Linguistic Data Annotation Specification:
Assessment of Fluency and Adequacy in Arabic-English and Chinese-English Translations
June 18, 2002

1. Goal

The goal of this effort is to evaluate the quality of TIDES research, human translations teams and commercial off-the shelf (COTS) systems. Translations are evaluated on the basis of adequacy and fluency. Adequacy refers to the degree to which the translation communicates information present in the original source language text. Fluency refers to the degree to which the translation is well-formed according to the grammar of the target language.

2. Data

The data evaluated includes multiple translations of 100 Chinese and 100 Arabic news stories. Data Selection information appears in table 1.

	TIDES	COTS	Human Ref. Text	Held Back Translations
Arabic	3	3	4	6
Chinese	8	2	4	0

In preparation for the translations, the original news stories are converted into a standard format that makes paragraph and segment boundaries explicit. Where an original story has explicit paragraph or sentence tags, these are also expressed in the new format. Otherwise, blank lines become paragraph boundaries while periods, question marks and exclamation points become sentence boundaries. On average, stories contain between 8 and 12 segments thus defined. The distribution of stories and segments across sources appears in Table 2.

Chinese Source	**Type**	**Stories**	**Segments**	**Segments /Story**
Xinhua	Newswire	70	546	7.80
Zaobao	News Web Pages	30	332	11.07
Arabic Source				
AFP	Newswire	50	376	7.52
Xinhua	Newswire	50	352	7.04

Each segment of each story is translated from the source language into English by multiple human translation teams and commercially available translation systems and research systems. Table 3 shows segments that are assisted under this effort by source and number of translations by each type of translator.

	Input Segments	Human Teams	Research Sites	Commercial Systems	Total Segments
AFP (Ara)	376	0	3	3	2256
Xinhua (Ara)	352	0	3	3	2112
Xinhua (Ch)	546	0	8	2	5460
Zaobao (Ch)	332	0	8	2	3320
Total Segments					13148
Decisions/Judge (20% sample)					2630
Hours/Judge (assuming 30 seconds/segment)					22

3. Method Overview

A team of human judges provide multiple assessments of adequacy and fluency for each sampled segment of each translation of each story. For *adequacy* assessments, judges compare each segment to a *reference translation*. A bilingual linguist and senior annotator chooses the best of the human translations to serve as the gold-standard(s). *Fluency* is assessed with respect to the grammar of Standard Written English and required no comparison. Judges view each translated sentence only once giving fluency and adequacy assessments in a single pass. Assessment is timed and judges are strongly encouraged to work as quickly as comfortably possible. Assessors are strongly encouraged to provide their intuitive reaction to each segment and strongly discouraged from pondering their decisions.

4. Sampling

Each judge reviews all segments of an equal-sized subset of translanted stories. Translations are assigned uniformly across judges. Each story is seen by at least two different judges. Each judge sees an even distribution of reference translations.

The following procedure is used to ensure a uniform distribution of judges across systems and documents while maintaining a random choice of judges, systems and documents.

```
First:
    Create an urn containing one token for each judge.
    Create a second urn containing one token for each translation system.
    Create a third urn containing one token for each translation.
    Create a fourth urn containing one token for each reference translation
Then, until all translations are chosen:
    Pick a system from the 2nd urn; if the urn is empty, refill it
    Pick a translation from the 3rd urn
    Pick two judge tokens from the 1st urn; if the urn is empty, refill it
    Assign judges to the chosen translation taking care that the two judges are
    different
    Pick two Reference Translations from the 4th urn, if the urn is empty, refill it;
    Assign
```

5. Judges

Judges are native speakers of English with at least some university level education and have been trained on an assessment interface designed specifically for this task. Judges are instructed to spend, on average, no more than 30 seconds assessing both the fluency and adequacy of a segment. Judges are further instructed to provide their intuitive assessments of fluency and adequacy and not to delay assessment by pondering their decisions.

6. Order of Presentation

Each judge assesses all translations of all segments from an equal subset of the stories. Judges assess the segments of a story in the order in which the segments appear in the story. However, the order of presentation of translation of stories is random. Specifically, judges do not see all translations of a story in sequence order nor do they see all translations by a single translator in order. All ordering is random except the ordering of segments within a story. Segments are presented in their original order to preserve the continuity within a story.

The 100 Arabic and 100 Chinese Stories will be divided into four groups. The stories will be presented to the judges in groups of fifty. This is accomplished by selecting 50 documents at random from each of the Arabic and Chinese stories. This will be group one. The bottom fifty will be group two. This is repeated for each language. The judges will first see 50 Chinese stories, then 50 Arabic stories, then 50 Chinese stories and then 50 Arabic stories.

