lowercase and do not include transcription of spontaneous speech phenomena.

An example of the three kinds of inputs is shown below:

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*

For Chinese-to-English direction, two kinds of text data were used as input:

1. The first one is the output of the automatic speech recognition systems. The common submission from LIMSI/UKA was used. No punctuation marks were provided. Again this year no manual segmentation in sentences was provided and the SLT output data was automatically aligned to the reference translations for scoring.
2. 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 was provided without punctuation.

As for the ASR evaluations, different training conditions were distinguished. The first one was the primary condition in which systems could only use the data produced within TC-STAR and the LDC Large Data listed in the Table 9. The aim is to have strict comparisons of systems.
In the secondary condition, any publicly available data before the cut-off date (May 31, 2005) could be used for training purposes.

3.2 Language Resources for Spoken Language Translation

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

3.2.1 SLT Training Data Sets

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

For the EPPS task, the training data consisted of the same data as for ASR training: the Final Text Editions (FTE), in Spanish and English, from April 1996 to January 2005, provided by RWTH. They were considered as reference translations from each other to train the systems. The EPPS data was sentence-aligned. Additionally, the manual verbatim transcriptions of the EPPS recordings in English and Spanish from May 2004 to January 2005 were provided by RWTH (English) and UPC (Spanish).

The Table 9 below summarizes all the data used for training by the participants.

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 and English Final Text Edition April 1996 to Jan 2005
En->Es	EPPS English verbatim transcriptions May 2004- Jan 2005
	EPPS English and Spanish Final Text Edition April 1996 to Jan 2005

Table 9 Training data used

3.2.2 SLT Development Data Sets

The SLT development set was built upon the ASR development data set, in order to enable end-to-end evaluation. Subsets of 25,000 words were 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 were 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 were translated into Spanish by 2 different agencies;
- EPPS Spanish verbatim transcriptions and FTE documents were translated into English by 2 different agencies;
- VOA verbatim transcriptions were translated into English by 2 different agencies;
- CORTES Spanish verbatim transcriptions and FTE documents were translated into English by 2 different agencies.

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

The development data for the ASR task were provided using the outputs of the ASR systems. A ROVER combination has also been provided. The corresponding references were those of the verbatim development data. All source text sets were 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 10.

Direction	**Data**	**Epoch**

Zh->En	VOA Verbatim transcriptions with 2 references translations	From December 14, 1998 to December 16, 1998
	VOA ASR transcriptions	
Es->En	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	
En->Es	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	

Table 10 Development data sets

3.2.3 SLT Test Data Sets

As for development, the same procedure was followed to produce the test data. The corresponding data sets used are summarized in Table 11.

Direction	Data	Epoch
Zh->En	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 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 September 7, 2005 to September 26, 2005
	EPPS ASR transcriptions	
	EPPS FTE documents with 2 reference translations	

Table 11 Test data sets

3.2.4 Summary

As a whole, we have 22 data sets. For a given set, there are:

- The data to be translated in the source language, organized 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 organized in documents and segments,
- Several candidate translations produced by the participants in the evaluation, following the same format of the source and reference sets.

3.2.5 Validation of Language Resources

SPEX validated the reference translations of the development and test sets for all three translation directions: English-to-Spanish, Spanish-to-English and Mandarin-to-English.

For each set of each translation direction, and for each reference translation (each set was 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 were taken from the target text). Half of the 1200 words were selected from the FTE sources and half from the VERBATIM sources.

Translation errors were then scored using the following penalty scheme:

Error	Penalty points
Syntactical	4 points
Deviation from guidelines	3 points
Lexical	2 points
Poor usage	1 point
Punctuation or spelling errors	0.5 point (with a maximum of 10 points)

Table 12 LRs translation errors

The validation criterion is that a reference translation must have less than 40 penalty points to be considered valid.

The validation results for the reference translations are reported in Table 13.

Direction	Dev		Test	
	Ref 1	Ref 2	Ref 1	Ref 2
En→Es	52.5	30.5	40 R	28.5 R
Es→En (EPPS)	34.5	40	39.5 R	38 R
Es→En (CORTES)	22 R	17 R	24.5 R	9 R
Zh→En	N/A	N/A	39.5 R	38 R

Table 13 Validation results for translation

Many of the reference translations did not pass the validation criterion at the first time. Therefore the translations were corrected and revalidated. R means that the results were obtained after correction and re-validation of the translations.

3.3 Schedule

The development phase took place from November 18, 2005 to February 14, 2006.
The SLT test run took place from February 15 to March 1, 2006.
The scoring was done in 2 phases:

- Automatic evaluation from March 1 to March 15 2006
- Human evaluation from March 21 to June 15 2006.

3.4 Participants and Submissions

The total number of participants in this second evaluation campaign was 12: 6 from the TC-STAR consortium and 6 external participants. External participants were:

- Deutsches Forschungszentrum für Künstliche Intelligenz, Germany (DFKI)
- Institute of Computing Technology, China (ICT)
- National Laboratory of Pattern Recognition, China (NLPR)
- National Research Council, Canada (NRC)
- University of Edinburgh, United Kingdom (UED)
- University of Washington, United States (UW)

All participants were allowed to submit for both conditions (Primary and Secondary), and for various versions of their systems. The total number of submissions was 116:

- 38 Submissions for English-to-Spanish
- 45 Submissions for Spanish-to-English
- 33 Submissions for Chinese-to-English.

There have been 4 submissions for the SLT ROVER (English to Spanish both FTE and Verbatim, Spanish to English both FTE and Verbatim).

The submissions received for both condition types are summarized

Site	En→Es			Es->En			Zh->En	
	ASR	FTE	Verbatim	ASR	FTE	Verbatim	ASR	Verbatim
IBM	4P	3P	3P	4P	4P	4P		
ITC-irst	1P + 1S	1P + 1S	1P	1P + 2S	1P + 2S	1P + 2S	1P + 3S	1P + 3S
LIMSI				1P		1P		
RWTH	2P	2P	3P	2P	2P	3P	3P	3P
UKA	1P	1P	1P	1P	1P	1P	2P	4P
UPC	1P	2P + 1S	1P	1P	1P + 1S	1P		
DFKI		1P			1P	1P		
ICT							3P	6P
NLPR							1S	1S
NRC								1P + 1S
UED		1P			1P			
UW		2P	2P		2P	2P		

Table 14 List of submissions in the Second TC-STAR Evaluation Campaign .Condition types: P: Primary; S: Secondary

In order to make a comparison with a real market product, we ran the evaluation for the English-to-Spanish direction with *Systran Professional Premium 5.0*. This software was bought at a regular store and no specific tuning on the software was done for this task. The results obtained with this system are shown in the following section, together with those of the other systems.

3.5 Evaluation Results

The following conditions were applied for evaluation:

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

3.5.1 Human Evaluation

3.5.1.1 Protocol

As planned within the project 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 their reference translations. For comparison purposes, we have also added the translation provided by a *Systran* product.

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 had to answer the question: “Is the text written in good Spanish?” A five-point scale was provided where only extreme marks were defined, ranging from “Perfect Spanish” to “Non understandable Spanish”.

For the evaluation of adequacy, evaluators had 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 was also provided to the evaluators, where, once again, only extreme cases were explicitly defined, going from “All the meaning” to “Nothing in common”.

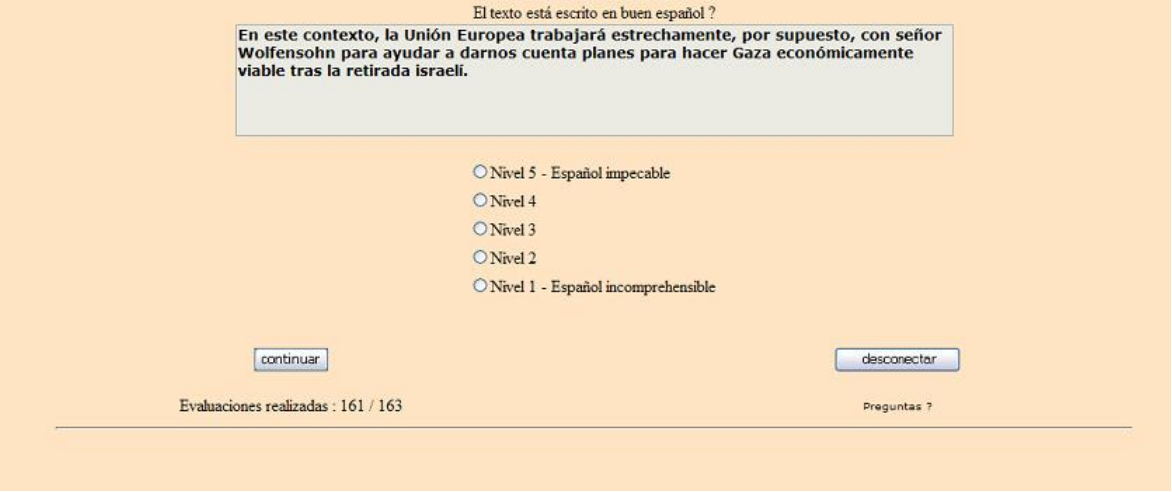
Two evaluations are carried out per segment, they are done by two different evaluators, and segments are distributed to evaluators randomly.

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

The segments are presented randomly, because evaluators should not build a “storyline” and preserve information between two adjoining segments.

3.5.1.2 Evaluation interface

In order to perform the evaluation, we re-use a specific web interface which has already been used for the human evaluation of the French CESTA project [13]. 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 3 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 israelí.

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

Evaluaciones realizadas : 161 / 163

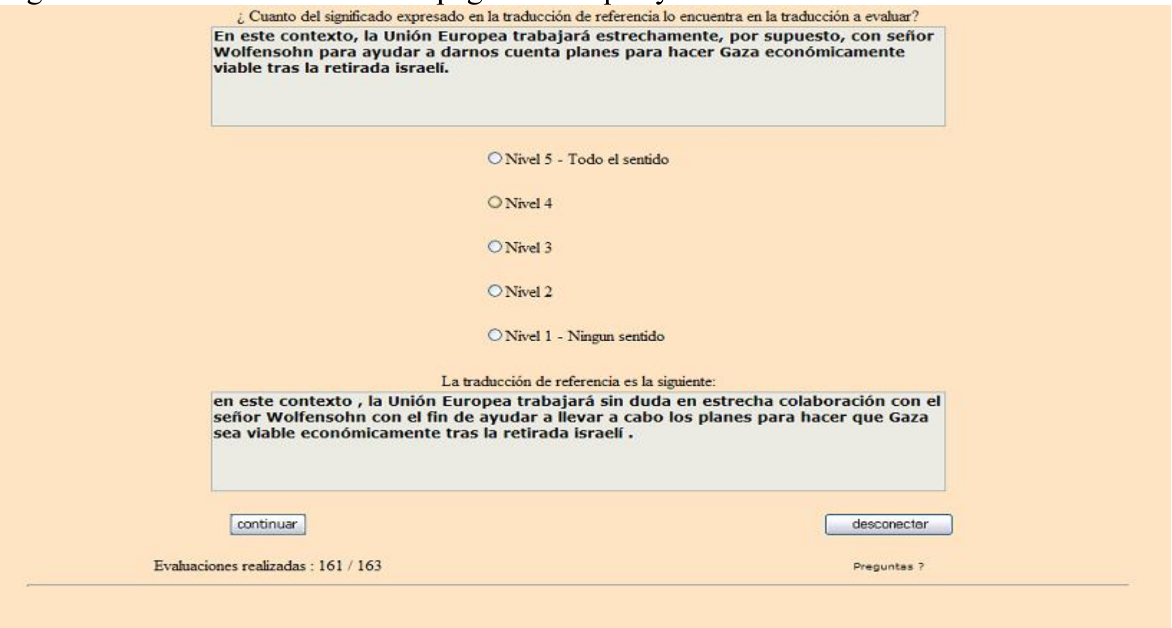
Figure 3: Fluency evaluation.

From top to bottom, 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”),
- 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.

Figure 4 illustrates the evaluation page for adequacy.



¿ 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.

From top to bottom, 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 done and the total of evaluations to do ("Evalautciones realizadas"),
- 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.

3.5.1.3 Set up

Data

Taking into account all the different SLT tasks considered (FTE, Verbatim, ASR), the ROVERs, the Systran product and the human reference translations (for Verbatim/ASR and FTE), there were 6 ASR outputs, 9 Verbatim outputs and 11 FTE outputs to evaluate. A subset of around 400 sentences or segments was randomly extracted for evaluation from each output, which corresponds to a third of the whole output.

Evaluators

The number of evaluators was selected according to three factors:

- the total number of segments to be evaluated,
- the duration of the evaluation,
- the number of evaluation per segment.

Experience shows that judges can undergo no more than about two hours of evaluation without any break, which means around a hundred and half segments. Regarding this constraint two evaluations were done per segment, and both were done by two different judges. Therefore, 125 evaluators were recruited. Finally, a total of 20,360 segments were evaluated, which corresponds to around 156 segments per evaluator.

The 125 evaluators had to be native speakers of Spanish. Table 15 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
125	2	FTE	392	11	8,624	162.88
		Verbatim	388	9	6,984	
		ASR	396	6	4,752	

Table 15: Figures about the human evaluation

3.5.1.4 Evaluators agreement

Each segment within the human evaluation has been evaluated twice, so as to measure consistency in the evaluations carried out and have significant number of judgments. This is done by first computing the ratio between those scores which are identical for two evaluations and the total number of segments. As shown in Table 16, this gives the total agreement between the evaluators.

	FTE + Verb. + ASR	FTE	Verbatim	ASR
Fluency	33.16	34.74	33.85	29.29
Adequacy	32.64	34.23	32.82	29.50

Table 16: Total agreement between the evaluators

The total agreement between the evaluators has proven to be rather good: about a third of the segments obtain identical evaluations with the two evaluators. When isolating the three tasks, the total agreement is different and reveals the difficulties to evaluate according to the task. The FTE agreement is slightly higher than the Verbatim agreement and the ASR agreement is 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.

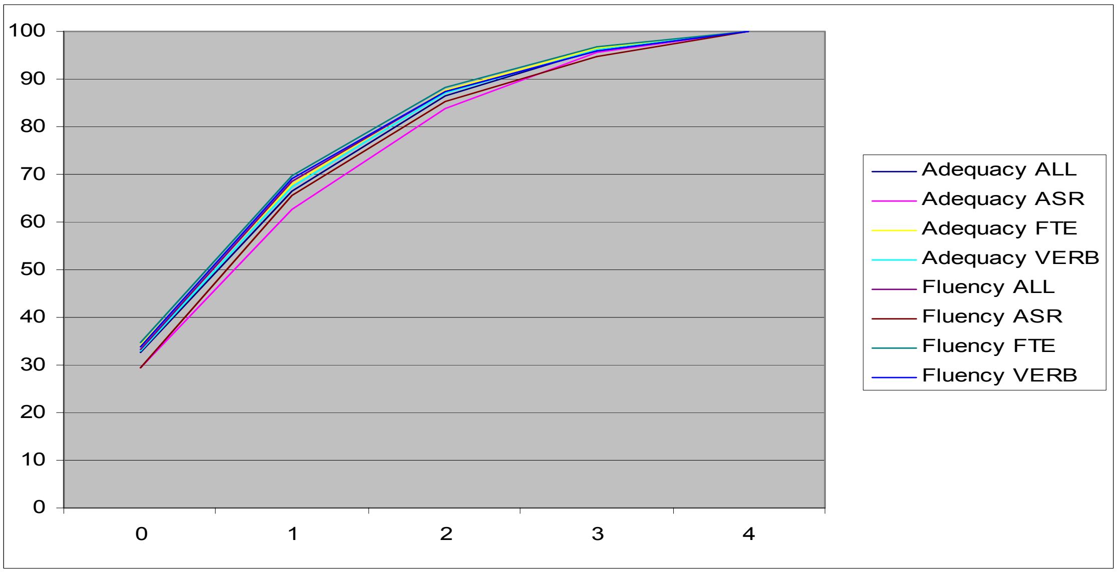


Figure 5: Total agreement between the 1st and 2nd evaluation passes

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 65% have obtained a score that do not differ more than 1 point between the first evaluation pass and the second.

Table 17 shows the mean of the deviance between two evaluations of a same segment (done by two different evaluators) and Table 18 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	1.15	1.11	1.14	1.25
Adequacy	1.18	1.14	1.17	1.28

Table 17: Mean of the deviance

	FTE + Verb. + ASR	FTE	Verbatim	ASR
Fluency	1.10	1.08	1.10	1.13
Adequacy	1.11	1.09	1.10	1.14

Table 18: Standard deviance of the deviances

Table 17 and Table 18 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.

3.5.1.5 Results

The results obtained for the different tasks are detailed below.

3.5.1.5.1 FTE Task

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

SYSTEM	**Fluency** 5 : good 1 : bad	**Adequacy** 5 : good 1 : bad	**Ranking Fluency**	**Ranking Adequacy**
Human Reference	*4.56*	*4.44*	*1*	*1*
UED	3.63	3.79	2	2
RWTH	3.58	3.74	3	3
IBM	3.50	3.60	4	8
UPC	3.48	3.68	5	5
IRST	3.46	3.67	6	6
ROVER	3.46	3.72	6	4
UW	3.40	3.62	8	7
DFKI	3.31	3.53	9	9
UKA	3.17	3.49	10	10
SYSTRAN[4]	*2.46*	*2.93*	*11*	*11*

Table 19: Human scoring and ranking for the FTE task

Table 19 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).

Regarding the general performance of the systems, after the human reference, which obtains the best score by far, the automatic system obtaining the highest score is UED. However, the difference between the human reference and the automatic systems is still considerable. The ROVER system does not obtain the best score, but is not very far from UED. Most of the systems have similar performance, with the exception of UKA, in what regards fluency, and Systran, with regard to both fluency and adequacy.

When considering the performance of systems for fluency and adequacy, all of them obtain higher scores for adequacy than for fluency. However, this is the opposite for the human reference whose fluency scores are higher.

[4] Translation with a commercial cd

The ranking shows some differences between fluency and adequacy: IBM presents a 4-rank difference between fluency and adequacy, while ROVER presents a 2-rank difference.

Then, the mean rank of the systems has been computed, according to each evaluator, for both fluency (Table 20) and adequacy (Table 21):

Fluency	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**	**8th**	**9th**	**10th**	**11th**	**Mean**	**Rank**
Human Reference	*88*	*14*	*1*	*5*	*6*	*2*	*4*	*3*	*1*	*1*	*0*	*2.02*	*1*
UED	13	15	23	15	18	14	6	7	9	4	1	4.61	2
RWTH	4	14	12	14	18	16	12	16	8	5	6	5.68	3
IBM	2	13	15	20	12	12	11	13	12	9	6	5.84	4
IRST	6	15	14	9	12	11	17	20	5	6	10	5.87	5
ROVER	1	15	16	13	13	14	15	9	9	15	5	5.94	6
UW	4	10	13	12	12	12	9	16	17	15	5	6.34	7
UPC	1	9	11	13	14	14	14	15	20	14	0	6.37	8
DFKI	3	8	9	11	9	13	21	12	15	19	5	6.69	9
UKA	2	6	11	9	7	12	8	8	17	27	18	7.45	10
SYSTRAN	*1*	*6*	*0*	*4*	*4*	*5*	*8*	*6*	*12*	*10*	*69*	*9.20*	*11*

Table 20: Mean rank for FTE fluency

Adequacy	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**	**8th**	**9th**	**10th**	**11th**	**Mean**	**Rank**
Human Reference	*68*	*15*	*5*	*6*	*5*	*9*	*6*	*4*	*3*	*1*	*3*	*2.88*	*1*
UED	13	16	12	18	17	21	7	8	4	6	3	4.85	2
RWTH	10	14	11	16	15	7	9	12	12	10	9	5.74	3
ROVER	5	9	19	16	14	9	16	9	13	10	5	5.79	4
IRST	6	12	14	17	8	17	9	13	14	13	2	5.82	5
UPC	5	11	13	11	16	5	20	19	12	5	8	6.06	6
UW	4	14	17	7	9	12	10	14	17	8	13	6.29	7
IBM	2	7	11	12	15	16	16	13	15	9	9	6.46	8
DFKI	8	6	9	10	10	13	13	11	21	20	4	6.62	9
UKA	3	16	7	7	9	11	11	17	7	29	8	6.79	10
SYSTRAN	*1*	*5*	*7*	*5*	*7*	*5*	*8*	*5*	*7*	*14*	*61*	*8.70*	*11*

Table 21: Mean rank for FTE adequacy

Tables show the ranking of the systems for each evaluator, i.e. for a system the number of evaluators who give the 1st position, the 2nd, etc.

