ACCURAT

Analysis and Evaluation of Comparable Corpora
for Under-Resourced Areas of Machine Translation

www.accurat-project.eu

Project no. 248347

Deliverable D4.5

Report on impact of data extracted from aligned comparable corpora on quality of MT

Version No. 1.0

29/06/2012

Document Information

Deliverable number:	D4.5
Deliverable title:	Report on impact of data extracted from aligned comparable corpora on quality of MT
Due date of deliverable:	31/03/2012 postponed to 30/06/2012
Actual submission date of deliverable:	29/06/2012
Main Author(s):	Bogdan Babych, Richard Forsyth, Fangzhong Su
Participants:	CTS
Internal reviewer:	Tilde
Workpackage:	WP4
Workpackage title:	Comparable corpora in MT systems
Workpackage leader:	DFKI
Dissemination Level:	PU
Version:	V2.0
Keywords:	evaluation, quality

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/ Approval Level
V0.1	28/06/2012	First draft	Bogdan Babych, Richard Forsyth, Fangzhong Su (CTS)	Initial contributions from partners	Submitted for internal review
V1.0	29/06/2012	Final	Tilde	Internal review, Reformatting	Submitted to PO

EXECUTIVE SUMMARY

This deliverable describes human evaluation experiment compared with the results of automated evaluation. Evaluated MT systems are built using data extracted from comparable corpora. We compare them with the baseline systems built using traditional parallel data.

Table of Contents

Abbreviations

Table 1. Abbreviations

Abbreviation	Term/definition
MT	Machine Translation
SMT	Statistical Machine Translation
CC	Comparable corpus
BLEU	Bilingual Evaluation Understudy (MT evaluation method)

1. Introduction

The goal of the ACCURAT project is to find and develop novel methods that exploit comparable corpora to compensate for the shortage of linguistic resources and to improve MT quality for under-resourced languages and narrow domains. WP4 measures improvements from applying the acquired data against results from the baseline SMT systems.

1.1 Comparing MT systems: baseline vs CC-enhanced

Baseline SMT systems were trained using MOSES toolkit on parallel data only. These systems are compared to systems enhanced with resources derived from comparable corpora, added to the baseline training set (CC-enhanced systems). We measured the differences in translation quality achieved with this addition of the data derived from comparable corpora.

1.2 Human evaluation scenario

We designed specific evaluation scenario for measuring the differences between baseline and CC-enhanced MT. This scenario was based on the following considerations:

1. The point of the evaluation is to characterise the changes relative to the baseline data. The absolute level of quality of compared MT systems is less important. This level was achieved with the parallel and comparable data collected within the limited time of the project, and is below the state-of-the-art systems built on large datasets. Instead we focussed on the question what difference in translation quality can be expected when translation models are trained on CC-enhanced data in addition to the baseline training sets. Higher absolute levels of quality can be achieved when larger datasets are used; but these are not the main focus of the evaluation experiment. Therefore we report relative differences in translation quality figures over the baseline as our main finding in the experiment.
2. The amount of changes in CC-enhanced system is limited, since the same baseline data is also included in its training set. The changes that occur due to addition of CC data normally cause only a few lexical or word order differences in the CC-enhanced system output. The majority of evaluated sentences are different, but these differences only concern several words per sentence: there are no cases where the same sentence received a completely different translation compared to the baseline. Therefore these differences are observable by human evaluators and can be presented to them and judged independently of the overall quality of the translated sentences.
3. The evaluators should compare changes in both MT systems, but they should also give their judgement about the absolute quality of each of the compared items, which would produce some numerical values and the possibility to correlate them with automated evaluation scores generated with BLEU
4. The evaluators should judge the overall translation quality of the compared segments, but also the quality of each of the lexical translation choices in cases where the baseline and enhanced translations are different.

The following scenario was designed to address the issues mentioned above:

- For each target language we recruited at least 3 evaluators, most of them had background in translation (either professional translators, or translation students, or linguists), who rated 120 sentences each. We obtained at least 3 independent scores for each of the compared sentences and lexical differences.

- Evaluators were asked to read the gold-standard translation of the segment, then to read each of the two compared translations: the baseline and the CC-enhanced. (The systems were anonymized and the order of presentation was random for each sentence). After that evaluators were asked to judge *overall translation quality* of the compared sentences.
- Finally, evaluators were asked to look into highlighted differences and rate the *quality of translation choices* of these highlighted words individually.

