Evaluation of Automatic Speech Recognition and Speech Language Translation within TC-STAR: Results from the first evaluation campaign

Djamel Mostefa*, **Olivier Hamon***†, **Khalid Choukri***

*Evaluations and Language resources Distribution Agency (ELDA)
55, rue Brillat Savarin 75013 Paris, FRANCE
† LIPN UMR 7030 - Université Paris 13 & CNRS
99 av. J.-B. Clément, 93430 Villetaneuse, FRANCE
{mostefa,hamon,choukri}@elda.org

Abstract

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

1. Introduction

The Speech-to-Speech Translation (SST) TC-STAR[1] project targets a selection of unconstrained conversational speech domains -speeches and broadcast news- and three languages: European English, European Spanish, and Mandarin Chinese. The long-term research goal of the project is effective speech to speech translation of unrestricted conversational speech on large domains of discourse.

To assess the advances in all SST technologies, annual competitive evaluations are organized. These evaluations are open to external participants.

This paper describes the evaluation activities in ASR and SLT (the TTS evaluation was organised later because no language resources were available for TTS evaluation at the moment) in the first TC-STAR Evaluation Campaign. The campaign took place in March 2005.

1.1. Evaluation tasks

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

- ***The EPPS task.*** The evaluation data consisted of audio recordings in English (En) and Spanish (Es), of the European Parliament Plenary Sessions (EPPS). The raw resources consisted of audio recordings of the parliamentary debates, and of the official documents published by the European Communities, that contained post-edited transcriptions of the sessions, in English and in Spanish. The focus was on the Parliament Members speaking in English and in Spanish, but the English and Spanish output of professional interpreters was also used in order to obtain enough data to work on. These corpora were used to evaluate translations from English into Spanish (En→Es) and from Spanish into English (Es→En).
- ***The VOA task.*** The evaluation data consisted of audio recordings in Mandarin Chinese (Zh), of the broadcast news of the Mandarin "Voice of America" (VOA) radio station. This data was used to evaluate translation from Mandarin into English (Zh→En).

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

1.2. Participants

There were 9 different participating sites including 2 outside TC-STAR consortium.

For ASR, there were 7 participating sites, all from the TC-STAR consortium. that submitted 30 different system results.

The total number of participants in the SLT evaluation campaign was 7; 5 from the TC-STAR consortium and 2 external participants. There were 97 submissions in total.

2. Automatic Speech Recognition Evaluation

2.1. Tasks and conditions

There were two tasks and three different training conditions for each task. For the EPPS task, automatic speech recognition systems were evaluated on recordings of the Parliament's sessions in English and Spanish from November 2004. For Mandarin, broadcast news recordings of December 1998 of the Mandarin radio "Voice of America" were used.

For each task, three training conditions were defined: the ***restricted*** training condition (participants could only use data produced within the TC-STAR project), the ***public*** data condition (all publicly available data could be used for training) and the ***open*** condition (any data before the cut-off date can be used). The cut-off date was October 16,

2004 for the EPPS task. For VOA, a black-out period that covered December 1998 was defined, instead of a cut-off date.

Classical evaluation metrics were used: Word Error Rate (WER) for the EPPS task and Character Error Rate (CER) for the VOA task.

2.2. Language resources for ASR

2.2.1. ASR Training Data Sets

Restricted: for the restricted condition, only data produced within TC-STAR could be used for training purposes (Van den Heuvel et al., 2006). This data was produced by using recordings of the European Parliament from 3 May to 14 October 2004. The audio files were recorded and manually transcribed. The Final Text Edition (FTE) of the documents published by the European Community, from April 1996 to June 2004, were also available.

Public: for the public condition, training data are data sets publicly available though various international Language Resources distribution agencies.

Open: for the open condition, any data before the cut-off date could be used.

2.2.2. ASR development data sets

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

For the VOA task, the development data comprised 3 hours of audio recordings from the broadcasted news of Mandarin "Voice of America" between 1 and 11 December 1998. It corresponds to approximately 42,000 Chinese characters or 30,000 running words.