As observed in these two tables, the human reference is far above the automatic systems: it obtains 88, the first rank for fluency (on the 125 evaluators), and then 68, the first rank for adequacy. The first automatic system is still UED, as for the scoring. Once again it should be observed that the systems are very close, except for UKA (for Adequacy) and Systran.

3.5.1.5.2 VERBATIM Task

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

SYSTEM	**Fluency** 5 : good 1 : bad	**Adequacy** 5 : good 1 : bad	**Ranking Fluency**	**Ranking Adequacy**

Human Reference	*4.31*	*4.31*	*1*	*1*
UPC	3.39	3.54	2	4
RWTH	3.38	3.55	3	2
IBM	3.35	3.51	4	6
IRST	3.35	3.54	4	4
ROVER	3.32	3.55	6	2
UW	3.14	3.43	7	7
UKA	3.07	3.36	8	8
SYSTRAN	*2.34*	*2.77*	*9*	*9*

Table 22: Human scoring and ranking for the Verbatim task

Results are still better for adequacy than fluency, but the scores for the human reference are the same. UPC is in the first position after the human reference for fluency. For adequacy RWTH and the ROVER are both in the first position after the human reference. But IBM and the ROVER have still strong differences between the ranking of fluency and the ranking of adequacy, as UPC. Except UW and Systran, the systems have quite the same scores, as for fluency there is a difference of 0.07 between the first automatic system and the fifth, and for adequacy a difference of 0.04. Those differences are bigger for the FTE results.

Then we compute the mean rank of the systems, according to each evaluator:

Fluency	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**	**8th**	**9th**	**Mean**	**Rank**
Human Reference	*87*	*18*	*6*	*2*	*2*	*4*	*4*	*2*	*0*	*1.82*	*1*
UPC	12	16	16	30	12	8	15	11	5	4.46	2
IBM	2	16	17	29	19	14	12	9	7	4.79	3
ROVER	7	21	20	11	10	16	17	16	7	4.87	4
IRST	2	15	23	12	20	24	10	15	4	4.95	5
RWTH	2	20	11	14	29	18	18	6	7	4.97	6
UW	6	8	12	11	14	17	26	24	7	5.69	7
UKA	4	9	12	11	15	13	14	30	17	5.97	8
SYSTRAN	*3*	*2*	*8*	*5*	*4*	*11*	*9*	*12*	*71*	*7.48*	*9*

Table 23: Mean rank for Verbatim fluency

Adequacy	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**	**8th**	**9th**	**Mean**	**Rank**
Human Reference	*71*	*20*	*9*	*12*	*3*	*5*	*0*	*5*	*0*	*2.168*	*1*
ROVER	9	23	18	12	19	11	12	15	6	4.608	2
UPC	7	17	17	21	15	15	15	12	6	4.768	3
UW	12	11	21	7	13	18	14	20	9	5.096	4
RWTH	4	11	16	23	18	14	20	10	9	5.128	5
IRST	6	16	11	12	18	24	17	12	9	5.192	6
IBM	8	13	9	18	19	18	15	15	10	5.208	7
UKA	6	10	18	17	9	16	16	17	16	5.448	8
SYSTRAN	*2*	*4*	*6*	*3*	*11*	*4*	*16*	*19*	*60*	*7.384*	*9*

Table 24: Mean rank for Verbatim adequacy

As for the FTE task, the human reference is by far the first, with 87 first ranks for fluency, and 71 for adequacy. The worst system for both fluency and adequacy is Systran. The second system after the human reference is UPC for fluency and the ROVER for adequacy.

3.5.1.5.3 ASR Task

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

SYSTEM	**Fluency** 5 : good 1 : bad	**Adequacy** 5 : good 1 : bad	**Ranking Fluency**	**Ranking Adequacy**
RWTH	3.06	3.13	1	1
IBM	3.04	3.05	2	4
UPC	3.04	3.09	2	2
IRST	2.99	3.09	4	2
UKA	2.84	2.97	5	5
SYSTRAN	*2.09*	*2.33*	*6*	*6*

Table 25: Human scoring and ranking for the ASR task

RWTH gets the first position for the ASR evaluation for both fluency and adequacy. The system is followed by IBM and UPC for the fluency and UPC for the adequacy.
Results are also better for Adequacy, but the scores are closer than for FTE and Verbatim tasks. The differences between the systems are even reduced comparing to FTE results and Verbatim results: for fluency the difference between the first system and the fourth is 0.07, and 0.04 for adequacy.

Then we compute the mean rank of the systems, according to each evaluator:

Fluency	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**	**Mean**	**Rank**
RWTH	37	25	27	20	12	4	2.66	1
IBM	26	30	21	25	15	8	2.98	2
IRST	17	27	26	27	13	15	3.30	3
UPC	21	13	29	26	26	10	3.42	4
UKA	13	22	15	17	47	11	3.77	5
SYSTRAN	*11*	*8*	*7*	*10*	*12*	*77*	*4.88*	*6*

Table 26: Mean rank for ASR fluency

Adequacy	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**	**Mean**	**Rank**
RWTH	31	28	25	18	17	6	2.84	1
UPC	21	28	24	14	27	11	3.25	2
IBM	25	21	20	27	19	13	3.26	3
IRST	23	22	19	29	23	9	3.27	4
UKA	17	20	24	23	28	13	3.51	5
SYSTRAN	*8*	*6*	*13*	*14*	*11*	*73*	*4.86*	*6*

Table 27: Mean rank for ASR adequacy

RWTH is also the best system for the ranking of the mean rank. Except for Systran, all the systems are very close, particularly for the three systems UPC, IBM and IRST.

3.5.1.5.4 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.

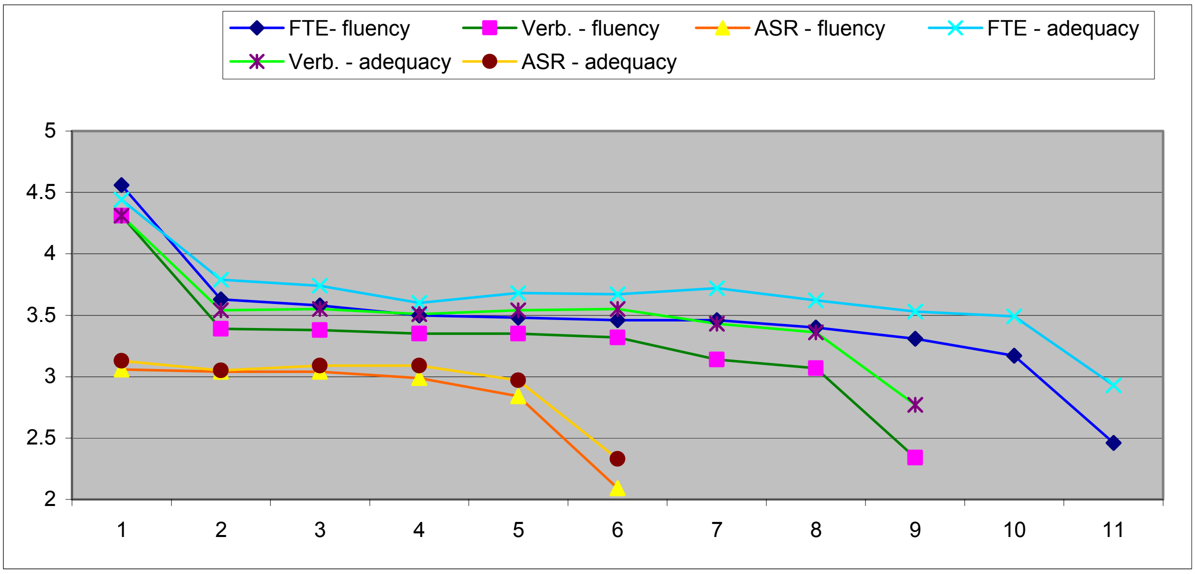


Figure 6: Differences between FTE, Verb. and ASR scores

Finally, Table 28 summarizes all the ranking of the human evaluation:

Task	Site	Score fluency ranking	Mean rank fluency ranking	Score adequacy ranking	Mean rank adequacy ranking
FTE	*Human Reference*	*1*	*1*	*1*	*1*
	UED	2	2	2	2
	RWTH	3	3	3	3
	IBM	4	4	8	8
	UPC	5	8	5	6
	ROVER	6	6	4	4
	IRST	6	5	6	5
	UW	8	7	7	7
	DFKI	9	9	9	9
	UKA	10	10	10	10
	SYSTRAN	*11*	*11*	*11*	*11*
Verbatim	*Human Reference*	*1*	*1*	*1*	*1*
	UPC	2	8	4	3
	RWTH	3	6	2	5
	IBM	4	3	6	7
	IRST	4	5	4	6
	ROVER	6	4	2	2

	UW	7	2	7	4
	UKA	8	9	8	8
	SYSTRAN	*9*	*7*	*9*	*9*
ASR	RWTH	1	1	1	1
	IBM	2	2	4	3
	UPC	2	4	2	2
	IRST	4	3	2	4
	UKA	5	5	5	5
	SYSTRAN	*6*	*6*	*6*	*6*

Table 28: Ranks summary

This table shows the differences between the ranking presented previously, comparing first adequacy and fluency, but also the score ranking and the mean rank ranking. Most of the systems obtain the same rank every time, especially at the first systems and the last systems. The differences appear between fluency and adequacy, while the two types of scores produce a similar ranking. So, a good fluency does not necessarily mean a good adequacy. See for example IBM, ROVER, and also UPC to a lesser extent. Indeed IBM has worse results for adequacy than fluency and loses four place in the FTE ranking and two places in the Verbatim and ASR ranking. The ROVER is conversely worse with fluency than adequacy and loses two places in the FTE ranking and four places in the Verbatim ranking.

3.5.2 Automatic evaluations

We used five different automatic metrics for the evaluation of the translation output:

3.5.2.1 Metrics

- **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.
- **BLEU/NIST, referred to as NIST,** is a variant metric of BLEU, which applies different weight for the n-grams, functions of information gain and length penalty.
- **BLEU/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 [9]). 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 BLEU/NIST. For BLEU/IBM, BLEU, 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.

3.5.2.2 Automatic results for English-to-Spanish

The statistics for the source documents are the following:

- Verbatim: 28 882 words for 1 155 sentences
- Text: 25 876 words for 1 117 sentences.
- ASR: 29 531 words.

As it can be seen, there is a higher number of words in the manual transcription (28 882) than in the final text edition (25 876). This is due to the hesitations, repetitions, etc. that can be found in the transcriptions. The number of words in the automatic transcription is also slightly higher (29 531) than the manual one (28 882).

Table 29 shows some statistics in terms of number of words for the submitted translations of the primary systems and for one reference translation. *Ref-1-ver* and *Ref-1-txt* are the first references for respectively ASR and Verbatim tasks, and FTE task.

Input	**Site**	**number of words**	**words per sentence**	**words src / words trans**
ASR	IBM	31 356	27.15	0.94
	ITC-irst	30 352	26.28	0.97
	RWTH	30 643	26.53	0.96
	UKA	29 368	25.43	1.01
	UPC	29 876	25.87	0.99
	Ref-1-ver	*31 243*	*27.05*	*0.95*
Verbatim	IBM	33 134	28.69	0.87
	ITC-irst	29 022	25.13	1.00
	RWTH	29 284	25.35	0.99
	UKA	27 658	23.95	1.04
	UPC	28 661	24.81	1.01
	UW	29 170	25.26	0.99
	ROVER	*28 802*	*24.94*	*1.00*
	Ref-1-ver	*29 114*	*25.21*	*0.99*
Text	IBM	31 556	28.25	0.82
	ITC-irst	27 419	24.55	0.94
	RWTH	26 945	24.12	0.96
	UKA	26 022	23.30	0.99
	UPC	27 568	24.68	0.94

	DFKI	27 312	24.45	0.95
	UED	26 892	24.08	0.96
	UW	27 539	24.65	0.94
	ROVER	*26 285*	*23.53*	*0.98*
	Ref-1-txt	*27 032*	*24.20*	*0.96*

Table 29: LRs statistics for English-to-Spanish EPPS task

The ratio between the source text in English and the reference translation in Spanish is 0.96, which outlines a strong correlation between the length of the source sentence and its corresponding translation. Then, IBM system which strongly moves away from this point of balance is clearly penalized by automatic metrics. The same occurs with the verbatim output, as the primary IBM output is 0.87 rather than 0.99 for the reference. All the other outputs from all the tasks are close to the reference file, and then are not penalized too much.

Table 30 presents the scoring results for English-to-Spanish EPPS. The best primary systems are in bold.

Task	**Site**	**BLEU/ NIST**	**BLEU**	**BLEU/ IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**
FTE	DFKI Primary	8.70	36.32	36.33	48.06	36.36	42.66	-
	IBM Primary	9.89	47.54	47.56	41.25	31.47	48.29	-
	ITC-irst Primary	10.23	49.81	49.00	39.31	30.21	48.54	-
	ITC-irst Secondary	10.23	49.79	49.12	39.17	30.10	48.32	-
	ROVER	*10.38*	*50.74*	*49.96*	*38.15*	*29.26*	49.50	-
	RWTH Primary	10.16	49.44	49.45	39.81	30.48	48.77	-
	UED Primary	10.11	49.50	49.42	39.69	30.51	48.37	-
	UKA Primary	9.56	44.04	42.95	43.61	33.66	45.95	-
	UPC Primary	10.00	48.20	47.69	40.89	31.49	46.89	-
	UPC Secondary	10.06	48.85	48.32	40.21	31.46	47.32	-
	UW Primary	10.01	48.50	48.05	40.37	30.95	47.98	-
	SYSTRAN	*8.57*	*36.29*	*36.31*	*47.79*	*37.36*	42.10	-
Verbatim	IBM Primary	9.61	45.12	45.12	43.56	32.60	46.30	-
	ITC-irst Primary	9.91	46.61	46.33	42.19	31.51	46.34	-
	ITC-irst Secondary	9.55	44.85	44.51	44.45	33.85	46.35	-
	ROVER	*10.06*	*47.53*	*46.99*	*40.92*	*30.39*	46.84	-
	RWTH Primary	9.71	45.42	45.42	43.12	32.09	46.21	-
	UKA Primary	9.08	40.10	39.59	47.63	36.13	44.61	-
	UPC Primary	9.50	44.06	43.47	44.66	33.68	44.97	-
	UW Primary	9.24	42.57	42.52	46.15	34.84	45.38	-
	SYSTRAN	*8.10*	*32.97*	*32.97*	*51.86*	*39.74*	39.38	-
ASR	IBM Primary	8.62	35.77	35.67	52.03	38.79	43.70	51.06
	ITC-irst	8.75	35.97	35.09	50.95	39.31	44.08	50.02

	Primary							
	ITC-irst Secondary	8.48	34.54	33.69	52.60	41.05	43.79	50.14
	RWTH Primary	8.72	35.91	35.02	50.52	38.66	43.44	50.05
	UKA Primary	8.10	31.32	30.58	55.48	43.15	41.93	56.38
	UPC Primary	8.56	34.76	34.02	51.79	40.01	43.23	50.87
	SYSTRAN	*7.03*	*23.93*	*23.86*	*62.15*	*47.84*	36.82	*61.80*

Table 30: Evaluation results for the English-to-Spanish EPPS task

Table 31 presents the ranking results for English-to-Spanish EPPS.

Task	**Site**	**BLEU/ NIST**	**BLEU**	**BLEU/ IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**
FTE	DFKI Primary	11	12	12	12	11	12	-
	IBM Primary	9	9	9	9	8	6	-
	ITC-irst Primary	2	2	5	3	3	3	-
	ITC-irst Secondary	3	3	4	2	2	5	-
	ROVER	*1*	*1*	*1*	*1*	*1*	*1*	-
	RWTH Primary	4	5	2	5	4	2	-
	UED Primary	5	4	3	4	5	4	-
	UKA Primary	10	10	10	10	10	10	-
	UPC Primary	8	8	8	8	9	9	-
	UPC Secondary	6	6	6	6	7	8	-
	UW Primary	7	7	7	7	6	7	-
	SYSTRAN	*12*	*11*	*11*	*11*	*12*	*11*	-
Verbatim	IBM Primary	4	4	4	4	4	4	-
	ITC-irst Primary	2	2	2	2	2	3	-
	ITC-irst Secondary	5	5	5	5	6	2	-
	ROVER	*1*	*1*	*1*	*1*	*1*	*1*	-
	RWTH Primary	3	3	3	3	3	5	-
	UKA Primary	8	8	8	8	8	8	-
	UPC Primary	6	6	6	6	5	7	-
	UW Primary	7	7	7	7	7	6	-
	SYSTRAN	*9*	*9*	*9*	*9*	*9*	*9*	-
ASR	IBM Primary	3	3	1	4	2	3	3
	ITC-irst Primary	1	1	2	2	3	1	1
	ITC-irst Secondary	5	5	5	5	5	2	4
	RWTH	2	2	3	1	1	4	2

	Primary							
	UKA Primary	6	6	6	6	6	6	6
	UPC Primary	4	4	4	3	4	5	5
	SYSTRAN	*7*	*7*	*7*	*7*	*7*	*7*	*7*

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

If we exclude the SLT ROVER system, which is most of the time in the first position, RWTH and ITC-irst share the best results for Verbatim and ASR tasks, while ITC-irst seems to have the best system for the FTE task.
The ROVER achieves the best results for both Verbatim and FTE inputs.

In general BLEU and BLEU/IBM return identical or very close scores. We can also observe that results for mPER are approximately 10% better than for mWER.
Results for Verbatim inputs are substantially better than FTE ones, which are likewise better than those from ASR input.

3.5.2.3 Automatic results for Spanish-to-English

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

- Text: 50 590 words, for 1782 sentences whereof
    - CORTES: 25 084 words, for 888 sentences
    - EPPS: 25 510 words, for 894 sentences
- Verbatim: 56 239 words, for 1 596 sentences whereof
    - CORTES: 28 370 words, for 699 sentences
    - EPPS: 27 873 words, for 897 sentences
- ASR: 54 708 words whereof
    - CORTES: 26 769 words.
    - EPPS: 28 939 words.

There are fewer words in the manual transcriptions (56 243 words for Verbatim CORTES and EPPS) than in the automatic ones (54 708 words for ASR CORTES and EPPS).

As with English-to-Spanish, we have computed some statistics about the average number of words per sentence that are shown in Table 32, for the whole CORTES and EPPS data.