7. Fluency Assessment

For each translation of each segment of each selected story, judges make the fluency judgment before the adequacy judgment. Fluency refers to the degree to which the target is well formed according to the rules of Standard Written English. A fluent segment is one that is well-formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker of English. A fluency judgment is one of the following:

	How do you judge the fluency of this translation? It is:
5	Flawless English
4	Good English
3	Non-native English
2	Disfluent English
1	Incomprehensible

Where English translations retain source language characters or words, judges are instructed to give a score between “1: Incomprehensible” and “3: Non-native English” depending upon the degree to which the un-translated characters, among the other factors, affect the fluency of the translation.

8. Adequacy Assessment

Having made the fluency judgment for a translation of a segment, the judge is presented with one of four reference translations. Comparing the target translation against the reference translation, judges determine whether the translation is adequate. Adequacy refers to the degree to which information present in the original is also communicated in the translation. Thus for adequacy judgments, the reference translation will serve as a proxy for the original source-language text. An adequacy judgment is one of the following:

	How much of the meaning expressed in the gold-standard translation is also expressed in the target translation?
5	All
4	Most
3	Much
2	Little
1	None

Where English translations retain Chinese and or Arabic characters from the original news stories, judges are instructed to give a score between “1: None” and “4: Most” depending upon the degree to which the un-translated characters, among the other factors, affect the adequacy of the translation.

9. Input File Format

The inputs to the translation assessment process are the multiple translations of the 100 Chinese and 100 Arabic news stories described in section 2. They have the following format.

```
<doc doc_id="official_docno" sys_id="sytem_name">
<hl>
<seg id="1"> Headline text when present </segment>
</hl>

<p>
<seg id="2"> Here is the first segment of paragraph 1. </segment>
<seg id="3"> This first paragraph has two segments. </segment>
</p>

<p>
<seg id="4"> Here is the first segment of paragraph 2. </segment>
<seg id="5"> [and so on...]
</p>
…
</doc>
```

In the original, source-language news stories the "system_name" is "source".

10. Output Format

The output file contains one record per assessment. The assessment record will have the following form.

<

Doc_ID =	official document number
Sys_ID =	unique identifier of translation system in input file
Seg_ID =	unique identifier of segment in input file
Judge_ID =	unique identifier (login name) of judge
RefTransID =	unique identifier of translation system used as Reference Translation file
Fluency =	integer from 1-5 containing fluency judgment
Adequacy =	integer from 1-5 containing adequacy judgment
Comments =	string containing comments entered by judge
Date_Time =	date and time of judgment