2.2.3. ASR evaluation data sets

The same general procedure was followed for the production of the test data.

For EPPS tasks, the Parliamentary sessions from which the audio recordings were selected ran from 15 to 18 November 2004. The selected EPPS audio segments were no longer contiguous. The selection strategy consisted of first transcribing all available English (resp. Spanish) speaking politicians and then transcribing up to 3 hours of interpreters' speeches. For VOA tasks, the data was selected from news broadcasts between 14 and 22 December 1998.

2.3. Evaluation results

The results in terms of Word Error Rate for English and Spanish and Character Error Rate for Chinese are shown in table 1.

	English	Spanish	Chinese
system1	11.6	12.2	-
system2	13.4	13.7	-
system3	10.6	11.5	10.7
system4	24.6	-	-
system5	14.1	12.7	-
system6	50.0	-	-
system7	14.0	-	10.7
ROVER	9.9	10.1	-

Table 1: ASR results in percentage of WER or CER

2.3.1. English ASR Results

A total of 21 different submissions from 7 participating sites were submitted.

The best WER for a single system was 10.6%.

In general, the best results were obtained in the public condition. Nevertheless, the difference between the restricted and public condition is not very impressive. For example, system1 trained in the public condition performs 0.7% better in absolute than the one trained in the restricted condition from 12.3% to 11.6%. Either the task is very specific and only few corpora are useful for training or the data provided in the restricted conditions was sufficient for acoustic and language modeling.

In addition to the 21 submissions, a Recognizer Output Voting Error Reduction (ROVER) combination of all system outputs was performed. The ROVER system is able to reduce error rates by exploiting differences in the nature of the errors made by multiple ASR systems (Fiscus, 1997). The ROVER combination gave the best results with 9.9% WER.

2.3.2. Spanish ASR Results

There were 8 submissions from four different sites. The performances of primary systems range from 11.5% to 13.7%. A ROVER combination of all hypotheses was performed one more time. The ROVER gave the best result with a WER of 10.1%

2.3.3. Chinese ASR results

There was a common submission from two different sites for the Mandarin Voice of America task. The first system produces a first hypothesis. This one is then used by the second system in order to adapt acoustic models and then to produce the final recognition output. For this task, the CER is 10.7%.

2.4. Error analysis

Short words. Most errors are substitutions when 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).

Speaking style.The performance of the systems is dependent on the speaker and his/her way of speaking. On a fairly clean prepared speech, systems perform well. For example, the European Commissioner Chris Patten's speeches are well recognized. The WER on Mr. Patten speeches is 5,1%. For Spanish, this is also the case for

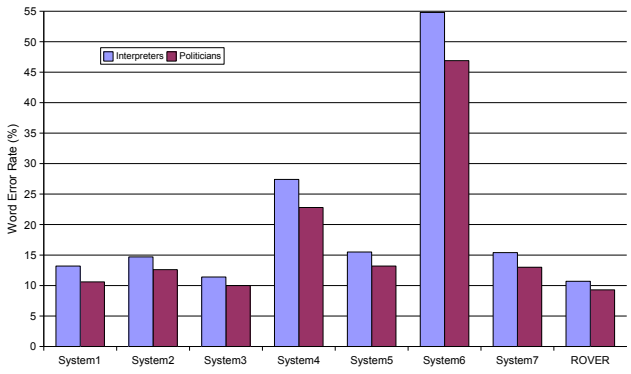

Figure 1: Recognition performance for the speech data of politicians and interpreters in the Spanish EPPS task

the President of the European Parliament, Josep Borrell. The WER is 7.79 % for Josep Borrell's speeches. On the contrary, the intervention of Robert Silk Kilroy had an error rate of 26.6%. This is mainly due to the context of the debate (the new Barroso Commission), the excitation and the words used by Mr Kilroy to express his opposition to the new Commission. Below is an alignment between the manual transcription of Mr. Kilroy's speech (denoted with REF) and the systems' output hypothesis (denoted with HYP).