Input	**Site**	**number of words**	**words per sentence**	**words src / words trans**
ASR	IBM	62 940	39.44	0.87
	ITC-irst	61 497	38.53	0.89
	LIMSI	57 647	36.12	0.95
	RWTH	60 775	38.08	0.90
	UKA	58 840	36.87	0.93
	UPC	62 222	38.99	0.88
	Ref-1-ver	*61 207*	*38.35*	*0.89*
Verbatim	IBM	62 407	39.10	0.90
	ITC-irst	56 584	35.45	0.99
	LIMSI	55 974	35.07	1.00
	RWTH	56 168	35.19	1.00

	UKA	54 921	34.41	1.02
	UPC	57 107	35.78	0.98
	DFKI	56 802	35.59	0.99
	UW	58 065	36.38	0.97
	ROVER	*56 510*	*35.41*	*1.00*
	Ref-1-ver	*59 583*	*37.33*	*0.94*
Text	IBM	58 964	33.09	0.86
	ITC-irst	52 856	29.66	0.96
	RWTH	52 407	29.41	0.97
	UKA	50 835	28.53	1.00
	UPC	53 423	29.98	0.95
	DFKI	53 139	29.82	0.95
	UED	51 940	29.15	0.97
	UW	54 121	30.37	0.93
	ROVER	*51 486*	*28.89*	*0.98*
	Ref-1-txt	*52 051*	*29.21*	*0.97*

Table 32: LRs statistics for the Spanish-to-English task

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 33 shows the scoring results for Spanish-to-English for the whole (EPPS+CORTES) corpus:

Task	**Site**	**BLEU/ NIST**	**BLEU**	**BLEU/ IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**
FTE	IBM Primary	10.49	48.16	48.16	41.68	30.18	44.81	-
	ITC-irst Primary	10.22	46.19	46.11	43.36	31.34	43.36	-
	ITC-irst Secondary	10.14	45.58	45.39	43.66	31.66	42.98	-
	ROVER	*10.50*	*48.07*	*48.07*	*41.64*	*30.03*	45.21	-
	RWTH Primary	10.36	47.11	47.12	42.89	30.93	44.55	-
	UED Primary	10.11	45.59	45.60	43.74	31.67	43.61	-
	UKA Primary	9.63	41.23	40.98	47.17	33.64	42.31	-
	UPC Primary	10.30	46.45	46.46	42.55	30.97	44.48	-
	DFKI Primary	9.06	37.24	37.24	63.15	34.95	39.95	-
	UW Primary	10.09	45.63	45.63	44.06	31.74	44.66	-
	SYSTRAN	*9.45*	*40.57*	*40.57*	*47.27*	*34.58*	38.39	-
Verbatim	IBM Primary	11.04	52.54	52.41	37.46	26.98	50.03	-
	ITC-irst Primary	10.57	48.85	48.49	39.94	28.66	46.06	-
	ITC-irst Secondary	10.27	47.16	46.63	41.89	30.71	45.60	-
	LIMSI Primary	9.72	42.59	42.08	44.50	31.76	41.80	-

	ROVER	11.09	52.55	52.08	36.82	26.78	50.55	-
	RWTH Primary	11.10	52.45	51.91	37.73	27.41	49.34	-
	UKA Primary	9.89	43.18	42.84	44.85	31.54	44.76	-
	UPC Primary	10.65	49.63	49.57	39.60	28.85	45.72	-
	DFKI Primary	9.09	37.50	37.42	68.61	33.53	40.34	-
	UW Primary	9.90	44.97	44.97	44.37	32.28	45.22	-
	SYSTRAN	9.89	43.73	43.74	43.54	31.19	39.66	-
ASR	IBM Primary	9.57	39.41	38.37	48.73	34.72	45.11	47.43
	ITC-irst Primary	9.03	34.30	33.92	50.50	38.58	42.96	49.04
	ITC-irst Secondary	8.64	32.52	32.24	52.52	40.82	42.46	49.43
	LIMSI Primary	8.48	32.60	32.13	52.58	38.18	38.28	51.72
	RWTH Primary	9.26	36.13	35.79	49.00	38.02	43.96	47.66
	UKA Primary	8.42	30.13	29.79	54.82	41.42	40.96	54.63
	UPC Primary	9.04	34.83	34.30	51.01	39.01	40.87	49.57
	SYSTRAN	8.06	29.22	29.22	62.27	47.77	36.60	*62.14*

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

Table 34 shows the ranking of systems for Spanish-to-English for the whole (EPPS+CORTES) corpus:

Task	**Site**	**BLEU/ NIST**	**BLEU**	**BLEU/ IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**
FTE	IBM Primary	2	1	1	2	2	2	-
	ITC-irst Primary	5	5	5	5	5	7	-
	ITC-irst Secondary	6	8	8	6	6	8	-
	ROVER	*1*	*2*	*2*	*1*	*1*	1	-
	RWTH Primary	3	3	3	4	3	4	-
	UED Primary	7	7	7	7	7	6	-
	UKA Primary	9	9	9	9	9	9	-
	UPC Primary	4	4	4	3	4	5	-
	DFKI Primary	11	11	11	11	11	10	-
	UW Primary	8	6	6	8	8	3	-
	SYSTRAN	*10*	*10*	*10*	*10*	*10*	11	-
Verbatim	IBM Primary	3	2	1	2	2	2	-
	ITC-irst Primary	5	5	5	5	4	4	-
	ITC-irst Secondary	6	6	6	6	6	6	-
	LIMSI Primary	10	10	10	9	9	9	-
	ROVER	2	1	2	1	1	1	-
	RWTH Primary	1	3	3	3	3	3	-
	UKA Primary	8	9	9	10	8	8	-
	UPC Primary	4	4	4	4	5	5	-
	DFKI Primary	11	11	11	11	11	10	-

	UW Primary	7	7	7	8	10	7	-
	SYSTRAN	9	8	8	7	7	11	-
ASR	IBM Primary	1	1	1	1	1	1	1
	ITC-irst Primary	4	4	4	3	4	3	4
	ITC-irst Secondary	5	6	5	5	6	4	5
	LIMSI Primary	6	5	6	6	3	7	6
	RWTH Primary	2	2	2	2	2	2	2
	UKA Primary	7	7	7	7	7	5	7
	UPC Primary	3	3	3	4	5	6	3
	SYSTRAN	*8*	*8*	*8*	*8*	*8*	8	*8*

Table 34: Ranking of systems for the Spanish-to-English task

For the ASR and Verbatim tasks, IBM and RWTH obtain the best results, while IBM is the best system for the FTE task.

Here the results for Verbatim inputs are better than the FTE ones, which are better than those from ASR inputs. As for the English to Spanish direction, the BLEU and BLEU/IBM scores are closer, and mPER is approximately 15% higher to mWER.

Table 35 shows the scoring results for Spanish-to-English EPPS.

Task	**Site**	**BLEU/ NIST**	**BLEU**	**BLEU/ IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**
FTE	IBM Primary	10.77	54.06	53.66	36.21	26.37	51.22	-
	ITC-irst Primary	10.56	52.40	52.13	37.43	27.21	50.56	-
	ITC-irst Secondary	10.48	51.81	51.44	37.73	27.62	50.24	-
	ROVER	*10.79*	*53.99*	*53.67*	*35.99*	*26.13*	51.42	-
	RWTH Primary	10.65	53.10	53.12	37.06	26.91	51.43	-
	UED Primary	10.48	51.87	51.88	37.65	27.47	51.38	-
	UKA Primary	9.98	47.05	46.57	41.52	29.70	49.23	-
	UPC Primary	10.60	52.30	52.20	36.97	27.12	51.48	-
	DFKI Primary	9.47	43.04	43.03	56.35	30.94	47.03	-
	UW Primary	10.53	52.61	52.61	37.57	27.18	52.95	-
	SYSTRAN	*9.72*	*45.72*	*45.73*	*42.08*	*30.76*	45.95	-
Verbatim	IBM Primary	10.89	55.08	54.60	36.35	25.84	53.39	-
	ITC-irst Primary	10.55	52.08	51.60	38.18	26.84	49.55	-
	ITC-irst Secondary	10.23	50.23	49.58	40.26	29.17	49.16	-
	LIMSI Primary	9.76	45.99	45.47	42.45	30.06	46.66	-
	ROVER	*10.99*	*55.55*	*54.89*	*35.61*	*25.31*	53.70	-
	RWTH Primary	10.94	55.06	54.53	36.45	25.85	52.35	-
	UKA Primary	9.85	46.00	45.48	43.20	30.33	48.15	-
	UPC Primary	10.45	52.00	52.00	38.84	27.95	50.01	-
	DFKI Primary	9.33	42.20	42.02	61.45	31.71	45.20	-

	UW Primary	9.85	47.86	47.86	42.73	30.97	49.15	-
	SYSTRAN	*9.68*	*45.28*	*45.28*	*43.12*	*30.57*	43.58	-
ASR	IBM Primary	9.63	42.65	41.61	45.98	32.65	49.76	45.37
	ITC-irst Primary	9.21	37.93	37.43	47.59	35.82	46.24	46.85
	ITC-irst Secondary	8.80	35.83	35.42	49.83	38.34	45.71	47.44
	LIMSI Primary	8.71	36.60	36.11	49.37	36.12	43.61	48.99
	RWTH Primary	9.38	39.44	39.01	46.53	35.57	47.88	46.03
	UKA Primary	8.53	33.02	32.75	52.21	39.27	44.69	53.47
	UPC Primary	9.15	38.33	37.67	48.75	36.58	45.74	48.04
	SYSTRAN	*8.58*	*33.79*	*32.95*	*53.16*	*38.80*	41.92	*54.60*

Table 35: Evaluation results for the Spanish-to-English EPPS task

Table 36 shows the ranking of systems for Spanish-to-English EPPS.

Task	**Site**	**BLEU/ NIST**	**BLEU**	**BLEU/ IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**
FTE	IBM Primary	2	1	2	2	2	51.22	-
	ITC-irst Primary	5	5	6	5	6	50.56	-
	ITC-irst Secondary	8	8	8	8	8	50.24	-
	ROVER	*1*	*2*	*1*	*1*	*1*	*51.42*	-
	RWTH Primary	4	3	3	4	3	51.43	-
	UED Primary	7	7	7	7	7	51.38	-
	UKA Primary	9	9	9	9	9	49.23	-
	UPC Primary	3	6	5	3	4	51.48	-
	DFKI Primary	10	11	11	11	10	47.03	-
	UW Primary	6	4	4	6	5	52.95	-
	SYSTRAN	*11*	*10*	*10*	*10*	*11*	*45.95*	-
Verbatim	IBM Primary	3	2	2	2	2	2	-
	ITC-irst Primary	4	4	5	4	4	5	-
	ITC-irst Secondary	6	6	6	6	6	6	-
	LIMSI Primary	9	9	9	7	7	9	-
	ROVER	*1*	*1*	*1*	*1*	*1*	*1*	-
	RWTH Primary	2	3	3	3	3	3	-
	UKA Primary	7	8	8	10	8	8	-
	UPC Primary	5	5	4	5	5	4	-
	DFKI Primary	11	11	11	11	11	10	-
	UW Primary	8	7	7	8	10	7	-
	SYSTRAN	*10*	*10*	*10*	*9*	*9*	*11*	-
ASR	IBM Primary	1	1	1	1	1	1	1
	ITC-irst Primary	3	4	4	3	3	3	3
	ITC-irst Secondary	5	6	6	6	6	5	4

	LIMSI Primary	6	5	5	5	4	7	6
	RWTH Primary	2	2	2	2	2	2	2
	UKA Primary	7	7	7	7	7	6	7
	UPC Primary	4	3	3	4	5	4	5
	SYSTRAN	*8*	*8*	*8*	*8*	*8*	*8*	*8*

Table 36: Ranking of systems for the Spanish-to-English EPPS task

Table 37 shows the scoring results for Spanish-to-English CORTES.

Task	Site	BLEU/ NIST	BLEU	BLEU/ IBM	mWER	mPER	WNM	AS-WER
FTE	IBM Primary	9.26	42.08	42.08	47.30	34.07	41.88	-
	ITC-irst Primary	8.96	39.66	39.66	49.43	35.58	39.86	-
	ITC-irst Secondary	8.90	39.02	38.97	49.74	35.81	39.42	-
	ROVER	*9.27*	*41.95*	*41.95*	*47.44*	*34.02*	*42.71*	-
	RWTH Primary	9.13	40.92	40.93	48.87	35.05	41.53	-
	UED Primary	8.85	39.04	39.04	49.98	35.97	39.97	-
	UKA Primary	8.45	35.17	35.10	52.96	37.67	38.78	-
	UPC Primary	9.06	40.37	40.38	48.26	34.91	41.14	-
	DFKI Primary	7.91	31.10	31.11	70.13	39.05	36.39	-
	UW Primary	8.76	38.30	38.30	50.71	36.4	39.78	-
	SYSTRAN	*8.32*	*35.02*	*35.01*	*52.74*	*38.67*	34.77	-
Verbatim	IBM Primary	10.20	50.14	50.14	38.54	28.11	50.16	-
	ITC-irst Primary	9.68	45.70	45.45	41.66	30.45	46.75	-
	ITC-irst Secondary	9.43	44.17	43.76	43.48	32.22	46.16	-
	LIMSI Primary	8.88	39.22	38.72	46.51	33.44	40.36	-
	ROVER	*10.20*	*49.65*	*49.35*	*38.01*	*28.23*	*50.45*	-
	RWTH Primary	10.25	49.88	49.33	38.98	28.94	49.19	-
	UKA Primary	9.11	40.45	40.27	46.47	32.73	44.97	-
	UPC Primary	9.91	47.28	47.06	40.34	29.73	45.71	-
	DFKI Primary	8.18	32.82	32.83	75.63	35.31	39.68	-
	UW Primary	9.11	42.13	42.13	45.98	33.57	45.44	-
	SYSTRAN	*9.26*	*42.20*	*42.22*	*44.01*	*31.81*	40.81	-
ASR	IBM Primary	8.71	36.06	35.03	51.55	36.84	44.69	49.50
	ITC-irst Primary	8.08	30.53	30.26	53.48	41.40	43.89	51.25
	ITC-irst Secondary	7.74	29.09	28.91	55.28	43.36	43.37	51.42
	LIMSI Primary	7.58	28.39	27.96	55.86	40.29	36.44	54.47
	RWTH Primary	8.34	32.70	32.46	51.73	40.76	44.01	49.48
	UKA Primary	7.63	27.12	26.72	57.50	43.63	41.25	55.80
	UPC Primary	8.14	31.19	30.77	53.33	41.49	40.88	51.10
	SYSTRAN	*7.86*	*28.48*	*27.72*	*57.46*	*42.83*	37.79	*56.31*

Table 37: Evaluation results for the Spanish-to-English CORTES task

Table 38 shows the ranking of systems for Spanish-to-English CORTES.

Task	Site	BLEU/ NIST	BLEU	BLEU/ IBM	mWER	mPER	WNM	AS-WER
FTE	IBM Primary	2	1	1	1	2	2	-
	ITC-irst Primary	5	5	5	5	5	6	-
	ITC-irst Secondary	6	7	7	6	6	8	-
	ROVER	*1*	*2*	*2*	*2*	*1*	*1*	-
	RWTH Primary	3	3	3	4	4	3	-
	UED Primary	7	6	6	7	7	5	-
	UKA Primary	9	9	9	10	9	9	-
	UPC Primary	4	4	4	3	3	4	-
	DFKI Primary	11	11	11	11	11	10	-
	UW Primary	8	8	8	8	8	7	-
	SYSTRAN	*10*	*10*	*10*	*9*	*10*	*11*	-
Verbatim	IBM Primary	3	1	1	2	1	2	-
	ITC-irst Primary	5	5	5	5	5	4	-
	ITC-irst Secondary	6	6	6	6	7	5	-
	LIMSI Primary	10	10	10	10	9	10	-
	ROVER	*2*	*3*	*2*	*1*	*2*	*1*	-
	RWTH Primary	1	2	3	3	3	3	-
	UKA Primary	8	9	9	9	8	8	-
	UPC Primary	4	4	7	4	4	6	-
	DFKI Primary	11	11	11	11	11	11	-
	UW Primary	9	8	8	8	10	7	-
	SYSTRAN	*7*	*7*	*7*	*7*	*6*	*9*	-
ASR	IBM Primary	1	1	1	1	1	1	2
	ITC-irst Primary	4	4	4	4	4	3	4
	ITC-irst Secondary	6	5	5	5	7	4	5
	LIMSI Primary	8	7	6	6	2	8	6
	RWTH Primary	2	2	2	2	3	2	1
	UKA Primary	7	8	8	8	8	5	7
	UPC Primary	3	3	3	3	5	6	3
	SYSTRAN	*5*	*6*	*7*	*7*	*6*	7	*8*

Table 38: Ranking of systems for the Spanish-to-English CORTES task

The results from EPPS inputs are better than those from CORTES inputs, and this is the case for all the systems. The ranking does not vary with very few exceptions. For the two tasks IBM obtains the best results, higher than the ROVER system (except for the Verbatim and the FTE EPPS tasks). The ROVER system is close to IBM with the Verbatim and FTE CORTES tasks. In most conditions RWTH is in third position. Then UPC, ITC and UED are often in the same scale of results in the next position. Finally, UKA, UW, DFKI and Systran are in a last group of systems whose the results are lower than for the other systems, especially for DFKI and Systran which obtain very low scores.

3.5.2.4 Automatic results for Chinese-to-English

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

- Verbatim: 27 370 words, for 1 232 sentences.

Some statistics about the average number of words per sentence are shown in Table 39.

Input	Site	number of words	words per sentence	words src / words trans
ASR	ITC-irst	30 584	24.82	0.89
	RWTH	30 198	24.51	0.91
	UKA	31 815	25.82	0.86
	ICT	29 618	24.04	0.92
	NLPR	32 216	26.15	0.85
	Ref-1-ver	31 184	25.31	0.88
Verbatim	ITC-irst	28 648	23.25	0.96
	RWTH	28 541	23.17	0.96
	UKA	27 996	22.72	0.98
	ICT	27 666	22.46	0.99
	NLPR	32 283	26.20	0.85
	NRC	29 971	24.33	0.91
	Ref-1-ver	30 707	24.92	0.89

Table 39: LRs statistics for the Chinese-to-English VOA task

Table 40 presents the scoring results for Chinese-to-English VOA.

Task	**Site**	**BLEU/ NIST**	**BLEU**	**BLEU/ IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**
Verbatim	ICT Primary	6.03	13.70	13.20	78.68	58.06	25.72	-
	ITC-irst Primary	6.01	14.04	13.49	79.76	59.42	25.73	-
	ITC-irst Secondary	6.00	13.92	13.42	80.05	59.66	25.70	-
	NLPR Secondary	4.35	7.30	7.30	102.92	79.82	22.25	-
	NRC Primary	5.49	12.24	12.25	84.50	63.07	26.43	-
	NRC Secondary	5.80	12.76	12.76	84.23	61.67	27.36	-
	RWTH Primary	6.45	16.07	15.32	78.08	56.34	27.58	-
	UKA Primary	5.51	10.81	10.30	82.22	61.72	24.21	-
	SYSTRAN	*4.28*	*6.53*	*6.53*	*95.37*	*74.77*	23.35	-
ASR	ICT Primary	4.90	10.86	10.46	77.79	62.46	24.56	83.31
	ITC-irst Primary	4.92	11.07	10.83	78.80	63.19	24.71	83.58
	ITC-irst Secondary	4.95	11.12	10.88	78.93	63.22	24.69	83.78
	NLPR Secondary	4.09	6.74	6.74	87.28	71.27	20.85	90.49
	RWTH Primary	5.17	12.39	12.09	77.99	61.98	26.47	83.02
	UKA Primary	4.59	8.46	8.46	82.92	66.89	23.86	88.28

	SYSTRAN	4.38	8.62	8.48	80.59	65.78	22.52	95.20

Table 40: Evaluation results for the Chinese-to-English EPPS task

Table 41 presents the ranking of systems for Chinese-to-English VOA.

Task	**Site**	**BLEU/ NIST**	**BLEU**	**BLEU/ IBM**	**mWER**	**mPER**	**WNM**	**AS-WER**
Verbatim	ICT Primary	2	4	4	2	2	5	-
	ITC-irst Primary	3	2	2	3	3	4	-
	ITC-irst Secondary	4	3	3	4	4	6	-
	NLPR Secondary	9	9	9	9	9	9	-
	NRC Primary	7	6	6	8	8	3	-
	NRC Secondary	5	5	5	7	5	2	-
	RWTH Primary	1	1	1	1	1	1	-
	UKA Primary	6	8	8	6	6	7	-
	SYSTRAN	*8*	*7*	*7*	*5*	*7*	8	-
ASR	ICT Primary	4	4	4	1	2	4	2
	ITC-irst Primary	3	3	3	3	3	2	3
	ITC-irst Secondary	2	2	2	4	4	3	4
	NLPR Secondary	7	7	7	7	7	7	6
	RWTH Primary	1	1	1	2	1	1	1
	UKA Primary	5	5	5	6	6	5	5
	SYSTRAN	*6*	*6*	*6*	*5*	*5*	6	7

Table 41: Ranking of systems for the Chinese-to-English EPPS task

Whatever the task is, RWTH is almost always in first position, except for the mWER and mPER metrics in the ASR condition.

3.6 Data analysis

3.6.1 Statistical analysis of the evaluation metrics

In Table 42 we present the metrics correlations. The used metrics to compute the Pearson correlation scores are BLEU, BLEU/IBM, WNM and mPER (as we see in the first evaluation [8] that mWER and mPER metrics are strongly correlate).