>

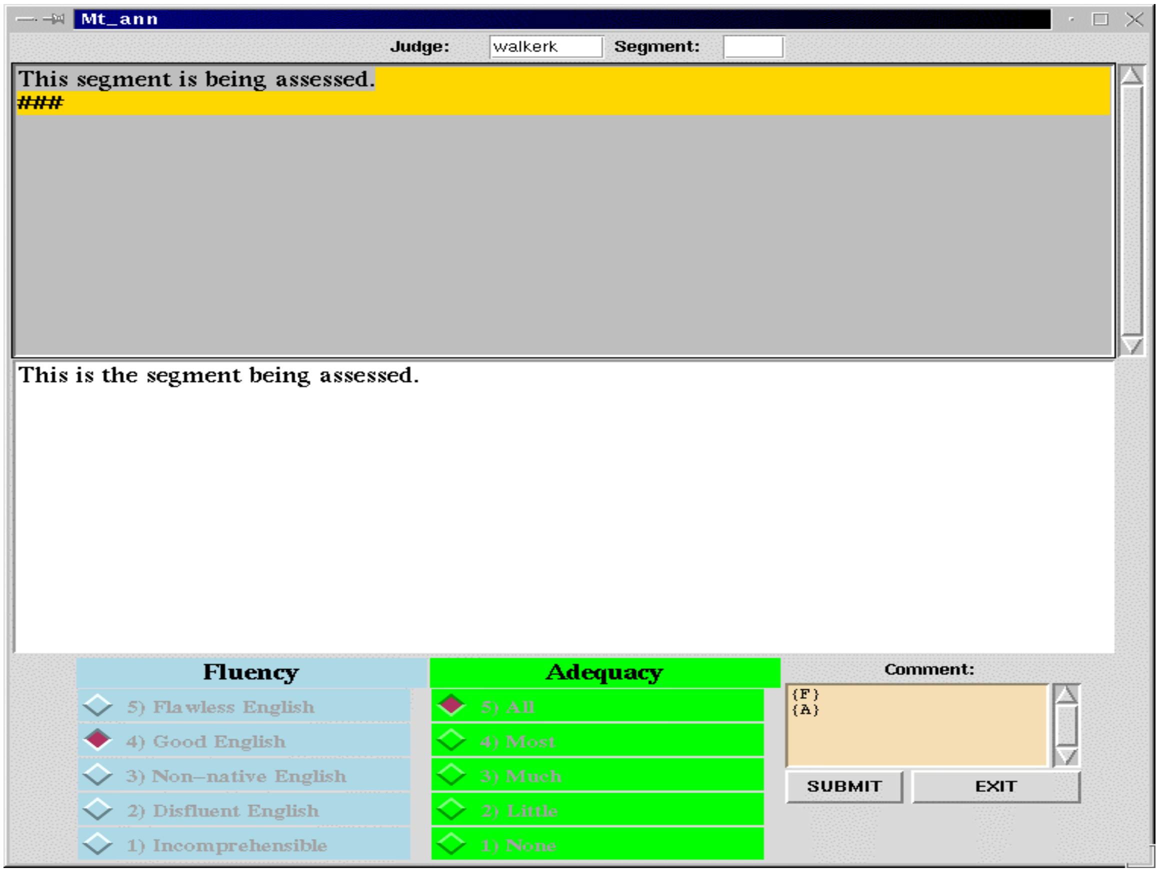
11. Assessment System

The “assessment system” is defined here as the collection of utilities, computer programs and graphical user interfaces that prepare the output of the human translation teams for assessment, assign translations to individual human judges, display segments of the translations, collect human judgments on them and output the human judgments in the output format specific above.

The assessment system accepts input in the form specified in Section 9 and delivers output as specified in Section 10. The assessment system distributes translations of the original news stories such that each reference translation is used proportionally across and such that two independent judges assess each translation of each story. The assessment system presents segments within a story in their naturally occurring order but otherwise provides all translation of all stories in a random order. The assessment system ensures that stories and translations of stories are distributed randomly across judges. Specifically, except as may occur in a random sampling, the assessment system does not assign any one judge a disproportionate percentage of either translations of one original story or of translation by a single translator.

The assessment system graphical user interface presents all segments of a selected translation in the order in which the segments appeared in the original news story. For each selection, the assessment system graphical user interface first presents the segment alone and acquires a fluency judgment. The interface then displays the corresponding gold-standard segment and acquires an adequacy judgment before progressing to the next segment. The assessment system graphical user interface does not display the gold-standard segment while the judge is making the fluency assessment.

Figure 1: Assessment System Graphical User Interface



Judge: walkerk Segment:

This segment is being assessed.
###

This is the segment being assessed.

Fluency

- ○ 5) Flawless English
- ● 4) Good English
- ○ 3) Non–native English
- ○ 2) Disfluent English
- ○ 1) Incomprehensible

Adequacy

- ● 5) All
- ○ 4) Most
- ○ 3) Much
- ○ 2) Little
- ○ 1) None

Comment:

{F}
{A}

Appendix A: Inputs files to the translation process

Xinhua Chinese	Zaobao Chinese	AFP Arabic	Xinhua Arabic
XIN20020316.0014.sgm	ZBN20020316.0001.sgm	artb_001.sgm	artb_500.sgm
XIN20020316.0092.sgm	ZBN20020316.0002.sgm	artb_002.sgm	artb_501.sgm
XIN20020317.0076.sgm	ZBN20020316.0003.sgm	artb_003.sgm	artb_502.sgm
XIN20020317.0152.sgm	ZBN20020317.0001.sgm	artb_004.sgm	artb_503.sgm
XIN20020318.0139.sgm	ZBN20020317.0003.sgm	artb_005.sgm	artb_504.sgm
XIN20020318.0154.sgm	ZBN20020318.0001.sgm	artb_006.sgm	artb_505.sgm
XIN20020319.0197.sgm	ZBN20020318.0002.sgm	artb_007.sgm	artb_506.sgm
XIN20020319.0205.sgm	ZBN20020318.0003.sgm	artb_008.sgm	artb_507.sgm
XIN20020320.0128.sgm	ZBN20020318.0004.sgm	artb_009.sgm	artb_508.sgm
XIN20020321.0027.sgm	ZBN20020318.0005.sgm	artb_010.sgm	artb_509.sgm
XIN20020321.0224.sgm	ZBN20020319.0001.sgm	artb_011.sgm	artb_510.sgm
XIN20020322.0066.sgm	ZBN20020319.0002.sgm	artb_012.sgm	artb_511.sgm
XIN20020322.0179.sgm	ZBN20020319.0003.sgm	artb_013.sgm	artb_512.sgm
XIN20020323.0163.sgm	ZBN20020319.0004.sgm	artb_014.sgm	artb_513.sgm
XIN20020324.0143.sgm	ZBN20020319.0005.sgm	artb_015.sgm	artb_514.sgm
XIN20020324.0145.sgm	ZBN20020319.0006.sgm	artb_016.sgm	artb_515.sgm
XIN20020325.0242.sgm	ZBN20020320.0001.sgm	artb_017.sgm	artb_516.sgm
XIN20020326.0110.sgm	ZBN20020320.0003.sgm	artb_018.sgm	artb_517.sgm
XIN20020326.0188.sgm	ZBN20020320.0004.sgm	artb_019.sgm	artb_518.sgm
XIN20020327.0070.sgm	ZBN20020321.0001.sgm	artb_020.sgm	artb_519.sgm
XIN20020327.0092.sgm	ZBN20020321.0002.sgm	artb_021.sgm	artb_520.sgm
XIN20020328.0091.sgm	ZBN20020321.0003.sgm	artb_022.sgm	artb_521.sgm
XIN20020328.0167.sgm	ZBN20020321.0004.sgm	artb_023.sgm	artb_522.sgm
XIN20020329.0043.sgm	ZBN20020321.0005.sgm	artb_024.sgm	artb_523.sgm
XIN20020329.0061.sgm	ZBN20020321.0006.sgm	artb_025.sgm	artb_524.sgm
XIN20020330.0063.sgm	ZBN20020322.0002.sgm	artb_026.sgm	artb_525.sgm
XIN20020330.0095.sgm	ZBN20020322.0003.sgm	artb_027.sgm	artb_526.sgm
XIN20020331.0102.sgm	ZBN20020322.0004.sgm	artb_028.sgm	artb_527.sgm
XIN20020401.0067.sgm	ZBN20020322.0005.sgm	artb_029.sgm	artb_528.sgm
XIN20020401.0085.sgm	ZBN20020322.0006.sgm	artb_030.sgm	artb_529.sgm
XIN20020402.0114.sgm		artb_031.sgm	artb_530.sgm
XIN20020402.0173.sgm		artb_032.sgm	artb_531.sgm
XIN20020403.0039.sgm		artb_033.sgm	artb_532.sgm
XIN20020403.0180.sgm		artb_034.sgm	artb_533.sgm
XIN20020404.0193.sgm		artb_035.sgm	artb_534.sgm
XIN20020404.0247.sgm		artb_036.sgm	artb_535.sgm
XIN20020405.0053.sgm		artb_037.sgm	artb_536.sgm
XIN20020405.0176.sgm		artb_038.sgm	artb_537.sgm
XIN20020406.0054.sgm		artb_039.sgm	artb_538.sgm
XIN20020406.0075.sgm		artb_040.sgm	artb_539.sgm
XIN20020407.0048.sgm		artb_041.sgm	artb_540.sgm
XIN20020407.0156.sgm		artb_042.sgm	artb_541.sgm
XIN20020408.0093.sgm		artb_043.sgm	artb_542.sgm
XIN20020408.0221.sgm		artb_044.sgm	artb_543.sgm
XIN20020409.0212.sgm		artb_045.sgm	artb_544.sgm
XIN20020409.0230.sgm		artb_046.sgm	artb_545.sgm
XIN20020410.0043.sgm		artb_047.sgm	artb_546.sgm
XIN20020411.0002.sgm		artb_048.sgm	artb_547.sgm
XIN20020411.0233.sgm		artb_049.sgm	artb_548.sgm

XIN20020412.0061.sgm	artb_050.sgm	artb_549.sgm
XIN20020412.0182.sgm	artb_051.sgm	artb_550.sgm
XIN20020413.0067.sgm	artb_052.sgm	artb_551.sgm
XIN20020413.0112.sgm	artb_053.sgm	artb_552.sgm
XIN20020413.0126.sgm	artb_054.sgm	artb_553.sgm
XIN20020414.0057.sgm	artb_055.sgm	artb_554.sgm
XIN20020415.0177.sgm	artb_056.sgm	artb_555.sgm
XIN20020415.0218.sgm	artb_057.sgm	artb_556.sgm
XIN20020416.0081.sgm	artb_058.sgm	artb_557.sgm
XIN20020416.0102.sgm	artb_059.sgm	artb_558.sgm
XIN20020417.0227.sgm	artb_060.sgm	artb_559.sgm
XIN20020417.0269.sgm	artb_061.sgm	artb_560.sgm
XIN20020418.0095.sgm	artb_062.sgm	artb_561.sgm
XIN20020419.0094.sgm	artb_063.sgm	artb_562.sgm
XIN20020419.0134.sgm	artb_064.sgm	artb_563.sgm
XIN20020420.0035.sgm	artb_065.sgm	artb_564.sgm
XIN20020420.0083.sgm	artb_066.sgm	artb_565.sgm
XIN20020421.0044.sgm	artb_067.sgm	
XIN20020421.0106.sgm	artb_068.sgm	
XIN20020422.0172.sgm	artb_069.sgm	
XIN20020422.0173.sgm	artb_S01.sgm	
	artb_S02.sgm	
	artb_S03.sgm	
	artb_S04.sgm	
	artb_S05.sgm	
	artb_S06.sgm	