As we expected, there are considerable differences between evaluation on the basis of overall translation quality of the segment, and the quality of translation choices of individual words and phrases: the later is generally higher. It is important to evaluate these individual differences since this allows us to exactly measure contribution of the CC-based data to translation quality.

1.3 Evaluation plan

Evaluation was performed according to the following steps:

1. System output was generated for the baseline and CC-enhanced MT systems for the following translation directions and domains

   De-en; ro-en; sl-en; hr-en; ro-de; lv-lt; en-lv; en-hr; en-el; de-ro; el-ro for the broader News domain

   De-en and *en-lv* – for Automotive domain

2. Evaluation set of 511 sentences (circa 11000 words) was used for all translation directions.
3. BLEU scores were generated for each of the evaluated systems
4. Evaluation packs for human evaluation were constructed using the following procedure
   a. Sentences different in the baseline vs. CC-enhanced output were identified
   b. Words different in the baseline vs. CC-enhanced output were automatically highlighted; if several consecutive words were highlighted all of them were evaluated together as a phrase.
   c. The order of presentation of the CC-enhanced vs baseline systems was randomised
   d. Evaluation packs were presented to evaluators within a web interface that automatically calculated submitted evaluation results (using CGI script). All evaluation packs are available now on:

      http://corpus.leeds.ac.uk/accurat2012/eval/

   e. The set of 120 sentences (those were the first 120 non-similar sentences out of the complete set of 511 sentences used for calculating BLEU scores) were typically used for human evaluation experiment, with at least 3 independent judgements collected for each sentence, and also – for each highlighted word or phrase that was different in the baseline vs. CC-enhanced translation.
5. Evaluators were recruited to rate the differences in the system
6. Evaluation results were collected
7. Scores were analysed and reported

1.4 Evaluation set-up

Evaluators were asked to evaluate translation quality of the compared sentences and the quality of translation choices of the highlighted words or phrases.
Instructions given to evaluators are given in Appendix 1.
The following table gives the number of evaluators recruited for rating each system.

Table 2 Numbers of evaluators and evaluated sentences per system / domain

SL	TL	No of evaluated sentences	No of evaluators
News			
de	en	120	3
ro	en	120	3
sl	en	120	3
hr	en	120	3
ro	de	120	2
lv	lt	120	5
en	lv	120	4
en	hr	360	6
en	el	120	3
de	ro	120	3
el	ro	120	3
Automotive			
de	en	120	3
en	lv	120	5
		Total	46

Words, which are different in the baseline vs CC-enhanced systems, are presented in bold italic font. This highlighting for different words was done automatically. Evaluation interface is presented on the following screenshot (Figure 1). Drop-boxes offer evaluators to choose one of the possible evaluation scores: 1 (lowest quality), 2, 3, 4 or 5 (highest quality).