Metric	**En->Es**			**Es->En**			**Zh->En**	
	ASR	**Text**	**Verb**	**ASR**	**Text**	**Verb**	**ASR**	**Verb**
BLEU ↔ IBM	99.75	99.74	99.85	99.86	99.97	99.90	99.91	99.74
BLEU ↔ mPER	98.74	98.76	97.87	86.90	98.23	96.11	95.94	93.96
BLEU ↔ WNM	98.58	97.79	96.61	80.68	88.49	90.30	94.95	87.46
IBM ↔ mPER	99.16	98.65	97.91	85.54	98.24	95.89	95.34	93.21

IBM ↔ WNM	97.85	97.93	96.52	81.06	88.32	89.72	95.68	89.90
mPER ↔ WNM	94.79	98.64	91.35	76.03	92.68	86.43	91.12	84.25

Table 42: Pearson correlation between metrics scoring

In Table 43 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 ↔ IBM	3	4	0	2	0	2	0	0
BLEU ↔ mPER	3	10	2	2	4	6	4	5
BLEU ↔ WNM	3	6	5	5	9	6	2	7
IBM ↔ mPER	3	10	2	3	4	8	4	5
IBM ↔ WNM	5	6	5	5	9	8	2	7
mPER ↔ WNM	5	8	5	5	8	3	3	7

Table 43: Number of systems with a different rank when comparing two metrics

All the metrics are strongly correlated. Most of the results obtained with BLEU and BLEU/IBM are almost the same for all the inputs. mPER is also correlated with the BLEU metrics.

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 was done on this direction only. For that we compute the correlations between the automatic metrics' scores and fluency/adequacy scores, and we also compute the differences in ranks between the automatic metrics' ranks and fluency/adequacy ranks.

Metrics	ASR scoring	Text scoring	Verb scoring	ASR ranking	Text ranking	Verb ranking
BLEU vs. Fluency	97.91	77.16	95.15	4	10	5
IBM vs. Fluency	97.26	78.35	94.70	4	9	5
mPER vs. Fluency	96.04	80.85	90.25	2	9	5
WNM vs. Fluency	98.66	78.82	95.69	3	9	5
BLEU vs. Adequacy	97.09	80.00	95.53	4	9	4
IBM vs. Adequacy	95.80	80.38	94.95	3	8	4
mPER vs. Adequacy	94.47	83.73	90.31	3	9	4
WNM vs. Adequacy	98.45	80.67	97.49	4	8	5

Table 44: Meta-evaluation of the automatic metrics

The correlations and distances are quite good except for Text, and curiously better for the ASR and the verbatim than for the FTE. More precisely the correlations seem better when translations have low quality. The FTE scores are better and the correlations are lower than for the Verbatim which has lower scores but better correlations, etc. The metric that correlates best with human judgments is the Weighted N-Gram Model

3.6.3 Impact of ASR errors

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

To obtain Figure 7 and 8, we computed the SLT-mWER as a function of the ASR-WER (blue curves) 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).

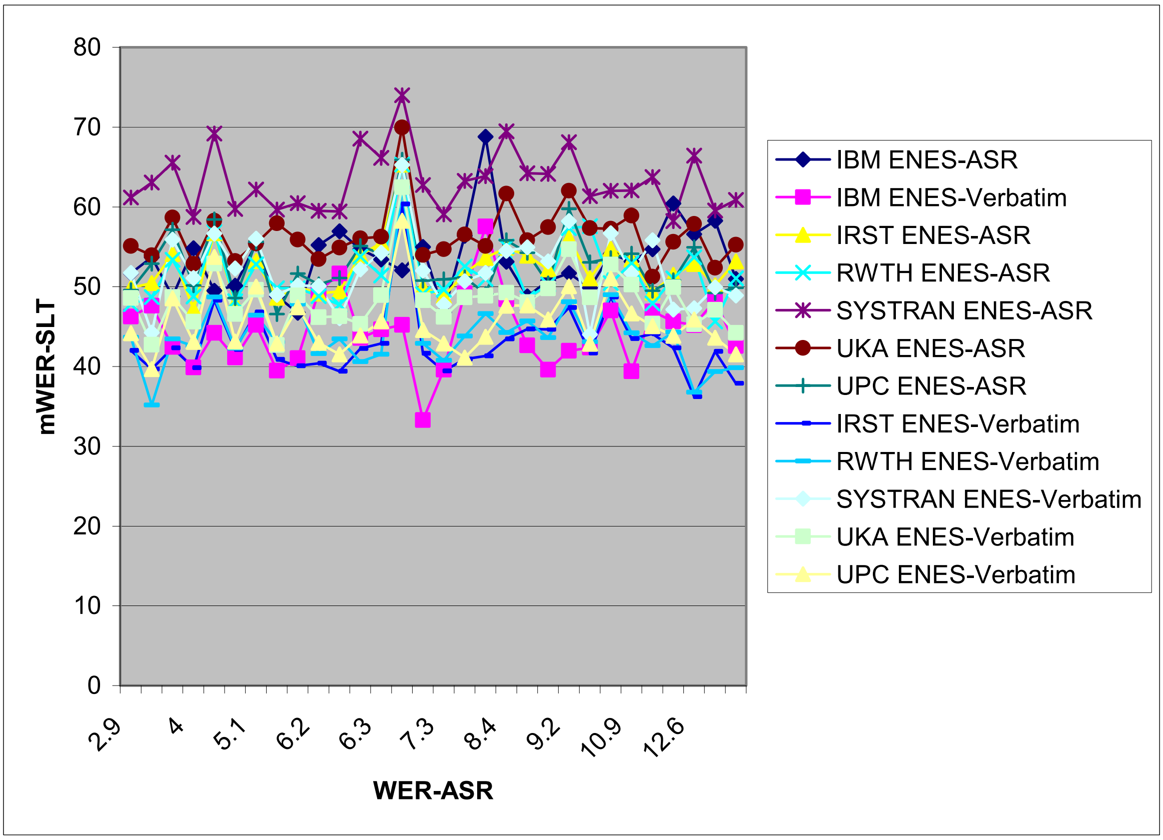


Figure 7: mWER-SLT as a function of WER-ASR for the English-to-Spanish EPPS task

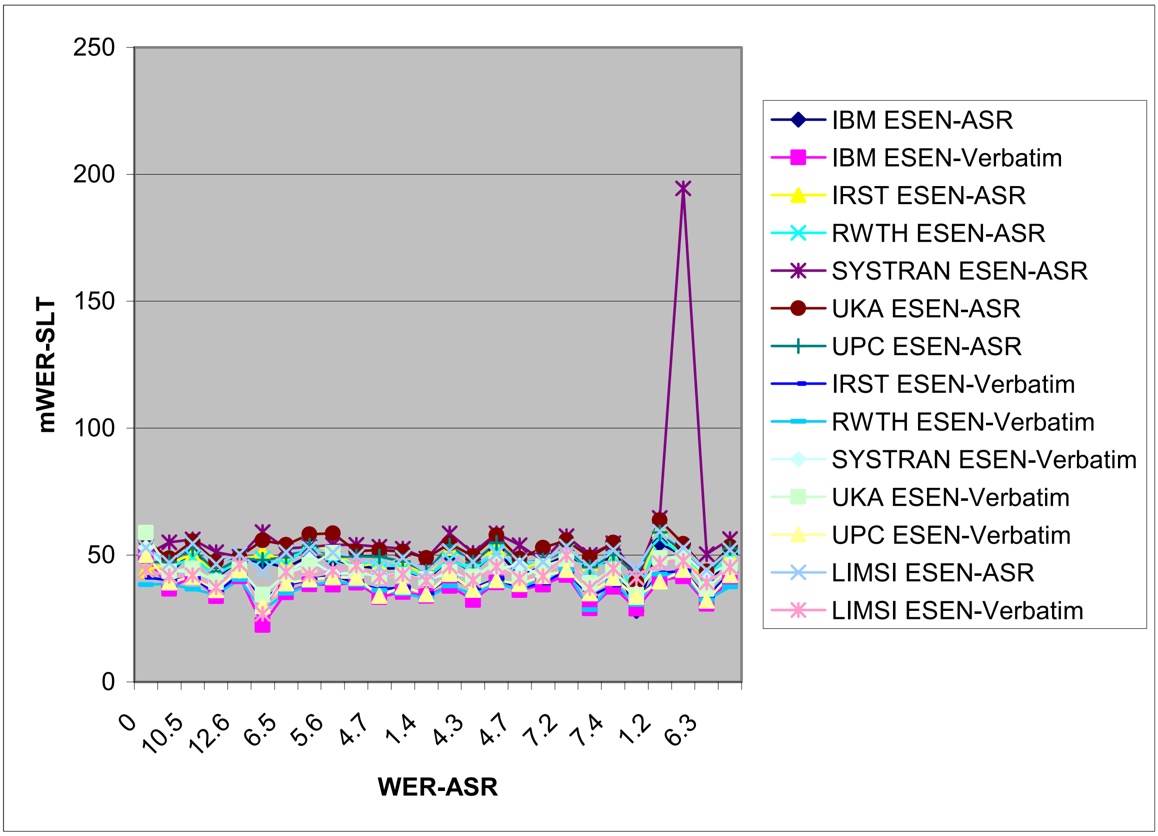


Figure 8: mWER-SLT as a function of WER-ASR for the Spanish-to-English EPPS task

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. As observed within the first evaluation campaign, there is no improvement for the SLT systems according to the improvement of the ASR output.

4 TTS Evaluation

The TTS evaluations were carried out in partnership with the ECESS[5] (European Center of Excellence on Speech Synthesis) consortium in order to define a better infrastructure and protocol and to attract a large number of external participants.

The state of art in TTS shows that there are less formalized worldwide evaluations than in ASR or SLT.

In order to help each participant benefits from TC-STAR infrastructure, it was decided to accommodate a large number of tasks across the 3 languages of TC-STAR. It is important to distinguish the major tasks aiming at evaluating globally TTS systems from diagnostic tests focusing on particular TTS modules. For more information, please refer to the TC-STAR Deliverable D8 [14].

4.1 Tasks and languages

17 different tasks and 3 languages have been considered for the evaluation of speech synthesis systems. The 3 languages are Chinese (Mandarin), English and Spanish. Table 45 gives an overview of the TTS evaluation tasks with the languages involved in each task.

Legend:	**CN**	Chinese Mandarin
	EN	English
	ES	Spanish

Evaluation task	Languages
Text processing	***CN, EN***
M1.1 Non Standard Word Normalization	EN
M1.2 End-of-sentence detection (EN) Words segmentation (CN)	CN, EN
M1.3 POS (Part-Of-Speech) Tagging	CN, EN
M1.4 Grapheme-to-phoneme conversion	CN, EN
Prosody Generation	***CN, EN, ES***
M2.1 Evaluation of prosody – Use of segmental information	EN, ES
M2.2 Evaluation of prosody – Rating of delexicalised utterances	CN, EN, ES
M2.3 Evaluation of prosody – Choice of a delexicalised utterance	CN, EN, ES
Acoustic Synthesis	***CN, EN***
M3.1 Intelligibility test (Semantically Unpredictable Sentences)	CN, EN
M3.2 Judgment test (Intelligibility and Naturalness)	CN, EN
Intra-lingual Voice Conversion (IVC)	***CN, EN, ES***
VC1 Comparison of speaker identities	
VC2 Evaluation of overall speech quality	
Crosslingual Voice Conversion (CVC)	***EN/ES, ES/EN***
VC1 Comparison of speaker identities	
VC2 Evaluation of overall speech quality	

Expressive speech	***ES***
ES1 Judgment test	ES
ES2 Comparison test	ES
TTS Component	***CN, EN, ES***
S1 Evaluation of the speech synthesis component (whole TTS System)	CN, EN, ES
S2 Evaluation of the speech synthesis component in the translation scenario. Intelligibility test using SUS sentences.	CN, EN, ES

Table 45: TTS evaluation tasks

The 17 TTS evaluation tasks and the corresponding evaluation methods and metrics are described in the following sections.
We used both automatic and semi-automatic metrics for the evaluation of the text processing module. The other modules were evaluated through subjective tests. Some information on how the subjective tests were carried out is given in section §4.5.

4.1.1 Normalization of Non-Standard-Words (M1.1)

This task consists in assessing the ability of the systems to disambiguate non-standard words (NSW). We used 11 NSW categories:

EXPN	abbreviations (e.g. “Mr.”)
LSEQ	letter sequence (e.g. “HCR”)
NDIG	number as digits (e.g. “Room 101”)
NUM	cardinal numbers
NORD	ordinal numbers (e.g. “the 3rd”)
NTIME	time (e.g. “11:45”)
NDATE	date (e.g. “23/09/05”)
NYEAR	year (e.g. “the 80ies”)
MONEY	money (e.g. “$3.45”)
PUNC	punctuation
PERC	percentage (e.g. “3,67%”)

Evaluation method: Manual check of the output of the text processing module.

Metric: Word Error Rate (number of not well normalized words per NSW category).

4.1.2 End-of-Sentence Detection / Words Segmentation (M1.2)

For English, this task consisted in evaluating the end-of-sentence (EOS) detection.

> **Evaluation method**: The submitted end-of-sentence boundaries were automatically compared to a segmentation of reference (manually segmented by an expert).
>
> **Metric**: Sentence Error Rate (SER) represents the percentage of sentences which are not correctly segmented.

For Chinese, it consisted in evaluating the segmentation of words.

> **Evaluation method**: The submitted word segmentations were automatically compared to a segmentation of reference (manually segmented by an expert).
>
> **Metric**: Word Error Rate (WER), which represents the percentage of words which are not correctly segmented.

4.1.3 POS Tagging (M1.3)

The POS (Part-Of-Speech) tagging consists of automatically tagging every word in a text with a grammatical tag reflecting the function of the word in the sentence (adverb, adjective, etc.).

Evaluation method: The participants processed the evaluation data set with their POS tagging algorithms. Submissions were then automatically compared to a reference POS tagging (the evaluation data set was manually tagged by an expert).

Metric: POS tags Error Rate (PER), i.e. the percentage of erroneous POS tags.

4.1.4 Grapheme-to-Phoneme Conversion (M1.4)

This task consists of automatically converting words into a sequence of phonemes, according to a predefined phoneme lexicon.

Evaluation method: The participants processed the evaluation data set with their grapheme-to-phoneme procedures. Submissions were then automatically compared to a phonetization of reference (i.e. the evaluation data set phonetized by an expert). The alignment between the submission and the reference is done using the SCTK software developed by NIST.

Metric: For Chinese, 3 metrics are used:

- CER (Character Error Rate) represents the percentage of Pinyin characters which were inserted, deleted or substituted for each word to be phonetized
- TER (Tone Error Rate) represents the percentage of Tones which were inserted, deleted or substituted for each word to be phonetized.
- CTER (Character and Tone Error Rate) represents the percentage of Pinyin characters and tones which were inserted, deleted or substituted for each word to be phonetized.

For English, 2 metrics are used:

- PhER (Phoneme Error Rate): computes the percentage of Phonemes which were inserted, deleted or substituted for each word to be phonetised.
- WER (Word Error Rate): computes the percentage of words which contain at least one erroneous phoneme.

4.1.5 Prosody – Use of segmental Information (M2.1)

The aim of this task is to evaluate the prosody module of each participating system. For each system and for N paragraphs (N=~3*Number of systems), ELDA generates synthetic speech based on the prosody descriptions submitted by each participants. To perform speech synthesis, ELDA uses a single generation toolkit developed by UPC.
The experimental protocol was the following:

- ELDA recorded the set of evaluation sentences and sent the text to the participants.
- The participants generated the prosody descriptions of the sentences and sent them back to ELDA.
- ELDA modified the prosody of sentences based on the prosody descriptions sent by the participants.

Evaluation method: Each subject is asked to rate the naturalness of the voice for N paragraphs, paying attention to the prosody. The evaluator is asked "How do you rate the naturalness of the sound of what you have just heard? (Pay attention to the melody of the sentences)" and must answer using a 5-point scale:

5 = Very natural

4 = Natural
3 = Neutral
2 = Unnatural (odd)
1 = Very unnatural (very odd)

Paragraphs are distributed randomly to the evaluators.

Metric: The average score is computed for each system.

4.1.6 Prosody – Rating of delexicalised utterances (M2.2)

For each system and for N paragraphs (N=~3*Number of systems), ELDA generates delexicalised utterances (i.e. lexical information is lost, only the melody and the temporal structure are kept and submitted to the evaluators) based on the prosody description generated by the participant systems and using a delexicalisation tool, developed at UPC.

Evaluation method: For each paragraph of the evaluation corpus, each subject listens to the delexicalised paragraph, reads the original text and judge if the prosody is good or not for that text. The evaluator is asked “How "good" is the melody of the sentences you heard for the text you read? (Pay attention to the melody of the sentences) (Does the melody fit to the text you heard?)” and must answer using a 5-point scale:

5 = Excellent
4 = Good
3 = Fair
2 = Bad
1 = Very bad

For more information on the evaluation method, please refer to the article of G. P. Sonntag [15].

Metric: The average score is computed for each system.

4.1.7 Prosody – Choice of a delexicalised utterance (M2.3)

For each paragraph of the evaluation corpus, each subject listens to the delexicalised paragraph and chooses from a set of 5 paragraphs which one is the most appropriate for the prosody he has just heard. The 5 paragraphs differ in phrase modality, boundaries, number of syllables, etc.

Evaluation method: A script automatically checks if the subject has chosen the right sentence. A score is computed for each sentence:

1 = Good answer
0 = Bad answer

For more information on the evaluation method, please refer to the article of G. P. Sonntag [15].

Metric: The average score is computed for each system.

4.1.8 Acoustic Synthesis - Intelligibility Test (Semantically Unpredictable Sentences) (M3.1)

Subjects are asked to listen to a set of Semantically Unpredictable Sentences (SUS) synthesized by the different participants and to write down what they have just heard. The SUS sentences were provided by ELDA.

Evaluation method: The sentences written by the evaluators are compared to the original texts of the SUS.

Metric: The WER (Word Error Rate) is computed for each system. It is the percentage of words that the subject did not recognize.

4.1.9 Acoustic Synthesis - Judgment Tests (Intelligibility and Naturalness) (M3.2)

For N paragraphs (N=~3*number of systems), each subject is asked to rate the intelligibility and the naturalness of the synthesized voices submitted by the participants.

Evaluation method: The evaluators are asked 2 questions:
Intelligibility: “How do you rate the intelligibility of the message you have just heard? (How "comprehensible" is it?)”. The answer is done using a 5-point scale:

5 = Excellent
4 = Good
3 = Fair
2 = Bad
1 = Very bad

Naturalness: **“**How do you rate the naturalness of the sound of what you have just heard?”. The answer is done using a 5-point scale:

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

Metric: The average *Intelligibility* and *Naturalness* scores are computed for each system.

4.1.10 Voice Conversion – Comparison of Speaker Identities (VC1)

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 synthesized voice *B* (target voice).
In the case of intra-lingual voice conversion (ICV) 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 synthesized voice *B*. For different pairs of voices, subjects are asked to judge if the 2 voices come from the same person.

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

- target voice versus transformed (converted) voice,
- and target voice versus source voice.

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 ICV and CVC tests were exactly the same (comparison of pairs of sentences spoken in the same language). Only the training data was different.

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 analyzing 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":

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

Metric: The average comparison score is computed for each voice conversion system.

4.1.11 Voice Conversion – Evaluation of Overall Speech Quality (VC2)

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 synthesized voices.

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

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

Metric: The average score is computed for each voice conversion system.

4.1.12 Expressive Speech – Judgment Tests (ES1)

For 8 different pairs "paragraph-systems", each subject is presented the synthetic speech and the context of the paragraph. Then the subject is asked to judge the expressiveness of the voice.

Evaluation method: The judgment scale is:

5 = the voice is very expressive but not appropriated in this context
4 = the voice is slightly expressive but is not appropriated in this context
3 = the voice is not expressive
2 = the voice is slightly expressive and appropriated in this context,
1 = the voice is very expressive and appropriated in this context.

Metric: The average score is computed for system.

4.1.13 Expressive Speech – Comparison Tests (ES2)

Each subject is asked to compare the expressiveness of 2 systems A and B, by listening to a set of sentences synthesized by A and B. Systems A and B are produced by the same site, including or not techniques to improve expressivity.

Evaluation method: The evaluator compares systems A and B using the following judgment scale:

5 = A is much more expressive than B
4 = A is a little more expressive than B
3 = A and B are equally expressive
2 = B is a little more expressive than A
1 = B is much more expressive than A

Metric: The average score is computed for each pair of system.

4.1.14 Evaluation of the whole TTS System (S1)

Each subject listens to N (N=~3*number of participating systems) synthesized sentences. Subjects are asked to rate a sentence according to the following categories, proposed by the ITU.P85 recommendations (see [16]).

Evaluation method: For each sentence they listen to, the evaluators are asked a series of 10 questions, to which they must answer using 5-point scales.
Overall Speech Quality: "How do you rate the quality of the sound of what you have just heard?"

5 = Excellent / 4 = Good / 3 = Fair / 2 = Poor / 1 = Bad

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

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

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

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

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

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

Articulation: "Were the sounds distinguishable?"