*REF: THEY are a GAGGLE of rejects failures has BEENS no marks liars dodgy ********* CHARACTERS COMMUNISTS EPITOMIZED mister president by the British commissioner MANDELSON who LIED on HIS MORTGAGE application form so that HE COULD LIVE ABOVE HIS means*
HYP: THERE are a GOAL of rejects failures has BEEN no marks liars dodgy CHARACTER IS COMMONEST LIFE mister president by the british commissioner ALSO who LIVE on IS MORE application form so that IT CAN DELIVER BY THESE means

Politicians versus interpreters In the EPPS data there are two main categories of speakers: the interpreters and the politicians. We have computed the results for each category for English. The results are shown in Figure 1. Speeches from politicians are better recognized than those of interpreters. This is the case for all submissions. On the one hand, interpreters speak quite fast and with a variant speech flow. They try to maintain certain synchronization with the original speech. Sometimes, when they are running late, they speed up and speak very fast. In that case, some words are not pronounced well. On the other hand, interpreters' speeches include a lot of hesitations, false starts and corrections.

3. Spoken Language Translation Evaluation

3.1. Tasks and conditions

The tasks for SLT are the same as the ones for ASR in order to be able to chain the two components.
Two different tasks and three translation directions have been taken into account:

- ***EPPS task*** for En→Es and for Es→En
- ***VOA task*** for Zh→En.

For each translation direction, three kinds of text data were used as input:

- ***ASR.***This is the output of the automatic speech recognition systems. The ROVER combination (see section 2.3.), which gave the lowest error rate, was used for English-to-Spanish and Spanish-to-English. For VOA, the common submission was used as input. The text was in lower case and no punctuation marks were used.

- ***Verbatim***.The second type of data is the verbatim transcriptions. These are manual transcriptions produced by ELDA. The transcriptions include spontaneous speech phenomena, such as hesitations, corrections, false-starts, etc. The annotations were produced for English, Spanish and Mandarin. As for the ASR output, the text data was provided without punctuation, but here capitalization was used.

- ***Text***. Final Text Editions (FTE) provided by the European Parliament were used for the EPPS task and the clean transcriptions were used for the VOA task. These text transcriptions differ slightly from the verbatim ones. Some sentences are rewritten. The text data includes punctuations, uppercase and lowercase and does not include transcription of spontaneous speech phenomena. An example for each of the three kinds of inputs is shown below:

Text: *I am starting to know what Frank Sinatra must have felt like,*

Verbatim: *I'm I'm I'm starting to know what Frank Sinatra must have felt like*

ASR: *and i'm times and starting to know what frank sinatra must have felt like*

As for the ASR evaluations, different training conditions were distinguished.

3.2. Language Resources for SLT

3.2.1. SLT training data sets

The training data for the VOA task are data sets publicly available through various international LRs distribution agencies. For the EPPS task, the training data consisted of the same data as ASR training: the manual transcriptions of EPPS recordings in English and Spanish from 3 May to 14 October 2004, and the Final Text Editions (FTE) from April 1996 to 14 October 2004. The EPPS data was sentence-aligned for English/Spanish.

3.2.2. SLT development data sets

The SLT development set was built upon the ASR development data set, in order to enable future end-to-end evaluation. Subsets of 25,000 words were selected from the EPPS verbatim transcriptions, and from the FTE documents, in

English and in Spanish. Subsets of 25,000 characters were selected from the VOA verbatim transcriptions.

For each source language (Spanish, Mandarin, English) and each kind of input (verbatim, text) two reference translations were produced by professional translation agencies.

3.2.3. SLT evaluation data sets

The same amount of data was available for the evaluation.

In total, we have 18 data sets (3 translation directions, 2 development/test, 3 inputs).