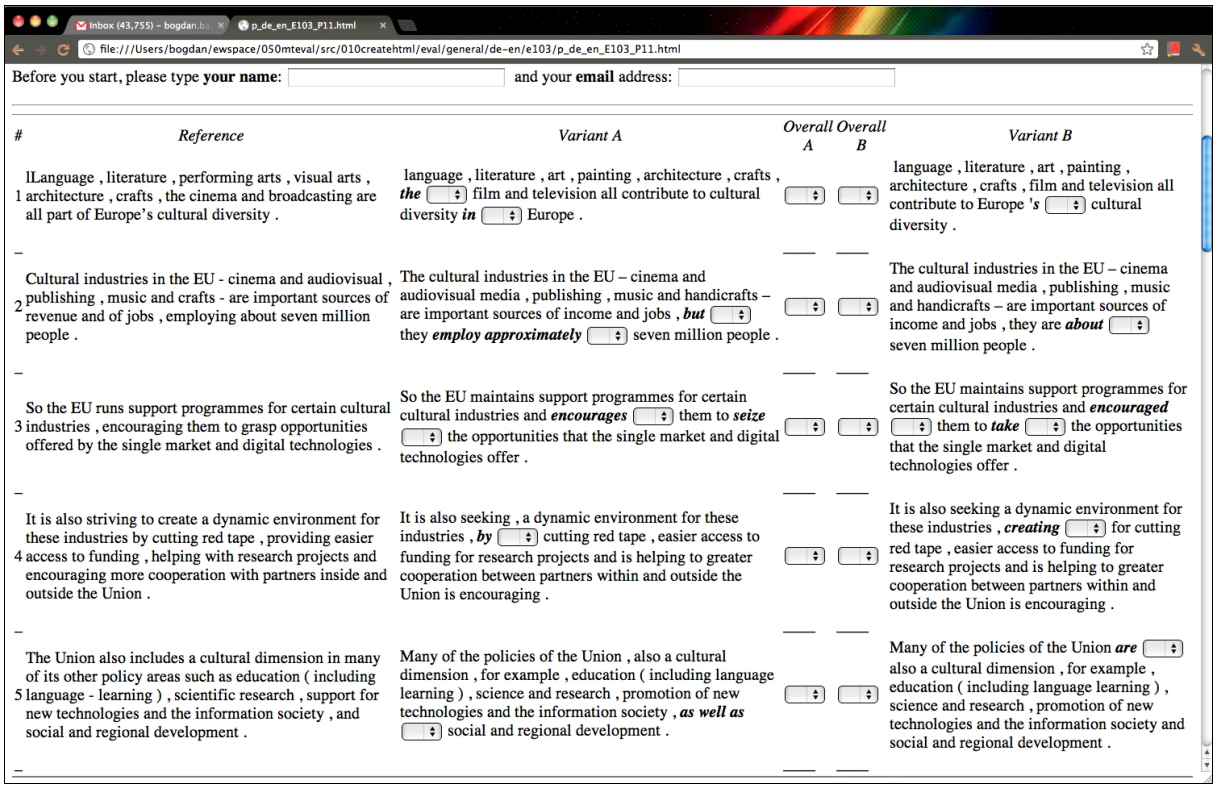

Figure 1 Evaluation interface

The interface and instructions are designed to be intuitive enough for evaluators without any technical background to follow up, the example can be found under the following URL: http://corpus.leeds.ac.uk/accurat2012/eval/automotive/de-en/e671/n_de_en_E671_P11.html

The interface was developed by CTS/Leeds team within ACCURAT project, specifically to measure improvements of CC-enhanced MT systems. However, this evaluation framework has potentially broader applications, in particular in cases where evaluation results need to highlight specific differences and improvements to MT systems which are less visible for standard evaluation methods, such as Adequacy and Fluency evaluation.

The source code for generating evaluation packs and the data used for ACCURAT evaluation are available from:

http://corpus.leeds.ac.uk/accurat2012/eval/code-and-data/

Raw evaluation results submitted by 46 evaluators are available at:

http://corpus.leeds.ac.uk/accurat2012/eval/results/

The evaluation packs, generation system and results are all open-source.

2. Evaluation results

Evaluation results are presented in two groups: overall evaluation results (Table 3) and evaluation results for lexical differences (Table 4).

Table 3 Evaluation results for overall translation quality

SL	TL	Baseline	CC-enhanced	Human scores for improvement
News				
de	en	2.269	2.1	-7.45%
ro	**en**	**1.826**	**2.721**	**49.01%**
sl	**en**	**1.869**	**2.025**	**8.35%**
hr	en	2.175	2.199	1.10%
ro	**de**	**1.692**	**1.846**	**9.10%**
lv	lt	2.157	2.095	-2.87%
en	lv	2.04	1.993	-2.30%
en	hr	2.107	1.864	-11.53%
en	el	2.212	2.362	6.78%
de	ro	1.942	1.914	-1.44%
el	ro	2.156	2.271	5.33%
Average				**4.92%**
Automotive				
de	en	2.201	2.893	31.44%
en	lv	2.177	2.5	14.84%
Narrow domain average				**23.14%**

Table 4 Evaluation results for lexical differences: baseline vs CC-enhanced MT

SL	TL	Baseline	CC-enhanced	Human scores for improvement
News				
de	en	2.773	2.774	0.04%
ro	**en**	**1.819**	**3.377**	**85.65%**
sl	en	2.642	2.867	8.52%
hr	**en**	**2.66**	**2.905**	**9.21%**
ro	de	2.351	2.376	1.06%
lv	lt	2.614	2.587	-1.03%
en	lv	2.564	2.618	2.11%
en	hr	2.507	2.118	-15.52%
en	el	3.026	3.271	8.10%
de	ro	2.399	2.365	-1.42%
el	**ro**	**2.757**	**3.458**	**25.43%**

SL	TL	Baseline	CC-enhanced	Human scores for improvement
Average				**11.10%**
Automotive				
de	en	2.628	3.835	45.93%
en	lv	2.604	2.956	13.52%
Narrow domain average				**29.72%**

The scores presented in the tables were computed as an *average* of evaluation scores of all the evaluators for all 120 sentences in each evaluation pack:

$$Average = \frac{(\sum Scores)}{(nSent \times nEvaluators)},$$

where the *Average* is the reported scores, *nSent* is the number of evaluated sentences (120) and *nEvaluators* is the number of evaluators (typically 3 per system).