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

Speaking Rate: The average speed of delivery was:"

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

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

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

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

5 = Very easy / 4 = Easy / 3 = Neutral / 2 = Difficult / 1 = Very difficult

Pleasantness: "How would you describe the pleasantness of the voice?"

5 = Very pleasant / 4 = Pleasant / 3 = Neutral / 2 = Unpleasant / 1 = Very unpleasant

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

5 = Very smooth / 4 = Smooth / 3 = Neutral / 2 = Discontinuous / 1 = Very discontinuous

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

4.1.15 Evaluation of Intelligibility of the TTS System in the Translation Scenario (S2)

For N segments (N= ~17 in English, N= ~20 in Spanish), subjects are asked to listen to the output of the synthetic speech and to write what they have just heard.

Scoring: WER (Word Error Rate) by system: computes the percentage of words that the subject did not recognize. This percentage is computed using the original text as a reference.

4.2 Language Resources

For tasks M1.1, M1.2, M1.3 and M1.4, two sets of data were used, corresponding to the two last classical phases of an evaluation: development and test. For all the other tasks, training, development and test data sets were used.
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. This material is composed of:

- TTS evaluation corpus for National High-Tech program 863 TTS evaluation in 2003. (ref 2003-863-002. Copyright ChineseLDC, CIPSC. See http://www.chineseldc.org).
- Word Segmentation and POS tagging corpus for National High-Tech program 863 TTS evamiatop, in 2003. (ref 2003-863-008. Copyright ChineseLDC, CIPSC).

These data have been adapted to the TC-STAR tasks.

Furthermore, a speech database produced within the project by UPC, Siemens and Nokia was used.

4.2.1 TTS Training Data

The training data was developed by the WP3 partners as described in D8 (see [14]). This data was used for many tasks: prosody, acoustic synthesis, voice conversion, complete TTS system, etc.

For VC, only the C33 corpus was used (see Deliverable D8 [14]). For CVC, the English-Spanish data was used.

For the complete TTS system, external partners (and also IBM for Mandarin) used their own training data.

4.2.2 TTS Development 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.

For each evaluation task except voice conversion tasks, one sample of test data was sent to the participants: these data include an example of data sent to the participant and an example of what the participant should sent back to ELDA. ELDA was in charge of the production of development data. Development data are detailed in Annex 7.1.1.

4.2.3 TTS Test Data

Test data are of the same nature and format as development data. They include data sent to the participants (evaluation corpora) and, for the evaluation of text processing, reference data used for the scoring. ELDA was in charge of the test data set production. Test data sets are detailed in Annex 7.1.2.

The evaluation corpora are data to be evaluated. These evaluation corpora are subsets of the whole data sent to the participants (the "Inputs"). Each participant processes the data and sends their results back to ELDA.

In the case of the evaluation of text processing, processed data are compared to a "Reference". For the other evaluation tasks, subjective tests are carried out.

For subjective tests, the amount of data depends on the number of systems to be evaluated. This amount is computed taking into account the number of human judges (we want to minimize the subject effect and the sentence/paragraph effect). For each task implying subjective tests, each system should be rated at least 40 times.

4.2.4 Overview

As a whole, we have 23 Test data sets and 15 Development data sets.

Regarding Test data sets, there are:

- 8 evaluation data sets for Chinese used for 7 evaluation tasks.
- 10 evaluation data sets for English used for 9 evaluation tasks.
- 5 evaluation data sets for Spanish used for 9 evaluation tasks.

4.2.5 Validation of LRs

For English and Spanish, SPEX validated (through human experts):

- The end of sentence detection of the test sets.
- The POS tagging of the test sets.
- The phonetic transcription of the test sets.
- The phonetic segmentation of the test sets.

For Chinese, SPEX validated (through human experts):

- The end of sentence detection of the test sets.
- The POS tagging of the test sets.
- The phonetic transcription of the test sets.
- The syllable strength tagging of the test sets.

4.3 Schedule

The TTS run took place from 24 February to 3 March 2006.

Subjective tests were conducted from 8 March to 14 March 2006.

A second run for the subjective tests were conducted from 27 March to 31 March 2006. This second run were conducted for completing the evaluation of voice conversion and the evaluation of TTS component (ATT submissions has been included).

Scorings and evaluation results were computed from 15 March 2006 to 31 March 2006.

4.4 Participants and Submissions

There were 10 participant sites in the TTS evaluation, 6 from the TC-STAR consortium (IBM, IBM China, NOKIA, NOKIA China, UPC, and SIEMENS) and 4 external participants (AT&T, Chinese Academy of Science, LMU, and TUD).

There were 61 submissions (9 in Chinese, 26 in English, 26 in Spanish)

Participants and Submissions are reported in Table 46.

Participants (#submissions)/ Mod (Task ID)	**Text processing (M1)**	**Prosody Generation (M2)**	**Acoustic synthesis (M31/ M32)**	**TTS component (S1/S2)**	**Intra-lingual VC**	**Cross-lingual VC**	**Expressive Speech**
ATT (4)				EN (1) +ES (1)/ EN (1) +ES (1)			
CAS (4)	CN (1)	CN (1)		CN (1)	CN (1)		
IBM (9)		EN(1) ES (2)		EN (1) +ES (2)/ EN (1) +ES (2)			
IBM China (3)				CN (1)	EN (1) CN (1)		
Nokia (3)			EN(1)/ EN(1)		EN (1)		
Nokia China (3)			CN(1)/ CN(1)	CN (1)			
Uni. Dresden (4)			EN(2)/ EN(2)				
Uni. Munich (1)	EN (1)						
UPC (23)		EN(1) ES (2)		EN (1) +ES (2)/ EN (1) +ES (2)	EN (3) ES (3)	EN (2) ES (2)	ES (4)
Siemens (7)	EN (1)			EN (1)	EN (1) ES (1)	EN (1)	
#SUBMISSIONs	**3**	**7**	**8**	**20**	**12**	**7**	**4**
TOTAL	**61**						

Table 46: Participants and submissions

For each module (text processing, prosody generation, acoustic synthesis), a unique submission was done for each language. ELDA selected the submitted parts to perform the different evaluations.

4.5 Subjective Test Settings

Subjective tests were carried out via the web. An access to high-speed/ADSL internet connection and good listening material were required. The duration of the tests for each language was about 2 hours (tests have been designed to avoid a longer duration).
The following sections provide some details about the TTS human evaluations for English, Spanish and Chinese.

In Table 47, Table 48 and Table 49, the signification of columns is:

Col 1 "Evaluation Task" is the ID of the evaluation task (cf. task description).

Col 2 "# of Subjects" gives the number of evaluators who took part to the evaluation task. Not all evaluators were used for each task.

Col 3 "# of Evaluation Data" gives the total number of evaluation sentences used for the evaluation task. The number of submissions per evaluated system is also given (the natural voices are considered as a system here).

Col 4 "Average # of Tests / Subject" is the average number of subjective tests performed by each evaluator who took part to the evaluation task.

Col 5 "Total # of Tests" is the total number of subjective tests performed for the evaluation task.

4.5.1 Subjective Test Settings for English

A total number of 17 judges were recruited and paid to perform the English subjective tests. They were 18 to 40 years old native English speakers with no known hearing problem. No one was a speech synthesis expert. More details are given in Table 47.

Evaluation Task	# of Evaluated Systems[6]	# of Subjects	# of Evaluation Data	Average # of Tests / Subject	Total # of Tests
M2.1	3	15	27 (9 sentences for each system)	6.7	101
M2.2	3	15	27 (9 sentences for each system)	6.5	98
M2.3	3	15	27 (9 sentences for each system)	6.7	101
M3.1	4	15	240 (60 sentences for each system)	12	180
M3.2	4	15	36 (9 sentences for each system)	18	270
S1	5	13	45 (9 sentences for each system, including ATT)	10.8	140
S2	4	12	168 (42 sentences for each system, including ATT)	16.8	202
VC1	10	14	200 (20 sentences for each system)	37.3	522
VC2	11	12	220 (20 sentences for each system)	30	360

Table 47: Information about subjective tests for English

[6] In some cases, this includes the set of natural voices used as a baseline system.

4.5.2 Subjective Test Settings for Spanish

A total number of 19 judges were recruited and paid to perform the Spanish subjective tests. They were 18 to 40 years old native Spanish speakers with no known hearing problem. No one was a speech synthesis expert. More details are given in Table 48.

Evaluation Task	# of Evaluated Systems[7]	# of Subjects	# of Evaluation Data	Average # of Tests / Subject	Total # of Tests
M2.1	3	19	45 (15 sentences for each system)	9	171
M2.2	5	19	90 (18 sentences for each system)	9	171
M2.3	5	19	90 (18 sentences for each system)	9	171
S1	6	16	108 (18 sentences for each system, including ATT)	13.2	212
S2	5	16	200 (40 sentences for each system, including ATT)	19.9	318
VC1	7	15	140 (20 sentences for each system)	28	420
VC2	8	19	160 (20 sentences for each system)	15	285
ES1	4	19	32 (8 sentences for each system)	8	152
ES2	6	19	48 comparison pairs (8 sentences for each comparison)	12	228

Table 48: Information about subjective tests for Spanish

4.5.3 Subjective Test Settings for Chinese

A total number of 19 judges were recruited and paid to perform the Chinese subjective tests. They were 18 to 40 years old native Mandarin Chinese speakers with no known hearing problem. No one was a speech synthesis expert. More details are given in Table 49.

[7] In some cases, this includes the set of natural voices used as a baseline system.

Evaluation Task	# of Evaluated Systems[8]	# of Subjects	# of Evaluation Data	Average # of Tests / Subject	Total # of Tests
M2.2	3	19	18 (6 sentences for each system)	9	171
M2.3	3	18	18 (6 sentences for each system)	8.6	155
M3.1	2	17	98 (49 sentences for each system)	14.8	251
M3.2	2	17	10 (5 sentences for each system)	5	85
S1	4	17	24 (6 sentences for each system)	12.6	214
VC1	3	17	45 (15 sentences for each system)	22.6	384
VC2	4	17	60 (15 sentences for each system)	11	187

Table 49: Information about subjective tests for Chinese

4.6 Evaluation Results

4.6.1 Results for English

4.6.1.1 Text processing module (M1.1/1.2/1.3/1.4)

Participants:	**LMU**	University of Munich (external participant)
	SIE	Siemens

The results of the text processing evaluation for English are given in Table 50.

M1.1 – Evaluation of Normalization of NSW (Non-Standard Words)		
System	**Amount of data**	**WER (%)**
LMU	~400 words	20.0%
SIE	~400 words	36.1%
M1.2 – Evaluation of end-of-sentence detection		
System	**Amount of data**	**SER (%)**
LMU	500 sentences	0.4%
SIE	500 sentences	1.8% (0%)[9]

[8] In some cases, this includes the set of natural voices used as a baseline system.
[9] For the evaluation of end of sentence detection, Siemens considered the character ':' as end of sentence. If we considered ':' as an end of sentence marker, the error rate is 0%.

M1.3 – Evaluation of POS tagging			
System	**Amount of data**	**Pos Tag ER (%)**	
LMU	~10 000 words	6.5%	
SIE	~10 000 words	4.5%	
M1.4 – Evaluation of grapheme-to-phoneme conversion			
System	**Type & amount of data**	**Phone ER PER (%)**	**Word ER WER (%)**
	Common Words		
LMU	500 words	5.5%	24.4%
SIE	500 words	4.5%	22.0%
	Proper Names		
LMU	286 words	16.4%	50.3%
SIE	286 words	15.9%	50.7%
	Geographic Location		
LMU	214 words	18.7%	60.7%
SIE	214 words	16.1%	56.5%

Table 50: Text processing evaluation results for English

4.6.1.2 Prosody module (M2.1/2.2/2.3)

Participants:	**IBM**	IBM
	UPC	Polytechnic University of Catalonia

The results of the prosody module evaluation for English are given in Table 51.
Remark: For English, only the female voice (TC-STAR voice) was available and has been tested.

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

System	**Prosody M21**		**Prosody M22**		**Prosody M23**	
	Score(1<5)	***Rank***	**Score (1<5)**	***Rank***	**Score (0<1*)***	***Rank***
NAT	4.10	*1*	3.97	*1*	0.65	*1*
IBM	2.46	*2*	2.09	*2*	0.37	*2*
UPC	2.20	*3*	2.04	*3*	0.36	*3*

Table 51: Prosody evaluation results for English

4.6.1.3 Acoustic synthesis module (M3.1/3.2)

Participants:	**NOK**	Nokia
	TUD	University of Dresden (made 2 submissions)

Results are given in Table 52 and Table 53.

Table 52 gives the results of the intelligibility test M3.1: subjects were asked to listen to synthesized Semantically Unpredictable Sentences (SUS) and to write down what they heard. Using the original text as a reference, the Word Error Rate (**WER**) and Sentence error Rate (**SER**) were computed and are both reported in Table 52. The ranking of systems is also given in the bottom part of the table.

Legend:

NAT	Natural voice, used as top-line in subjective tests.
TUD1	1st TUD submission
TUD2	2nd TUD submission

M3.1				
System	**WER**		**SER**	
	Score	**Rank**	**Score**	**Rank**
NAT	2.8	*1*	20.0	*1*
NOK	16.5	*2*	57.8	*2*
TUD1	30.4	*4*	86.7	*4*
TUD2	28.9	*3*	84.4	*3*

Table 52: Results of the SUS intelligibility tests M3.1 for English

Table 53 gives the results of the judgment tests M3.2. Subject had to use two 5 point-scales rating *Intelligibility* and *Naturalness* of the synthesized voices (in both cases: '5' represents the best score and '1' the worse). These results are reported in Table 53 together with the ranking of systems (bottom part).

M3.2				
System	**Intelligibility**		**Naturalness**	
	Score (5>1)	**Rank**	**Score (5>1)**	**Rank**
NAT	4.73	*1*	4.56	*1*
NOK	3.51	*2*	2.46	*2*
TUD1	2.59	*3*	2.38	*3*
TUD2	2.07	*4*	1.60	*4*

Table 53: Results of the judgment tests M3.2 for English

In both tests M3.1 and M3.2, the best results were obtained by Nokia.

4.6.1.4 TTS component (S1, S2)

Participants:	**IBM**	
	SIE	Siemens
	UPC	Polytechnic University of Catalonia
	ATT	AT&T

The results are reported in Table 54 and Figure 9. 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.

Annex 7.2 provides a more detailed presentation of these results, including the 10 judgment categories of task S1.

It should be noted that IBM, Siemens and UPC participated with data produced within the project, while ATT submitted their own voices.

Legend:

NAT	Natural voice, used as top-line in subjective tests.
WER	Word Error Rate.
IC	Interval of Confidence (at 95%)
WSI	Wilson Score Interval (at 95%)

TTS Component Evaluation						
System	**S1 (Overall Quality)**			**S2**		
	Score	**IC**	**Rank**	**WER(%)**	**WSI**	**Rank**
NAT	4,79	± 0.17	*1*	-	-	-
IBM	3,13	± 0.24	*3*	6.8	[4.9 -9.3]	*1*
SIE	1,65	± 0.27	*5*	26.2	[22.6 - 30.2]	*4*
UPC	2,79	± 0.28	*4*	9.0	[6.8 - 11.8]	*2*
ATT	3,41	± 0.29	*2*	7.3	[5.3 - 9.9]	*3*

Table 54: Results of the TTS component evaluation tasks S1 and S2 (English).

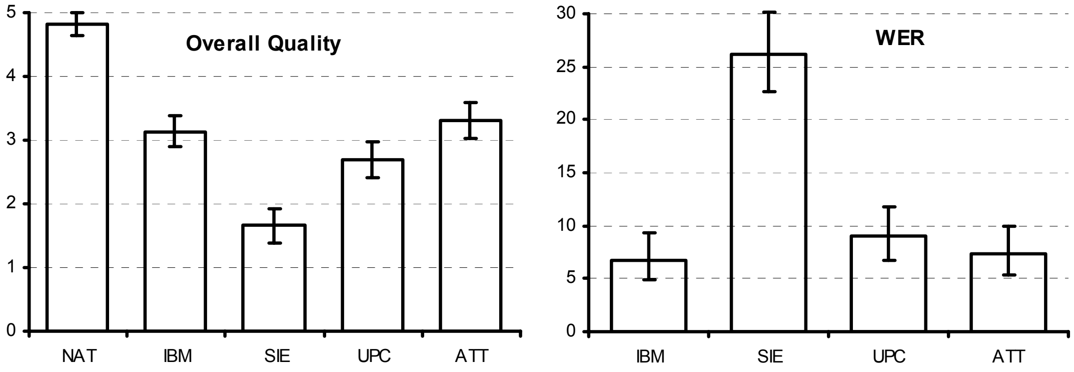


Figure 9: Overall Quality (S1) and WER (S2) results with intervals of confidence (English).

4.6.1.5 Voice conversion (VC1, VC2)

Participants:		
	SIE	Siemens
	UPC	Polytechnic University of Catalonia
	IBMc	IBM China
	NOK	Nokia

Results of the comparative tests (VC1) and the overall quality judgment tests (VC2) are reported in Tables below. Table 55 and Table 56 refer to the intra-lingual voice conversion task, Table 57 and Table 58 to the cross-lingual voice conversion task.

Legend:

IVC Intra-lingual voice conversion (English to English)
There were 5 IVC submissions: 1 from IBMc (IVC_IBMc), 1 from NOK (IVC_NOK) and 3 from UPC (IVC_UPC1, IVC_UPC2 and CVC_UPC3).

CVC Cross-lingual voice conversion (Spanish to English)
There were 3 CVC submissions: 1 from SIE (CVC_SIE) and 2 from UPC (CVC_UPC2 and CVC_UPC3). For training, only Spanish data of the target

speaker was available. With respect to source, UPC used bilingual data (which requires bilingual source) while Siemens only used English data.

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* is an English voice. Source voice *A* is an English voice (in the case of IVC) or a Spanish voice (in the case of CVC).
The *A->B* conversion consists in synthesizing voice *B* from the natural voice *A*. The conversion evaluation score results from comparing the natural voice *B* with the synthesized voice *B*.

SRC-TGT This result (last line) corresponds to the comparison of the natural source voice with the natural target voice (no conversion).

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_SIE	2.73	*5*	2.02	*6*	2.38	*4*	2.15	*4*
IVC_UPC1	3.63	*1*	4.30	*1*	3.67	*1*	3.70	*1*
IVC_UPC2	3.47	*2*	3.60	*2*	3.57	*2*	3.27	*2*
IVC_UPC3	2.88	*4*	3.17	*3*	2.57	*3*	3.07	*3*
IVC_IBMc	2.22	*6*	2.07	*5*	1.47	*6*	1.73	*6*
IVC_NOK	3.10	*3*	3.05	*4*	2.20	*5*	1.77	*5*
SRC-TGT	2.47	-	1.83	-	1.60	-	1.87	-

Table 55: Results of the intra-lingual voice conversion comparison tests VC1 for English

System	**IVC2 (Overall Quality)**	
	Score (1<5)	**Rank**
SOURCE	4.80	-
TARGET	4.78	-
IVC_SIE1	3.12	*2*
IVC_UPC1	1.61	*6*
IVC_UPC2	1.78	*5*
IVC_UPC3	2.23	*3*
IVC_IBMc	4.09	*1*
IVC_NOK	2.09	*4*

Table 56: Results of the intra-lingual voice conversion quality judgment tests VC2 for English

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_SIE	2.20	*3*	1.78	*3*	1.87	*2*	2.23	*3*
CVC_UPC2	2.53	*2*	2.25	*2*	1.48	*3*	2.57	*1*
CVC_UPC3	2.63	*1*	2.63	*1*	2.58	*1*	2.52	*2*
SRC-TGT	2.47	-	1.83	-	1.60	-	1.87	-

Table 57: Results of the cross-lingual voice conversion comparison tests VC1 for English

System	CVC2 (Overall Quality)	
	Score (1<5)	Rank
SOURCE	4.80	-
TARGET	4.78	-
CVC_SIE	3.40	*1*
CVC_UPC2	1.58	*3*
CVC_UPC3	2.13	*2*

Table 58: Results of the cross-lingual voice conversion quality judgment tests VC2 for English

Figure 10 (below) plots the VC2 scores versus the averaged VC1 scores for the different participating systems. As it can be seen, there is a trade-off between the conversion rate (identity) and the quality.