For a given set, there is: the data to be translated in the source language and organized in documents and segments, the reference translations of the source data, prepared by professional translators, also organized in documents and segments and several candidate translations produced by the participants in the evaluation, that follow the same format of the reference set.

3.3. Evaluation Results

The same ASR input was used for all systems. It was the case-insensitive result of the ROVER combination of ASR hypotheses. Case information was used by evaluation metrics. Punctuation marks were present in the text input, but not in the ASR and verbatim transcriptions.

Metrics

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

BLEU, which stands for BiLingual Evaluation Understudy (Papineni et al., 2001), counts the number of word sequences (n-grams) in a sentence to be evaluated, which are common to one or more reference translations. A translation is considered to be better if it shares a larger number of n-grams with the reference translations. In addition, BLEU applies a penalty to those translations whose length significantly differs from that of the reference translations.

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

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, Mutli 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 (Babych and Hartley, 2004) is the same metric as BLEU but gives a higher importance to words such as names, events, terminological lexemes which are statistically more relevant.

All scores are given in percentages. For BLEU/NIST, WNM, the higher the percentage, the better the translation is. On the other hand, for mPER and mWER, which are error rate scores, the lower the percentage the better the translation, i.e. a 0 would represent a supposedly perfect translation.

The results of the Top 1 system in each direction is given in table 2.

3.3.1. Results for English-to-Spanish

The ratio between the source text in English and the reference translation in Spanish is 0.99, which outlines a strong correlation between the length of the source sentence and its corresponding translation. It is clear that systems which move away from this point of balance are penalized by automatic metrics.

We observe that results for mPER are approximately 15% better than those for mWER. This is understandable if we bear in mind that the best speech recognition outputs contain 9.9% of Word Error Rate (see section 2.3.). Furthermore, no capitalization or punctuation was provided in the ASR output. This makes the detection of named entities (proper names, country names, etc) more difficult for SLT systems. The Verbatim texts are case-sensitive and no recognition error is present in the sentences, even if no punctuation is provided. The Text input is the easiest data to deal with for SLT systems, since sentences are semantically and syntactically correct. Moreover, punctuation and capitalization are used.

3.3.2. Results for Spanish-to-English

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

3.3.3. Results for Chinese-to-English

The overview of table 2 indicates that the results obtained are considerably worse than for the other evaluation tasks. mPER and mWER show a difference of about 30%, which is also higher than before. Table 2 does not show a variation between the three inputs undertaken, namely ASR, Verbatim and Text. This can be explained by the fact that Verbatim and Text inputs are closer for VOA than for EPPS. The VOA text data contains clean transcriptions (no hesitation and fragment words) but is not rewritten as the EPPS text data.

3.4. Error Analysis

3.4.1. Impact of ASR errors

To obtain figure 2, we computed the mWER-SLT as a function of the ASR-WER for the Top1 system for En→Es. The second curve shows the result obtained for 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).

Both curves behave in a very similar manner. As anticipated, mWER results are worse when taking into account the translation of the ASR output. SLT mWER for the Verbatim task is always better than SLT mWER for the ASR task. However, to our surprise, the increase in SLT mWER for the ASR task was not as divergent as expected. After revising a number of translations to examine the type of errors produced and how they had been handled by WER, we came to the conclusion that mWER severely penalizes word order diversions in the translation, as can be seen in the example below:

Direction	Input	BLEU	NIST	mPER	mWER	WNM (f-measure)
En→Es	ASR	38.7	8.73	35.8	49.8	34.3
	Verbatim	42.5	9.33	32.7	46.1	36.9
	Text	46.3	9.66	31.1	41.2	36.3
Es→En	ASR	41,5	9,12	32,3	46,6	72,9
	Verbatim	45,9	9,75	28,6	42,5	73,7
	Text	53,3	10,5	25,6	35,1	78,1
Zh→Es	ASR	15.0	5.61	59.7	80.0	82,2
	Verbatim	15.7	5.80	58.5	79.4	78,0
	Text	12.5	5.40	62.7	83.6	74,3