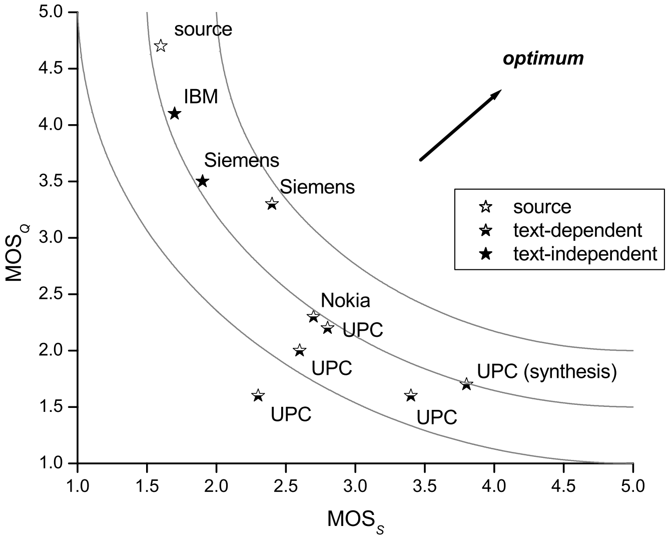


Figure 10: Plot of the VC2 scores (MOS_Q) vs. VC1 scores (MOS_S).

4.6.2 Results for Spanish

4.6.2.1 Prosody module (M2.1/2.2/2.3)

Participants:	**UPC**	Polytechnic University of Catalonia
	IBM	IBM

The results of the prosody module evaluation for Spanish are given in Table 59. These results are first given mixing the male and female data, then the detailed results, obtained with male voices only and female voices only are reported. The systems' ranking is given in each case.

Legend:

NAT	Natural voice, used as top-line in subjective tests.
IBM_F	IBM submission using female voice.
IBM_M	IBM submission using male voice.
UPC_F	UPC submission using female voice.

UPC_M UPC submission using male voice.

System	Prosody M21		Prosody M22		Prosody M23	
	Score(1<5)	*Rank*	Score (1<5)	*Rank*	Score (0<1*)*	*Rank*
NAT	4.19	*1*	3.58	*1*	0.75	*1*
IBM_F	-	-	3.19	*4*	0.53	*5*
IBM_M	2.77	*2*	3.17	*5*	0.75	*2*
UPC_F	-	-	3.30	*2*	0.69	*3*
UPC_M	2.48	*3*	3.25	*3*	0.61	*4*

Table 59: Prosody evaluation results for Spanish

If we don’t take into account the natural voice, the best results were obtained by the male voice of IBM, in the M21 and M23 tasks, followed by the female voice of UPC,. In the M23 task, IBM even obtains the same result as the natural voice.
In the M22 task, the female voice of UPC obtains the best results.

4.6.2.2 TTS component (S1, S2)

Participants: **IBM**
UPC Polytechnic University of Catalonia
ATT AT&T

The results are reported in Table 60 and Figure 11. 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.
Annex 7.2 provides a more detailed presentation of these results, including the 10 judgment categories of task S1.
It should be noted that IBM and UPC participated with data produced within the project, while ATT submitted their own voices.

Legend:
NAT Natural voice, used as top-line in subjective tests.
IBM_F/M IBM submission using female / male voices
UPC_F/M UPC submission using female / male voices
ATT_F ATT submission using female voices (no submission with male voices).
WER Word Error Rate.
IC Interval of Confidence (at 95%)
WSI Wilson Score Interval (at 95%)

TTS Component Evaluation						
System	**S1 (Overall Quality)**			**S2**		
	Score	**IC**	**Rank**	**WER(%)**	**WSI**	**Rank**
NAT	4.66	±0.22	*1*	-	-	-
IBM_F	3.92	±0.25	*4*	4.8	[3.3 - 6.9]	*2*
IBM_M	4.33	±0.24	*2*	7.7	[5.8 - 10.2]	*4*
UPC_F	3.32	±0.43	*6*	4.7	[3.2 - 6.8]	*1*
UPC_M	4.22	±0.23	*3*	5.0	[3.5 - 7.1]	*3*
ATT_F	3.78	±0.28	*5*	8.5	[6.5 - 11.1]	*5*

Table 60: Results of the TTS component evaluation tasks S1 and S2 (Spanish)

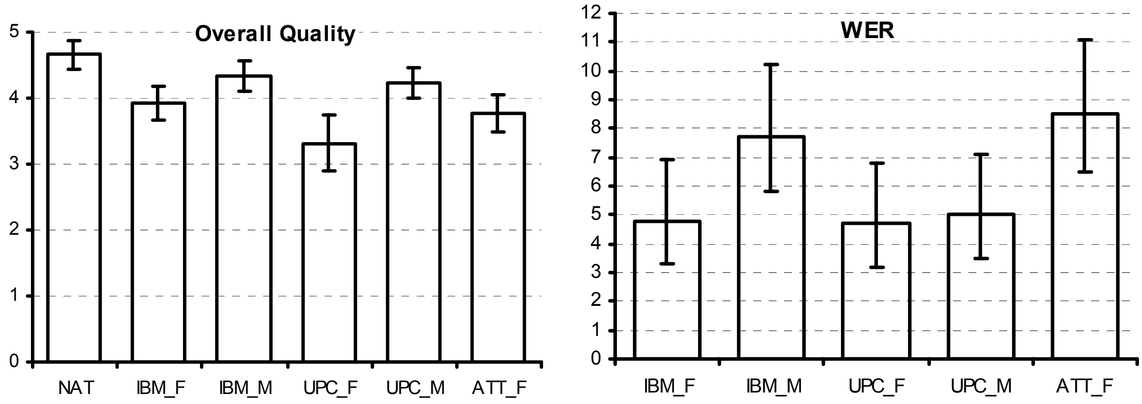


Figure 11: Overall Quality (S1) and WER (S2) results with intervals of confidence (Spanish).

4.6.2.3 Voice conversion (VC1, VC2)

Participants: **SIE** Siemens
UPC Polytechnic University of Catalonia

Results of the comparative tests (VC1) and the overall quality judgment tests (VC2) are reported in Tables below. Table 61 and Table 62 refer to the intra-lingual voice conversion task, Table 63 and Table 64 to the cross-lingual voice conversion task.

Legend:

IVC Intra-lingual voice conversion (Spanish to Spanish)
There were 4 IVC submissions: 1 from Siemens (IVC_SIE) and 3 from UPC (IVC_UPC1, IVC_UPC2 and CVC_UPC3).

CVC Cross-lingual voice conversion (English to Spanish)
There were 2 CVC submissions, both from UPC (CVC_UPC2 and CVC_UPC3).

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* is a Spanish voice. Source voice *A* is a Spanish voice (in the case of IVC) or an English voice (in the case of CVC).

The *A*->*B* conversion consists in synthesizing voice *B* from the natural voice *A*. The conversion evaluation score results from comparing the natural voice *B* with the synthesized voice *B*.

SRC-TGT This result (last line) corresponds to the comparison of the natural source voice with the natural target voice (no conversion).

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_SIE	2.48	*4*	2.08	*4*	2.32	*4*	2.28	*4*
IVC_UPC1	3.20	*1*	3.80	*2*	3.65	*1*	2.73	*3*
IVC_UPC2	3.13	*2*	3.95	*1*	2.93	*3*	3.85	*1*
IVC_UPC3	3.12	*3*	3.60	*3*	3.10	*2*	2.88	*2*
SRC-TGT	2.47	-	1.83	-	1.60	-	1.87	-

Table 61: Results of the intra-lingual voice conversion comparison tests VC1 for Spanish

System	**IVC2 (Overall Quality)**	
	Score (1<5)	**Rank**
SOURCE	4.80	-
TARGET	4.63	-
IVC_SIE	3.03	*2*
IVC_UPC1	3.20	*1*
IVC_UPC2	2.25	*4*
IVC_UPC3	2.38	*3*

Table 62: Results of the intra-lingual voice conversion quality judgment tests VC2 for Spanish

If we do not take into account the source and the target voices, which are natural voices, the best results, with respect to quality, were obtained by UPC for their 1st submission, followed by Siemens.

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_UPC2	3.50	*1*	3.45	*1*	2.60	*2*	3.27	*1*
CVC_UPC3	3.00	*2*	2.78	*2*	2.87	*1*	2.50	*2*
SRC-TGT	2.47	-	1.83	-	1.60	-	1.87	-

Table 63: Results of the cross-lingual voice conversion comparison tests VC1 for Spanish

System	**CVC2 (Overall Quality)**	
	Score (1<5)	**Rank**
SOURCE	4.80	-
TARGET	4.63	-
CVC_UPC2	1.63	*2*
CVC_UPC3	2.33	*1*

Table 64: Results of the cross-lingual voice conversion quality judgment tests VC2 for Spanish

4.6.2.4 Expressive Speech (ES1, ES2)

Participants: **UPC** Polytechnic University of Catalonia

Legend:

UPC_F1	UPC female baseline voice (without introducing features for expressive speech)
UPC_F2	UPC “expressive” female voice
UPC_M1	UPC male baseline voice (without introducing features for expressive speech)
UPC_M2	UPC “expressive” male voice

The results of the ES1 judgment tests are presented in Table 65. The judgment scale used by the evaluators was:

5 = the voice is very expressive but not appropriated in this context
4 = the voice is slightly expressive but is not appropriated in this context
3 = the voice is not expressive
2 = the voice is slightly expressive and appropriated in this context,
1 = the voice is very expressive and appropriated in this context.

System	Judgment Test ES1
	Score
UPC_F1	2.64
UPC_F2	2.10
UPC_M1	2.47
UPC_M2	2.81

Table 65: Results of the ES1 judgment tests for Spanish

The results of the ES2 comparison tests are given in Table 66. The evaluators compared systems A and B using the following judgment scale:

5 = A is much more expressive than B
4 = A is a little more expressive than B
3 = A and B are equally expressive
2 = B is a little more expressive than A
1 = B is much more expressive than A

System	ES2 score (comp test)	Conclusion
UPC_M1-UPC_M2	3.00	UPC_M1 is as expressive as UPC_M2.
UPC_F1-UPC_F2	2.51	UPC_F2 is a little more expressive than UPC_F1.

Table 66: Results of the ES2 comparison tests for Spanish

All the submissions were appropriated for the context. The expressivity features improved the result for the female speaker, but not for the male speaker. Results from ES2 are consistent with those obtained with tests ES1.

4.6.3 Results for Chinese

4.6.3.1 Text processing module (M1.2/1.3/1.4)

Participants: **CAS** China Academy of Sciences (external participant)

The results of the text processing evaluation for Chinese are given in Table 67.

Legend:

WER	Percentage of words which are not correctly segmented.
CER	Pinyin character Error Rate
CTER	Pinyin character and Tone Error Rate
TER	Tone Error Rate

M1.2 – Evaluation of word segmentation		
System	**Amount of eval. Words**	**WER (%)**
CAS	~2000 words	- [10]
M1.3 – Evaluation of POS tagging		
System	**Amount of eval. Words**	**Pos Tags ER (%)**
CAS	~2000 words	- [10]
M1.4 – Evaluation of grapheme-to-phoneme conversion		
System	**Amount of eval. Words**	**CER / CTER / TER (%)**
CAS	~2000 words	0.68% / 4.98% / 4.75%

Table 67: Text processing evaluation results for Chinese

4.6.3.2 Prosody module (M2.2/2.3)

Participants:	**CAS**	China Academy of Sciences (external participant)
	NOK	Nokia China

The results of the prosody module evaluation for Chinese are given in Table 68.
Remark: In the evaluation of the Chinese prosody module, the human speaker was not a professional speaker.

Legend:

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

System	**Prosody M22**		**Prosody M23**	
	Score (1<5)	***Rank***	**Score (0<1*)***	***Rank***
NAT	1.85	*3*	0.38	*2*
CAS	2.39	*2*	0.28	*3*
NOK	3.58	*1*	0.50	*1*

Table 68: Prosody evaluation results for Chinese

In these 2 tasks, Nokia's submission was the same as in the first evaluation. This voice obtained the best results.

4.6.3.3 Acoustic synthesis module (M3.1/3.2)

Participants:	**NOK**	Nokia China

[10] There are several ways to segment words in Chinese. In this evaluation, the reference text provided for word segmentation and POS tagging was not based on the same word segmentation conventions as the results submitted by CAS. Therefore the results obtained for these 2 evaluations are not meaningful and are not mentioned in this document.

Results are given in Table 69 and Table 70.

Table 69 gives the results of the intelligibility test M3.1: subjects were asked to listen to synthesized Semantically Unpredictable Sentences (SUS) and to write down what they heard. Using the original text as a reference, the Word Error Rate (**WER**) and Sentence error Rate (**SER**) were computed and are both reported in Table 69. The ranking of systems is also given in the bottom part of the table.

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

M3.1				
System	**WER**		**SER**	
	Score	**Rank**	**Score**	**Rank**
NAT	11.0	*1*	46.3	*1*
NOK	28.9	*2*	82.8	*2*

Table 69: Results of the SUS intelligibility tests M3.1 for Chinese

Table 70 gives the results of the judgment tests M3.2. Subject had to use two 5 point-scales rating *Intelligibility* and *Naturalness* of the synthesized voices (in both cases: '5' represents the best score and '1' the worse). These results are reported in Table 70 together with the ranking of systems (bottom part).

M3.2				
System	**Intelligibility**		**Naturalness**	
	Score (5>1)	**Rank**	**Score (5>1)**	**Rank**
NAT	4.39	*1*	4.29	*1*
NOK	3.24	*2*	2.61	*2*

Table 70: Results of the judgment tests M3.2 for Chinese

4.6.3.4 TTS component (S1)

Participants:	**CAS**	China Academy of Sciences (external participant)
	IBM	IBM China
	NOK	Nokia China

The results are reported in Table 71 and Figure 12. Only the Overall Quality test results are reported here. The intervals of confidence (at 95%) are also reported.
Annex 7.2 provides a more detailed presentation of these results, including the 10 judgment categories of task S1.
It should be noted that Nokia participated with data produced within the project, while CAS and IBM used their own data (but the same amount of data, approximately 8h, was used in the 3 cases).

Legend:
NAT Natural voice, used as top-line in subjective tests.
WER Word Error Rate.
IC Interval of Confidence (at 95%)

TTS Component Evaluation			
System	**S1 (Overall Quality)**		
	Score	**IC**	**Rank**
NAT	4.44	±0.28	*1*
CAS	3.58	±0.24	*3*
IBM	3.84	±0.20	*2*
NOK	2.77	±0.21	*4*

Table 71: Results of the TTS component evaluation tasks S1 (Chinese)

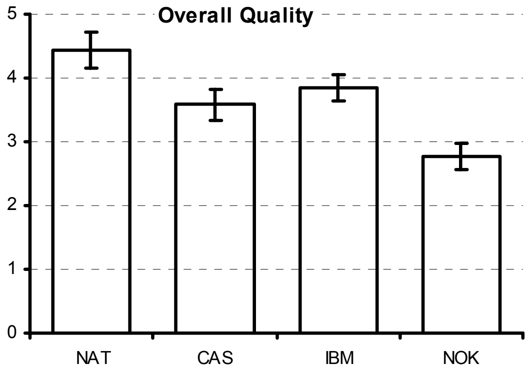


Figure 12: Overall Quality (S1) results with intervals of confidence (Chinese).

4.6.3.5 Voice conversion (VC1, VC2)

Participants:	**CAS**	China Academy of Sciences (external participant)
	IBM	IBM China

Results of the comparative tests (VC1) and the overall quality judgment tests (VC2) are reported in Table 72 and Table 73 respectively. There were only intra-lingual voice conversion (IVC: Chinese to Chinese) in this case.

Legend:

IVC Intra-lingual voice conversion (Chinese to Chinese)
There was no cross-lingual voice conversion (CVC) for Chinese.
The 2 IVC submissions are IVC_CAS (from CAS) and IVC_IBM (from IBM China).

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 Chinese voices.
The *A->B* conversion consists in synthesizing voice *B* from the natural voice *A*. The conversion evaluation score results from comparing the natural voice *B* with the synthesized voice *B*.

SRC-TGT This result (last line) corresponds to the comparison of the natural source voice with the natural target voice (no conversion).

Comparison test VC1 (*Scoring: 5>1*)						
Conversion System	**Conversion F(01)->M(02)**		**Conversion F(01)->F(03)**		**Conversion M(02)->F(03)**	
	Score	**Rank**	**Score**	**Rank**	**Score**	**Rank**
IVC_CAS	2,54	*1*	2,92	*1*	2,85	*1*
IVC_IBM	1,97	*2*	2,76	*2*	2,22	*2*
SRC-TGT	1,27	-	3,17	-	3,00	-

Table 72: Results of the voice conversion comparison tests VC1 for Chinese

In all conversion directions, CAS gets better results than IBM.

The VC2 results are reported in Table 73.

System	**VC2 (Overall Quality)**	
	Score (1<5)	**Rank**
SOURCE	4.28	-
TARGET	3.57	-
IVC_CAS	2.39	*2*
IVC_IBM	3.68	*1*

Table 73: Results of the voice conversion quality judgment tests VC2 for Chinese

IBM obtained better results than CAS and even better than the target voices, which are human voices.

4.7 Evaluation packages

As for ASR & SLT, 3 evaluation packages corresponding to the three languages will be available. They include development data, test data and scoring tools and are distributed by ELDA. They enable external players to test their systems and run the same evaluation but offline.

5 End-to-end evaluation

5.1 Tasks and conditions

In the second evaluation campaign of TC-STAR, an end-to-end evaluation has been planned. 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 for these questions to translation experts, it is preferable to use *adequacy* and *fluency* questionnaires, to be filled by human judges acting as potential users. In particular, we believe it is very difficult for an expert to make a *judgment* about the adequacy, based on the listening of the synthetic speech in the target language and the source speech. Instead, we use a *functional test* were the understanding is rated.

- **Adequacy**: comprehension test on potential users allows measuring the intelligibility rate.
- **Fluency**: judgment 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

Since this is the first time that end-to-end evaluation is conducted, only test data are produced, for English-to-Spanish.

Input/Reference
EPPS Spanish Domain. Input: Audio data: 20 * 3 minutes of speech in English. For each speech: - The corresponding ASR ROVER output (English). - The corresponding RWTH Primary output (Spanish). - The translated word alignment (Spanish)

5.3 Schedule

The end-to-end run and the evaluation took place in May 2006.

5.4 Participants and Submissions

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

Component	Input
ASR	ROVER
SLT	RWTH
TTS	UPC

Table 74: Test data

5.5 Protocol

ELDA recruited 20 subjects that were native Spanish speaker, 18-40 years old and with no hearing problem. They were not experts in speech synthesis and they were paid for the task. Subjects were required to have access to high-speed/ADSL internet connection and good listening material.

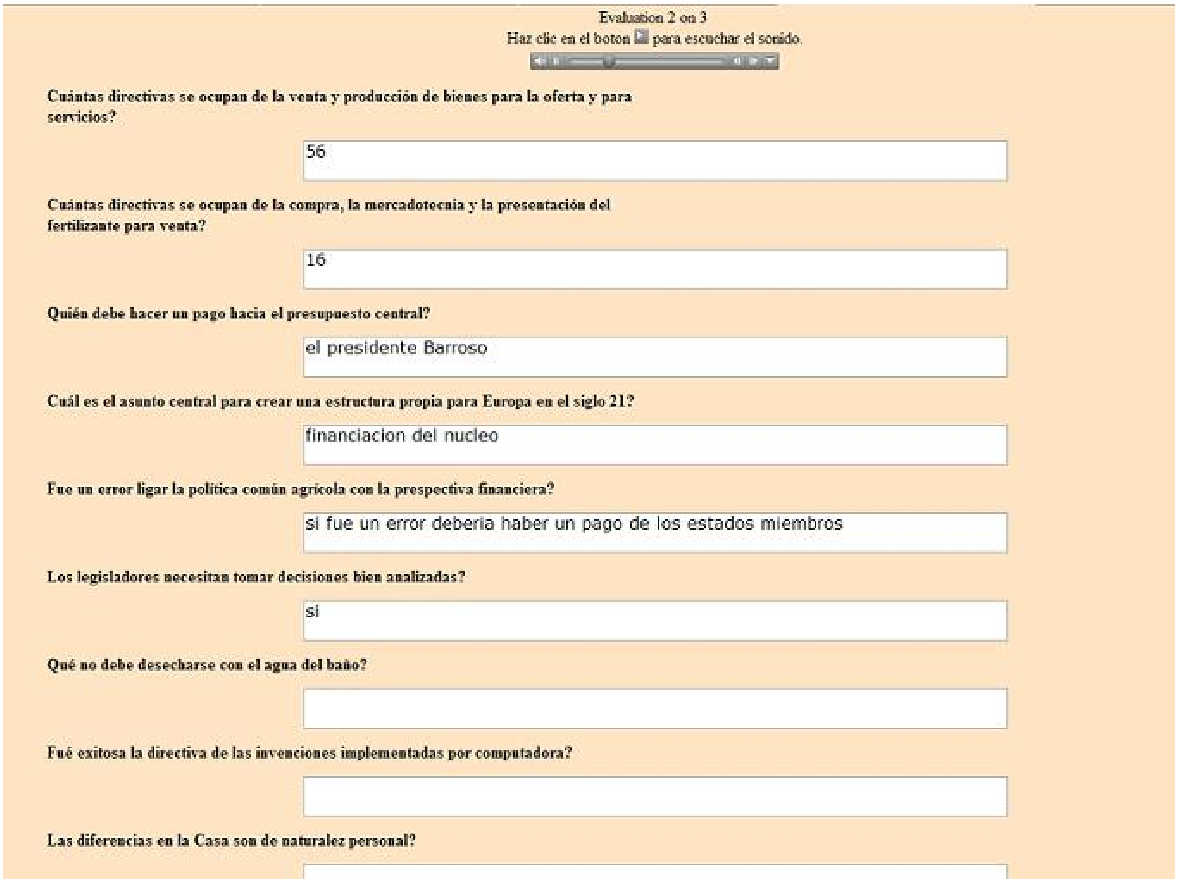
Subjective tests are carried out via the web. A specific interface has been developed, similar to the interface used for the SLT human evaluation.