Table 2: Evaluation results for the Top 1 system for each direction

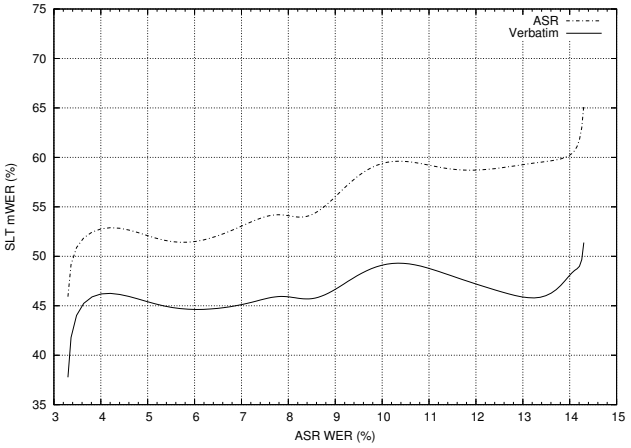

Figure 2: mWER-SLT as a function of WER-ASR for the En→Es EPPS task

Verbatim SRC: NATO's SFOUR was criticized by Amnesty International for human rights violations
TRANSL-Verb: de la OTAN Sfour fue criticada por Amnistía Internacional por violaciones de los derechos humanos
ASR SRC : nato's for was criticised by amnesty international for human rights violations
TRANSL-ASR : por la OTAN fue criticada por Amnistía Internacional por violaciones de los derechos humanos
REF-1: la SFOR de la OTAN fue criticada por Amnistía Internacional por violar los derechos humanos
REF-2: Amnistía Internacional criticó la SFOR de la OTAN por violaciones de los derechos humanos

In this example, both translations for the ASR and Verbatim inputs have obtained the same mWER score, however, we can see that the translation resulting from the Verbatim input (TRANSL-Verb) only suffers from a word order error (*de la OTAN Sfour* instead of *Sfour de la OTAN*), while the one resulting from the ASR input (TRANSL-ASR) has a word order problem among with incorrect lexical entry (preposition *por* instead of *de*) and a missing word (SFOUR).
This problem with mWER was most significant for the English-to-Spanish translation part, where we realized that the mWER metric was critical even with relatively correct output. This seems to be partly due to the fact that Spanish is a flexible word-order language, especially when compared to English. That means that 2 references are not sufficient for the evaluations of the possible translations into Spanish as word order can vary considerably in a Spanish sentence while being equally correct. That flexibility in word order seems to cause a constant increase in the mWER, resulting in the curve shown in Figure 2.

Impact of spontaneous speech phenomena Further to the evaluation results, where it was explained that the best SLT results are achieved with the Text data, then with the Verbatim transcriptions and, finally, with the ASR output, we would like to illustrate this with an example which also shows the impact of spontaneous speech phenomena in translation. If we consider the following example (initially introduced in Section 3.1.):

Text SRC: I am starting to know what Frank Sinatra must have felt like,
REF: empiezo a imaginar lo que debió sentir Frank Sinatra
TRANSL: Empiezo a saber qué Frank Sinatra debe han experimentado
Verbatim SRC: I'm I'm I'm starting to know what Frank Sinatra must have felt like
REF: empiezo a imaginar lo que debió de sentir Frank Sinatra
TRANSL: me estoy estoy comenzando a saber qué debe Frank Sinatra han sido
ASR SRC: and i'm times and starting to know what frank sinatra must have felt like
REF: empiezo a imaginar lo que debió de sentir Frank Sinatra
TRANSL: y estoy veces y comenzar a saber qué ebe Frank Sinatra han sido

We can see that translation is best for the Text data input, where all basic pieces of content are present and errors derive mostly from incorrect flexion (*han* instead of *haber*) and incorrect word ordering caused by the choice of interrogative pronoun *qué* (in the case of using this pronoun the subordinate construction should have been *qué debe haber experimentado Frank Sinatra*). However, if we move on to the translation of the Verbatim transcriptions, we find further problems, like the repetition of *estoy* derived from the translation of repetitions in the input text and the loss of some meaning with the wrong translation of *have felt like* into *han sido*. Finally, the translation of the ASR output