Three evaluations by evaluator are given. As there are a total of 40 audio, some of the excerpts have been evaluated twice.

They are explained the TC-STAR system and the evaluation procedure. Furthermore, they listen to one minute of synthetic speech to become familiar with the voice, and complete one evaluation as a training session, which is not considered thereafter. Within the interface, the evaluator can play the sound corresponding to either TC-STAR speech or interpreter speech, during the evaluation session. Each evaluator assess at least one TC-STAR audio and one interpreter audio.

They are instructed to:

- read the questionnaire;
- listen the whole excerpt;
- listen a second time. They are allowed to stop the playback to write down the answers to the adequacy 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 política común agrícola 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 13: Interface for the end-to-end evaluation

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

Adequacy questionnaire:

For each excerpt, 20 comprehension questionnaires have been prepared, based on the English speech of the Final Text Edition, by a native English speaker. For each excerpt, 10 questions are asked to the subject about the excerpt he has just heard.

To prepare the questionnaire, the whole 200 questions have been created from the English Final Text Edition, 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.

After all the evaluations were done, a native Spanish person compared 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 an 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.

Fluency questionnaire:

After each excerpt, the evaluator has to rate the following questions (here in English):

Test	Question / Answers
Understanding	Do you think that you have understood the message? 1: not at all ....5: yes, absolutely
Fluently	The system is fluent? 1: No, it is very bad! ...... 5: Yes, it is perfect Spanish.
Effort	Rate the listening effort 1: very high .... 5: low, as natural speech
Overall Quality	Rate the overall quality of this translation system 1: Very bad, unusable; ...... 5: It is very useful

Table 75: Fluency questionnaire

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

5.6 Results

5.6.1 Fluency evaluation (subjective evaluation)

Speech	Audio	Understanding 1: low quality 5: better quality	Fluently 1: low quality 5: better quality	Effort 1: low quality 5: better quality	Overall Quality 1: low quality 5: better quality
Interpret	1	5	5	4	4
	2	4	3	2	4
	3	5	5	5	4
		4	5	4	5

	4	4	5	4	5
	5	3	3	3	3
		3	5	3	4
	6	2	1	1	1
		1	1	1	1
	7	2	3	3	2
		3	3	2	4
	8	4	4	4	5
	9	2	2	2	2
	10	5	5	4	5
	11	3	4	2	3
	12	2	1	5	1
		3	3	4	4
	13	3	1	3	2
		2	4	2	3
	14	3	3	3	3
		3	2	1	2
	15	4	4	4	5
		5	5	5	5
	16	3	1	2	2
		4	4	3	4
	17	4	4	4	4
		5	5	5	5
	18	3	4	4	4
	19	4	4	3	4
	20	5	5	4	5
		4	4	3	4
	mean	**3.45**	**3.48**	**3.19**	**3.52**
TCSTAR	1	3	1	2	2
	2	3	5	3	4
		1	1	1	1
	3	1	2	1	1
	4	1	2	1	2
		2	1	1	1
	5	3	2	1	2
		3	2	3	3
	6	3	1	2	1
	7	4	4	3	4
	8	4	3	2	2
	9	1	2	1	1
		2	1	1	1
	10	2	3	2	2
	11	4	3	2	4
	12	2	1	1	2
	13	3	1	1	1
	14	2	2	1	1

		1	1	1	1
	15	2	1	1	2
	16	3	2	3	2
		2	2	1	2
	17	2	1	1	1
		1	1	1	1
	18	3	2	2	3
		2	2	1	2
	19	2	2	1	2
		3	3	3	3
	20	3	2	1	2
	mean	**2.34**	**1.93**	**1.55**	**1.93**

Table 76: Fluency evaluation results

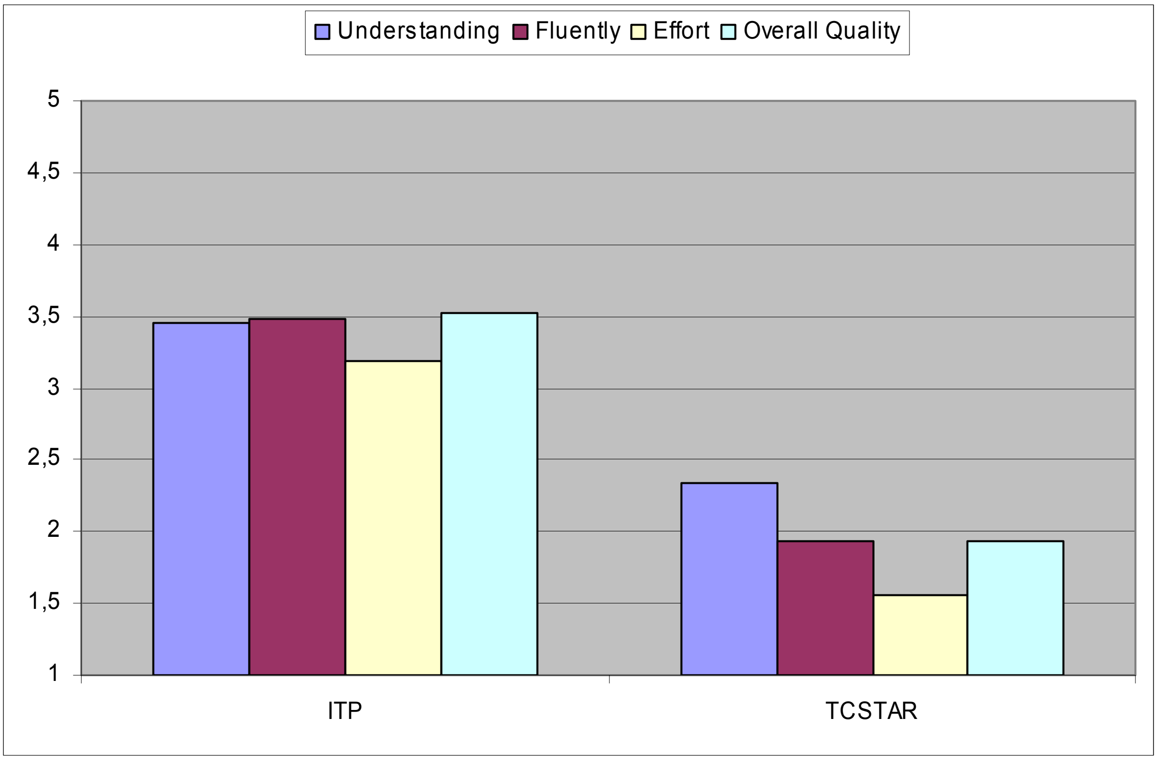


Figure 14: Mean of the fluency scores

With some exceptions, the audio are better for the interpreter than for the TC-STAR system, which has clearly a wide margin to improve. But the interpret results are not as good as supposed. This could be explained by the translation context of the interpreter, as they need to translate quickly.

For the TC-STAR system the message is not understood for less than the half of the information, and a quarter of the excerpts are not fluent and very hard to listen.

The four fluency scores are correlated, so we separate the fluency questions in the further evaluations, as it is done with the SLT human evaluation.

We tried to compare some of the excerpts and so as to understand why the TC-STAR system is better than the interpreter for those. Actually, the audios 2, 6, 7, 11 and 16 are clearly better for the TC-STAR system. For those audios we denote some deviations between the two systems:

- Noise interferences and back speaker: interpreter audio do not contain the single interpreter speak. We often hear the original speaker (speaking in English), more or less loud: it is very disturbing to have a good idea of what the interpreter say, in particular in the audio 7 whose we can not understand very well the hisses from the English speaker. Moreover, various noises are also present in interpreter audios, like knock table, low into the microphone, breathing, etc. Obviously there are not those noises in the TC-STAR audio, what facilitate the listening.
- Male or female voice: the TC-STAR system voice is always the same voice and the tone of voice is homogeneous and at a good level, while the interpreter can be a male or a female, and with a tone of voice which is not always the same between the different audio files. The voice could be high-pitched, or not, but it seems to make no difference for the scoring: for the interpreter, there is 11 female speeches and 9 males speeches (representing 16 female evaluations and 15 male evaluations), and scores are quite equivalent, as the table below shows.

Voice	**Understanding** 1: low quality 5: better quality	**Fluently** 1: low quality 5: better quality	**Effort** 1: low quality 5: better quality	**Overall Quality** 1: low quality 5: better quality
Male	3.47	3.20	3.33	3.33
Female	3.44	3.75	3.06	3.69

Table 77: Differences between male and female speak for the fluency evaluation

- Speaker hesitations: it often happen that the interpret hesitates. Contrary to the TC-STAR system, the interpreter listen to the speaker in the same time he speaks, and it can be really difficult to have a fluent speak. A good example is the audio 6 where the interpreter hesitates too much, makes pauses, take back himself, lengthens words, makes errors with the beginning of certain words, etc. This audio obtains the lower scores for all the fluency questions. But of course the hesitations could also be due two the quality of the English speaker, making errors for the interpreter.
- Break between two sentences: the interpreter makes pauses between two sentences as much the TC-STAR system inclined to concatenate the sentences. Then it is more difficult for an evaluator to follow the course of the speech, with the sequence of the sentences: the discourse is not clear.
- Problems of grammatical agreement (for gender, number and verb tense) as well as of sentence syntax: lack of agreement can pose a real problem towards understanding within the TC-STAR system. For example, in the following

sentence, we can find a clear case of combination of errors, which increases the difficulty of the understanding task of the evaluator:
Source sentence (in English): “that this debate is attended by French, German, Austrian, Belgian, British and other colleagues.”
Translated and synthesized sentence (in Spanish): “que este debate es ***la*** que ***asistieron*** francés alemán ***austríaca*** belga y británico y colegas de otros .”
The lack of agreement between "debate" (masculine) and "la" (feminine) triggers having to invest a bigger effort to understand the Spanish text. This is worsened by the fact that the verb is in 3rd person plural while the noun immediately following (beginning of the subject noun phrase) is in 3rd person singular but does not contain a determiner to indicate so. In fact, this noun represents the beginning of a list of country representatives, so it does not actually require the determiner. However, its lack of number agreement with its related verb form seems to impose our expecting it. Furthermore, one of the nouns in this subject noun phrase list is feminine, thus without gender agreement with the rest either.
All this is further complicated by the fact that the syntactic structure has been changed. The originally passive construction "is attended by" has been transformed into the inversed-order active-voice "es la que asistieron" (is that which attended) in Spanish, which would be actually more common than a passive in Spanish, but has probably represented a higher difficulty for the system to achieve a correct final gender and number agreement. Last but not least, there is also a change in syntactic structure between the English "other colleagues" and the Spanish "colegas de otros" (colleagues of others), which changes the meaning at the end of the sentence.
In conclusion, although all these may look like minor problems when considered individually, they pose serious problems for the evaluators when they come in a block. Evaluators need to make a much bigger effort to follow and understand the translation as they are easily distracted by trying to find the association between the non-agreeing words or by trying to restructure the sentence so as to accomplish a correct meaning.

- less natural connection words: the TC-STAR system produce connection words less natural, as “well”, “so”, etc. It also disturbs the listener, because in case the word has not the good tone, it seems to form a part of the significant information of the sentence, while it should be not. So the listener has to make additional effort to “delete” the word in order to understand the meaning of the sentence.

Unfortunately those problems can be applied to the adequacy evaluation, what makes the evaluation harder too.

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 subjective evaluation. It shows the percentage of good answers;
- fair E2E: objective verification of the question answers presence: the audio files had 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 ITP 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 was lost, files from each component (recognized files for ASR, translated files for SLT, and synthesized 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	**SLT** 0 : low 1 : better	**ASR** 0 : low 1 : better
ITP	1	0.70	0.90	--	--
	2	0.20	0.40	--	--
	3	0.65	0.70	--	--
	4	0.60	0.80	--	--
	5	0.35	0.60	--	--
	6	0.30	0.50	--	--
	7	0.30	0.60	--	--
	8	0.40	0.70	--	--
	9	0.30	0.80	--	--
	10	0.70	0.90	--	--
	11	0.40	0.50	--	--
	12	0.30	0.90	--	--
	13	0.25	0.70	--	--
	14	0.45	0.60	--	--
	15	0.75	0.80	--	--
	16	0.65	0.80	--	--
	17	0.75	0.80	--	--
	18	0.80	0.80	--	--
	19	0.40	0.50	--	--
	20	0.75	1.00	--	--
	mean	**0.50**	**0.72**	--	--
TCSTAR	1	0.80	1.00	1.00	1.00
	2	0.60	1.00	1.00	1.00
	3	0.50	0.90	0.90	1.00
	4	0.55	0.90	0.90	0.90
	5	0.60	0.90	0.90	1.00
	6	0.70	0.90	0.90	0.90
	7	0.50	0.80	0.80	0.90
	8	0.80	0.90	0.90	1.00
	9	0.30	0.90	0.90	1.00
	10	0.50	0.50	0.50	0.60
	11	0.70	0.90	0.90	0.90
	12	0.50	0.90	0.90	0.90
	13	0.60	0.60	0.60	0.60

	14	0.55	0.60	0.60	0.70
	15	1.00	1.00	1.00	1.00
	16	0.55	0.70	0.70	1.00
	17	0.25	0.70	0.70	0.80
	18	0.65	0.80	0.80	0.90
	19	0.60	0.70	0.70	1.00
	20	0.40	0.90	0.90	1.00
	mean	**0.58**	**0.83**	**0.83**	**0.91**

Table 78: Adequacy evaluation results

At first sight, the TC-STAR system is better than the interpreter. But it seems the evaluators can only answer half of the question whether the audio is the interpreter, and as a less important manner the TC-STAR output.

With respect to the subjective evaluation itself, there is a strong difference between the answers of the evaluators and the answers that were possible to find (i.e. between the subjective evaluation and the fair evaluation). The first conclusions that can be drawn from this are: it is difficult for the evaluators to find the answers. There is a difference of 22% for the ITP results and a difference of 25% for the TC-STAR results, with a similar difference rate. This could be due both to the quality of the audio output or to the subjectivity of the tests.

For the ITP, evaluators can only answer 50% of the questions, and the results are not really better for the TC-STAR system for which the evaluators can answer 58% of the questions. Only one audio (audio 18) has all the answer correct for the ITP, while there is two audio (audios 10 and 15) for the TC-STAR system. Most of the audio has a low-level of good answer, of course regarding the possible good answers of the fair evaluation.

The ITP contains 72% of correct answers, and so the overall loss is 28% of the information, while for the TC-STAR system the overall loss is 17% of the information.

Again for the ITP, 10 audios contain more than 80% of the answers considering the fair evaluation. An open question is then to find where the loss of information is. For that, there are some hypotheses:

- questions are too difficult, or badly asked;
- some human evaluators were not enough motivated,
- interpreters filter the information, due to the assigned time to translate speaker discourse,
- interpreters reformulate or paraphrase speaker discourse, which causes some ambiguous questions.

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 were present within the SLT text within the ASR text, in case the answer was not find before (actually we considered if information was found within a component - including subjective evaluation- information was also in the component upstream). Of course, for an objective comparison the person who checked the files had the reference answers in plain view.

Then we get the overall loss, which is 9% for the ASR component, and 8% more for the SLT component, while the subjective evaluation causes a 25% loss in addition. In conclusion, the original information (recovered by the ASR component) is really important, as the information within the two others components is quite stable after the recognition part.

There are three audio files containing 100% of the correct answers (while there is only one for the ITP). 14 audios contain more than 80% of the answers (4 audios more than ITP): 17 audios for the ASR component and 14 for the SLT component. Those differences can easily be explained: 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 summarizes the comparison between the two systems about the information loss.

	ITP	TC-STAR	
		SLT	ASR
Objective loss	28%	17%	9%
Subjective loss	50%	42%	-
Audios > 80%	10	14	17

Table 79: Information loss fot he two systems

To objectively compare ITP and TC-STAR, we 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 ITP. The same study as before has been done for the three components.

Speech	**Audio**	**subj. E2E** 0 : low 1 : better	**SLT** 0 : low 1 : better	**ASR** 0 : low 1 : better
TCSTAR (ITP 1.00 only)	1	0.89	1.00	1.00
	2	0.63	1.00	1.00
	3	0.43	0.86	1.00
	4	0.56	0.88	0.88
	5	0.75	1.00	1.00
	6	1.00	1.00	1.00
	7	0.50	0.67	0.83
	8	0.86	0.86	1.00
	9	0.31	0.88	1.00
	10	0.56	0.56	0.56
	11	0.60	1.00	1.00
	12	0.56	1.00	1.00
	13	0.86	0.86	0.86
	14	0.58	0.67	0.83
	15	1.00	1.00	1.00
	16	0.56	0.75	1.00
	17	0.31	0.75	1.00
	18	0.75	0.75	1.00
	19	0.57	0.71	1.00

	20	0.40	0.90	1.00
	mean	**0.63**	**0.86**	**0.95**

Table 80: Limited evaluation results

Here again the preserved information decrease, but the results are better in absolute. If we assume that the objective information loss is null for ITP, the TC-STAR system is not so good since the system lose 14% of the original information. The subjective loss is 37%, while the subjective loss for the ITP is 33%., and so the ITP quality could be slightly better than those of the TC-STAR system.

The ASR component loses 5% of the information and the SLT component loses 9% more.

	ITP	TC-STAR	
		SLT	ASR
Objective evaluation	100%	86%	95%
Subjective evaluation	67%	63%	-
Audios > 80%	10	14	17

Table 81: sum up of the evaluation

TC-STAR system needs to improve, but we get promising results, while it recovers 86% of the information that the interpreter could give. The quality of the system needs also to improve, but is quite the same than the one of the interpreter.

However the subjective impression is very bad and fluency needs to improve.

6 References

[1] Hunt M. « Figures of merit for assessing connected word recognizers» Speech Communication 9, p329-336

[2] Gauvain J.L. et *al* *ASR progress report* http://www.tc-star.org/documents/deliverable/tcstar_d7.pdf

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

[4] Ney H et *al* SLT progress report http://www.tc-star.org/documents/deliverable/Deliv_D5_Total_21May05.pdf

[5] Westphal M. *TC-STAR Recognition Baseline Results* http://www.tc-star.org/documents/deliverable/deliverable_updated14april05/D6.pdf

[6] Papineni, K; Roukos, S; Ward, T; Zhu, W-J. 2001. *Bleu : a Method for Automatic Evaluation of Machine Translation*, IBM Research Report RC22176 (W0109-022). Proceedings of the 40th Annual Meeting of the Association for the Computational Linguistics (ACL), Philadelphia, July 2002, pages 311-318.

[7] Babych, B; Hartley, A. 2003. *Modelling legitimate translation variation for automatic evaluation of MT quality*. In : Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisboa, Portugal, 27th may 2004. pages 833-836.