	ASR	Text	Verb.
BLEU / NIST	1	0.98	1
	1	0.98	0.99
BLEU / mPER	0.94	0.80	0.9
	-0.97	-0.91	-0.96
BLEU / mWER	0.89	0.92	1
	-0.91	-0.94	-0.99
BLEU / WNM (f-measure)	0.26	0.46	-0.5
	-0.06	0.51	-0.65
mPER / mWER	0.77	0.71	0.9
	0.94	0.75	0.97
mPER / WNM (f-measure)	0.37	0.43	-0.6
	0.24	-0.24	0.61
mWER / WNM (f-measure)	-0.09	0.66	-0.5
	0.42	-0.92	0.59

Table 3: Correlation metrics for Es→En

is certainly the one that poses the most problems. The elements incorrectly recognized by the ASR system have caused the translation system to generate the translation of non-existing elements such as *y...veces y*, together with the loss of meaning already suffered by the translation of the Verbatim transcriptions (translation of *have felt like* into *han sido*).

3.4.2. Statistical analysis of the evaluation metrics

Table 3 presents the metrics correlations, with, up to a cell, the rank correlation, and down the score correlation. As BLEU and NIST are strongly correlated (see the first line of the table below), we have decided not to compute the correlation between NIST and the other metrics. Score correlation between mPER (or mWER) and the other metrics must be inversed. Therefore a -1 correlation score with mPER (or mWER) means that the metrics are totally correlated. Of course this is not the same for rank correlation, nor for the two metrics between them.It is obvious that BLEU, NIST, mWER and mPER are strongly correlated, with a majority of rank and score correlation between 0.9 and 1.0. The last remark is about the WNM correlation with the other metrics. This correlation was not as high as expected. One explanation could be the use of a non-adapted statistical corpus (needed to recompute weights of the n-grams).

4. Conclusion

For ASR, thirty system outputs were submitted. Rover combinations were performed for English and Spanish. The best word error rate of is 9.9% for English, 10.6% for Spanish and 10.7% for Chinese. A large number of SLT systems participated in this evaluation.

Nevertheless, the WER must still be reduced further, as SLT systems need even lower error rates, especially as machine translation models improce. Around 50% of sentences contain one ore more recognition errors.

For SLT, we observed a surprisingly good performance by the EPPS tasks. The best results were obtained for the Spanish-to-English direction, while those of the Chinese-to-English VOA task were by far the worst, not reflecting the results obtained from the ASR part, with a CER of only 10.7%.

The next evaluations will take place in February 2006 and will focus on the same tasks and languages. More training data will be available and more external participants will be invited to the evaluations.

Evaluation packages

Evaluation packages that include resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during this campaign are available[2]. Six evaluation packages are available:

- TC-STAR 2005 Evaluation Package - ASR English
- TC-STAR 2005 Evaluation Package - ASR Spanish
- TC-STAR 2005 Evaluation Package - ASR Mandarin
- TC-STAR 2005 Evaluation Package - SLT English-to-Spanish
- TC-STAR 2005 Evaluation Package - SLT Spanish-to-English
- TC-STAR 2005 Evaluation Package - SLT Mandarin-to-English

The aim of each evaluation package is to enable external players to evaluate their own system and compare their results with those obtained during thecampaignitself.

5. References

B. Babych and A. Hartley. 2004. Modelling legitimate translation variation for automatic evaluation of mt quality. pages 833–836.

J. G. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). pages 347–352.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2001. Bleu : a method for automatic evaluation of machine translation. pages 311–318.

H. Van den Heuvel, K. Choukri, C. Gollan, A. Moreno, and D. Mostefa. 2006. Tc-star: New language resources for asr and slt purposes.

[2]http://www.elda.org/article204.html