[8] Mostefa D. et *al TC-STAR Deliverable D12 Evaluation Report*

[9] Salton, G. and Buckley, C. 1988. *Term-weighting approaches in automatic text retrieval*. Information Processing & Management 24(5): 513–523

[10] Hamming, R. 1950. Error-detecting and error-correcting codes, Bell System Technical Journal 29(2):147-160

[11] van den Heuvel H. and Senders E. 2006. *Validation of Language Resources in TC-STAR* In proceedings of the TC-STAR Evaluation Workshop pp165-170 http://www.elda.org/tcstar-workshop/

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

[13] Hamon O., Popescu-Belis A., Choukri K., Dabbadie M., Hartley A., Mustafa El Hadi W., Rajman M., Timimi I., 2006. "CESTA: First Conclusions of the Technolanguage MT Evaluation Campaign". In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, May 2006, p.179-184.

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

[15] Sonntag G. P., Portele T., 1998. "PURR – a Method for Prosody Evaluation and Investigation", Journal of Computer Speech and Language, Vol. 12, No. 4, October 1998.

[16] ITU-T Recommendations P.85, "A Method for Subjective Performance Assessment of the Quality of Speech Output Devices", International Telecommunication Union publication 1994.

7 Annexes

7.1 TTS Data Sets

7.1.1 TTS 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.

For each evaluation task except voice conversion tasks, one sample of test data was sent to the participants: these data include an example of data sent to the participant and an example of what the participant should sent back to ELDA. ELDA was in charge of the production of development data. Development data are listed in Table 82.

Languages	*Dev data*
English	• TC-STAR XML DTD (Document Type Definition) for English • Input of the text processing module (SSML format). • Input of the prosody module (XML format). • Input of the acoustic synthesis module (XML format). • Input of the TTS component (SSML format) • ASR output, SLT output
Spanish	• TC-STAR XML DTD for Spanish • Input of the text processing module (SSML format). • Input of the prosody module (XML format). • Input of the acoustic synthesis module (XML format). • Input of the TTS component (SSML format) • ASR output, SLT output • Expressive speech development data set: input of ASR component , corrected output of ASR component, SLT input/output word alignment
Chinese Mandarin	• TC-STAR XML DTD for Chinese • Input of the text processing module (SSML format). • Input of the prosody module (XML format). • Input of the acoustic synthesis module (XML format). • Input of the TTS component (SSML format)

Table 82: TTS development data

7.1.2 TTS Test Data Sets

Test data are of the same nature and format as development data. They include data sent to the participants (evaluation corpora) and, for the evaluation of text processing, reference data used for the scoring. ELDA was in charge of the test data set production.

Test data sets are reported in Table 83.

Eval tasks	*Input/Reference*	*Amount of data : Input / Evaluation corpus*
	ENGLISH	
M1.1	Domain: English EPPS (European Parliamentary Plenary Session) FTE (Final Text Edition).	~150 000 words / 400 words
M1.2	Domain: English EPPS FTE. Ref: manually segmented sentences.	~150 000 words / 500 sentences
M1.3	Domain: English EPPS FTE. Ref: manually POS tagged words.	~150 000 words / 10 000 words
M1.4	Domain: English EPPS FTE. Ref: manually phonetised words.	~150 000 words / 1000 words (~50% common words, ~25% proper names, ~25% geographic locations)
M2.1 M2.2 M2.3	Domain: English EPPS FTE. Input: Input of the prosody module (paragraph, sentence, token, and word segmentations, POS tagging and phonetisation) Format: UTF-8 encoding, XML format (TC-STAR DTD)	9 paragraphs
M3.1	Input: Input of the acoustic synthesis module: (paragraph, sentence, token, and word segmentations, POS tagging, phonetisation, prosodic information (For each phoneme: duration, fundamental frequency, energy), syllabic information). These inputs are produced for 60 SUS (Semantically Unpredictable Sentences) sentences, which means for sentences syntactically correct but without meaning (semantically anomalous). Format: UTF-8 encoding, XML format (TC-STAR DTD)	60 SUS sentences
M3.2	Domain: English EPPS FTE. Input: Input of the acoustic synthesis module: (paragraph, sentence, token, and word segmentations, POS tagging, phonetisation, prosodic information (For each phoneme: duration, fundamental frequency, energy), syllabic information) Format: UTF-8 encoding, XML format (TC-STAR DTD)	9 paragraphs
S1	Domain: English EPPS FTE. SSML format.	9 paragraphs

S2	Domain: English EPPS FTE. ASR + SLT outputs.	202 segments / 42 segments (sentences).
VC	There are 4 conversion directions for English: (source -> target, F: female voice, M: male voice) 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, ELDA sends to the participants: - 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	5 files per conversion direction
SPANISH		
M2.1 M2.2 M2.3	Domain: Spanish EPPS FTE. Input: Input of the prosody module (paragraph, sentence, token, and word segmentations, POS tagging and phonetisation) Format: UTF-8 encoding, XML format (TC-STAR DTD)	18 paragraphs
S1	Domain: Spanish EPPS FTE. SSML inputs	18 paragraphs
S2	Domain: Spanish EPPS FTE. ASR + SLT outputs.	149 segments / 40 segments (sentences)
VC	There are 4 conversion directions for Spanish: (source -> target, F: female voice, M: male voice) 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 UPC Input data: For each source voice, ELDA sends to the participants: - audio files (channel 1, 96kHz, 24 bits) - xxL files: laringograph output (text files with the time of epoch closure) corresponding to the audio files	5 files per conversion direction

	- xxP files: phoneme segmentation corresponding to the audio files - xxS files: SAM files (text, prosodic information, etc.) corresponding to the audio files	
ES1 ES2	Domain: Spanish EPPS. Input: a) text of a document in the target language (Spanish); b) text of a selected paragraph (subset of the document) in the source (English) and target language; c) Reading of the selected paragraph, in the source voice and the labelling (corrected ASR output + translated word alignment).	8 documents
CHINESE		
M1.2	863 program data – GB2312 encoding Reference: manually segmented words.	~400 000 Chinese char. (~200 000 words)/ 2000-3000 words
M1.3	863 program data – GB2312 encoding Reference: manually POS tagged words.	~400 000 Chinese characters (~200 000 words)/ 2000-3000 words
M1.4	863 program data – GB2312 encoding Reference: manually phonetised words.	~21000 Chinese characters (~10000 words) / ~2000 words
M2.2 M2.3	Domain: 863 program data Input: Input of the prosody module (paragraph, sentence, token, word, and syllable segmentations, POS tagging and phonetisation) Format: UTF-8 encoding, XML format (TC-STAR DTD)	6 paragraphs (1 paragraph = ~ 120 Chinese char. (~60 words))
M3.1	Domain: 863 program data Input: Input of the acoustic synthesis module: (paragraph, sentence, token, word, and syllable segmentations, POS tagging, phonetisation (Pinyin characters), prosodic information (For each Pinyin character: duration, fundamental frequency, energy), syllabic information) Format: UTF-8 encoding, XML format (TC-STAR DTD)	50 short sentences
M3.2	Domain: 863 program data Input: Input of the acoustic synthesis module: (paragraph, sentence, token, word, and syllable segmentations, POS tagging, phonetisation (Pinyin characters), prosodic information (For each Pinyin character: duration, fundamental frequency, energy), syllabic information) Format: UTF-8 encoding, XML format (TC-STAR DTD)	6 paragraphs
S1	Domain: 863 program data Format: UTF-8 encoding, SSML format	12 paragraphs

VC	There are 3 conversion directions for Chinese: (source -> target, F: female voice, M: male voice) 01 (F) -> 03 (F) 01 (F) -> 02 (M) 02 (M) -> 03 (F) 01 is the Chinese female voice produced by Nokia 02 is the Chinese male voice produced by Nokia 03 is the Chinese female voice produced by Siemens Input data: For each source voice, ELDA sends to the participants: - 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	5 files per conversion direction

Table 83: TTS test data sets

7.2 TTS: Detailed Results of the TTS component Evaluation (S1, S2)

This section is a more detailed presentation of the TTS component evaluation results. For Spanish and English, 2 evaluation runs were performed, with and without the participation of ATT. The results of both runs are given. For the 3 languages, we also detail the results obtained in the 10 judgment categories of the S1 evaluation.

7.2.1 Detailed Results for English

7.2.1.1 1st evaluation without AT&T

Participants:	**IBM**	
	SIE	Siemens
	UPC	Polytechnic University of Catalonia

AT&T decided to take part to this evaluation task after this first evaluation run had been performed, involving IBM, Siemens and UPC. As a result, a new subjective evaluation run was performed with the submissions from the 4 participants (AT&T, IBM, Siemens, and UPC).
The next section will give the results of the 2nd evaluation run, involving AT&T.

Results are reported in Table 84 and Table 85.

Table 84 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 test data only comprised female voices.

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.

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

	S1									
System	**OQ**	**LE**	**Pr**	**C**	**A**	**SR**	**N**	**EL**	**Pl**	**AF**
Scores (1<5)										
NAT	4.58	4.59	4.77	4.84	4.61	4.67	4.48	4.41	4.31	4.33
IBM	3.42	3.63	3.55	3.94	3.79	4.43	2.42	2.71	3.18	2.70
SIE	1.85	2.26	2.38	2.4	2.31	3.73	1.67	1.61	2.16	1.50
UPC	2.84	2.92	3.02	3.49	3.25	3.83	2.26	2.13	2.82	2.28
Ranking										
NAT	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
IBM	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*
SIE	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*
UPC	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*3*

Table 84: Results of the TTS component evaluation S1 without the participation of ATT (English)

Whatever the category, IBM gets the best results, followed by UPC and Siemens (SIE).

Table 85 gives the results of intelligibility tests S2. Judges had to listen to synthesized Semantically Unpredictable Sentences (SUS) and to write down what they heard. Using the original text as a reference, the Word Error Rate (**WER**) and Sentence error Rate (**SER**) were computed and are both reported in Table 85. The ranking of systems is also given in the bottom part of the table.

S2				
System	**WER**		**SER**	
	Score	**Rank**	**Score**	**Rank**
IBM	7.3	*1*	47.3	*1*
SIE	25.6	*3*	88.9	*3*
UPC	12.6	*2*	62.5	*2*

Table 85: Results of the TTS component evaluation S2 without the participation of ATT (English)

The best results are obtained by IBM, followed by UPC and Siemens (SIE).

7.2.1.2 2nd evaluation with AT&T

Participants:	**IBM**	
	SIE	Siemens
	UPC	Polytechnic University of Catalonia
	ATT	AT&T

AT&T participated to this 2nd TTS component evaluation run. The tests of the previous section were carried out again. The results are reported in Table 86 and Table 87.

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.

Legend: see previous section.

S1										
System	**OQ**	**LE**	**Pr**	**C**	**A**	**SR**	**N**	**EL**	**Pl**	**AF**
Scoring (1<5)										
NAT	4,79	4,91	5,00	4,98	4,95	4,79	4,62	4,48	4,48	4,66
IBM	3,13	3,68	3,64	3,79	3,50	4,11	3,06	2,90	3,15	2,74
SIE	1,65	2,41	2,82	2,57	2,36	3,56	1,68	1,73	2,00	1,63
UPC	2,79	3,19	3,34	3,49	3,44	3,84	2,54	2,54	2,89	2,31
ATT	3,41	3,44	3,51	3,87	3,56	3,80	2,57	2,78	2,99	2,55
Ranking										
NAT	1	1	1	1	1	1	1	1	1	1
IBM	3	**2**	**2**	3	3	**2**	**2**	**2**	**2**	**2**
SIE	5	5	5	5	5	5	5	5	5	5
UPC	4	4	4	4	4	3	4	4	4	4
ATT	**2**	3	3	**2**	**2**	4	3	3	3	3

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

S2				
System	**WER**		**SER**	
	Score	**Rank**	**Score**	**Rank**
IBM	6.8	***1***	38.0	***1***
SIE	26.2	*4*	87.8	*4*
UPC	9.0	*3*	46.0	*2*
ATT	7.3	*2*	49.0	*3*

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

These results are consistent with the ones of the first evaluation run (i.e. without the participation of AT&T). The original system ranking remains unchanged (1st IBM, 2nd UPC and 3rd Siemens).
Regarding the S1 results, AT&T performs better than IBM in 3 categories (Overall Quality, Comprehension and Articulation). AT&T gets better results than UPC and Siemens in all categories, except Speaking Rate, where it is outperformed by UPC.

7.2.2 Detailed Results for Spanish

7.2.2.1 1st evaluation without AT&T

Participants:	**IBM**	IBM
	UPC	Polytechnic University of Catalonia

AT&T decided to take part to this evaluation task after this first evaluation run had been performed, involving IBM and UPC. As a result, a new subjective evaluation run was performed with the submissions from the 3 participants (AT&T, IBM, and UPC).
The next section will give the results of the 2nd evaluation run, involving AT&T.

Results are reported in Table 88 and Table 89.

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 test data only comprised female voices.

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.

Legend:

NAT	Natural voice, used as top-line in subjective tests.
IBM_F	IBM submission using female voices
IBM_M	IBM submission using male voices
UPC_F	UPC submission using female voices
UPC_M	UPC submission using male voices
Eva1_1_UPC_F	UPC submission of the 1st evaluation campaign using female voices
Eva1_1_UPC_M	UPC submission of the 1st evaluation campaign using male voices

S1										
System	**OQ**	**LE**	**Pr**	**C**	**A**	**SR**	**N**	**EL**	**Pl**	**A**
Scoring (1<5)										
NAT	4.61	4.89	4.89	4.94	4.67	4.97	4.58	4.36	4.28	4.33
IBM_F	3.56	4.33	4.11	4.69	4.14	4.53	3.33	3.5	3.72	3.28
IBM_M	4.33	4.61	4.56	4.69	4.31	4.47	3.86	3.89	4.00	3.44
UPC_F	3.89	4.36	4.14	4.56	3.64	4.08	3.25	3.17	3.56	2.97
UPC_M	4.00	4.28	4.00	4.44	4.11	4.17	3.36	3.47	3.67	3.25
Eval_1_UPC_F	3.67	3.92	3.92	4.25	3.96	4.17	3.21	3.17	3.67	2.75
Eval_1_UPC_M	3.67	4.17	4.08	4.5	4.04	4.37	3.37	3.37	3.67	3.17
Ranking										
NAT	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
IBM_F	*7*	*4*	*4*	*2*	*3*	*2*	*5*	*3*	*3*	*3*
IBM_M	*2*	*2*	*2*	*2*	*2*	*3*	*2*	*2*	*2*	*2*
UPC_F	*4*	*3*	*3*	*4*	*7*	*7*	*6*	*6*	*7*	*6*
UPC_M	*3*	*5*	*5*	*6*	*4*	*5*	*4*	*4*	*4*	*4*
Eval_1_UPC_F	*5*	*7*	*7*	*7*	*6*	*5*	*7*	*6*	*4*	*7*
Eval_1_UPC_M	*5*	*6*	*6*	*5*	*5*	*4*	*3*	*5*	*4*	*5*

Table 88: Results of the TTS component evaluation S1 without the participation of ATT (Spanish)

About Male & Female results:
If we do not take into account results obtained with the natural voice, the IBM male voice (IBM_M) yields the best results for all categories (except the "Speaking Rate" category). This confirms the good results obtained with the IBM male voice in the evaluation of prosody.
The IBM female voice (IBM_F) yields irregular results. Its score in the "overall quality" category is surprising: it is classified as the worst system, which is not confirmed by the results obtained in all the other categories. This has to be investigated.

The baseline voices of UPC from the first evaluation campaign were part of these tests. Like in the first evaluation, the male voice (Eval_1_UPC_M) gets overall better scores than the female voice (Eval_1_UPC_F).

S2				
System	**WER**		**SER**	
	Score	**Rank**	**Score**	**Rank**
IBM_F	4.4	*3*	33.8	*2*
IBM_M	3.0	*1*	25.0	*1*
UPC_F	8.4	*4*	42.1	*4*
UPC_M	3.2	*2*	34.2	*3*

Table 89: Results of the TTS component evaluation S2 without the participation of ATT (Spanish)

The best results were obtained with the male voice of IBM.

7.2.2.2 2nd evaluation with AT&T

Participants: **IBM**
UPC Polytechnic University of Catalonia
ATT AT&T

AT&T took part to this 2nd TTS component evaluation run, only for tests on female voices. The tests of the previous section were carried out again. The results are reported in Table 90 and Table 91.

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.

Legend: see previous section.
ATT_F ATT submission using female voices (no submission with male voices).

S1										
System	**OQ**	**LE**	**Pr**	**C**	**A**	**SR**	**N**	**EL**	**Pl**	**A**
Scoring (1<5)										
NAT	4.66	4.84	4.88	4.88	4.88	4.72	4.53	4.44	4.44	4.44
IBM_F	3.92	3.92	3.69	4.50	4.25	4.03	3.00	3.28	3.47	3.14
IBM_M	4.33	4.36	4.19	4.56	4.39	4.64	3.67	3.78	3.97	3.81
UPC_F	3.32	4.00	3.91	4.26	3.97	4.29	3.18	3.24	3.76	2.79
UPC_M	4.22	4.47	4.08	4.61	4.33	4.61	3.28	3.42	3.67	3.14
ATT_F	3.78	3.92	3.81	4.28	3.97	4.08	3.08	3.22	3.44	3.03
Eval_1_UPC_F	3.42	4.13	3.75	4.00	3.92	4.21	2.71	2.83	3.46	2.67
Eval_1_UPC_M	3.73	3.95	3.95	4.32	4.05	4.55	3.09	3.27	3.41	2.91

Ranking										
NAT	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
IBM_F	*4*	*7*	*8*	*4*	*4*	*8*	*7*	*4*	*5*	*3*
IBM_M	***2***	*3*	***2***	*3*	***2***	***2***	***2***	***2***	***2***	***2***
UPC_F	*8*	*5*	*5*	*7*	*6*	*5*	*4*	*6*	*3*	*7*
UPC_M	*3*	***2***	*3*	***2***	*3*	*3*	*3*	*3*	*4*	*3*
ATT_F	*5*	*7*	*6*	*6*	*6*	*7*	*6*	*7*	*7*	*5*
Eval_1_ UPC_F	*7*	*4*	*7*	*8*	*8*	*6*	*8*	*8*	*6*	*8*
Eval_1_ UPC_M	*6*	*6*	*4*	*5*	*5*	*4*	*5*	*5*	*8*	*6*

Table 90: Results of the TTS component evaluation S1 with the participation of ATT (Spanish)

Male Voices:
The results are consistent with the ones obtained in the previous case (i.e. without the participation of AT&T).
The IBM male voice scores better than UPC (except in the Listening Effort and Comprehension categories, this time).
Female Voices:
It is less obvious in the case of female voices. The IBM female voice is still ranked 1st in 5 categories out of 10. However, the overall performance of the UPC female voice (UPC_F) is comparable to one of IBM_F. The ATT voice is ranked 3rd.

S2				
System	**WER**		**SER**	
	Score	**Rank**	**Score**	**Rank**
IBM_F	4.8	*2*	25.4	***1***
IBM_M	7.7	*4*	38.1	*4*
UPC_F	4.7	***1***	32.3	*3*
UPC_M	5.0	*3*	31.7	*2*
ATT_F	8.5	*5*	49.2	*5*

Table 91: Results of the TTS component evaluation S2 with the participation of ATT (Spanish)

It should be noted that these results are in contradiction with those obtained with the previous human judges (test without AT&T) that ranked IBM first and UPC second.

7.2.3 Detailed Results for Chinese

Participants:	**CAS**	China Academy of Sciences (external participant)
	IBM	IBM China
	NOK	Nokia China

Table 92 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).

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.

Legend:

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

S1										
System	**OQ**	**LE**	**Pr**	**C**	**A**	**SR**	**N**	**EL**	**Pl**	**A**
Scoring (1<5)										
NAT	4.44	4.34	4.59	4.49	4.63	4.46	4.09	3.93	3.97	4.35
CAS	3.58	3.86	3.39	4.33	3.88	4.36	3.01	2.96	2.99	3.06
IBM	3.84	3.89	3.77	4.17	3.95	4.07	3.23	3.14	3.11	3.14
NOK	2.77	3.03	2.65	3.81	3.22	3.80	2.69	2.52	2.43	2.61
Ranking										
NAT	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
CAS	*3*	*3*	*3*	***2***	*3*	***2***	*3*	*3*	*3*	*3*
IBM	***2***	***2***	***2***	*3*	***2***	*3*	***2***	***2***	***2***	***2***
NOK	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*	*4*

Table 92: Results of the TTS component evaluation S1 (Chinese)

The best overall results were obtained by IBM.