3. Analysis of the results

3.1 Comparison with automated evaluation

Table 5 presents the comparison between human and automated BLEU scores.

Table 5 Human and automated evaluation scores

Overall sentence-level evaluation							
Human evaluation scores					**Automated BLEU scores**		
SL	**TL**	**Baseline**	**CC-enhanced**	**Human improvement**	**BLEU-base**	**BLEU-CC-enh**	**BLEU imprv.**
de	en	2.269	2.1	-7.45%	27.9	28.62	2.58%
ro	**en**	**1.826**	**2.721**	**49.01%**	21.54	30.35	40.90%
sl	**en**	**1.869**	**2.025**	**8.35%**	26.28	27.46	4.49%
hr	en	2.175	2.199	1.10%	20.78	21.91	5.44%
ro	**de**	**1.692**	**1.846**	**9.10%**	10.22	11.21	9.69%
lv	lt	2.157	2.095	-2.87%	12.12	12.69	4.70%
en	lv	2.04	1.993	-2.30%	12.74	13.25	4.00%
en	hr	2.107	1.864	-11.53%	10.91	11.45	4.95%
en	el	2.212	2.362	6.78%	19.06	23.67	24.19%
de	ro	1.942	1.914	-1.44%	9.66	10.14	4.97%
el	ro	2.156	2.271	5.33%	15.81	17.25	9.11%
					r correlation: BLEU vs Human		
		Average =		4.92%	0.2309	0.6929	**0.8922**
de	en	2.201	2.893	31.44%			
en	lv	2.177	2.5	14.84%			
		Narrow domain average		23.14%			
Word-level evaluation							
Human evaluation scores					**Automated BLEU scores**		
de	en	2.773	2.774	0.04%	27.9	28.62	2.58%
ro	**en**	**1.819**	**3.377**	**85.65%**	21.54	30.35	40.90%
sl	en	2.642	2.867	8.52%	26.28	27.46	4.49%
hr	**en**	**2.66**	**2.905**	**9.21%**	20.78	21.91	5.44%
ro	de	2.351	2.376	1.06%	10.22	11.21	9.69%
lv	lt	2.614	2.587	-1.03%	12.12	12.69	4.70%
en	lv	2.564	2.618	2.11%	12.74	13.25	4.00%
en	hr	2.507	2.118	-15.52%	10.91	11.45	4.95%
en	el	3.026	3.271	8.10%	19.06	23.67	24.19%
de	ro	2.399	2.365	-1.42%	9.66	10.14	4.97%
el	**ro**	**2.757**	**3.458**	**25.43%**	15.81	17.25	9.11%
					r correlation: BLEU vs Human		
		Average =		11.10%	0.1732	0.6784	**0.8550**
de	en	2.628	3.835	45.93%			
en	lv	2.604	2.956	13.52%			
		Narrow domain average =		29.72%			

Table 5 compares the average values of human scores presented in Table 3 and Table 4 with BLEU evaluation figures generated for the same dataset that was used for human evaluation.

For measuring agreement between BLEU results and human evaluation scores we computed Pearson's correlation coefficient *r* between the two ranges, calculated as:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

This value was calculated between the following ranges (see Table 4):

- Baseline and BLEU-base
- CC-enhanced and BLEU-CC-enh
- Human improvement and BLEU improvement

For the first two ranges, since BLEU scores for different target languages were present in these ranges, there was no strong correlation between them, as we expected[1].

The reason is that absolute values of BLEU change considerably across different target languages because of different type/token ratios in them, and in general are not comparable.

It can be seen from the table that raw BLEU do not correlate with the range of raw scores either in the case of the baseline, or the CC-enhanced systems.

However, for the third range there is a strong positive correlation ***r=0.89*** – for the overall quality and ***r=0.85*** for word-level improvements between the relative improvement in terms of BLEU scores and relative improvement in terms of human evaluation scores.

For the system improvement figures BLEU can predict the amount of improvements in terms of human scores, but only on the larger scale, e.g., if the improvement is around 10% or more. In case of improvements of around 5% or less the BLEU cannot capture the differences in translation quality and there are serious mismatches in terms of human and automated evaluation.

3.2 Features that generate statistically significant splits of evaluation data

Taking human evaluation scores as a response variable, we generated the graph of which features split all human evaluation results on using wilcoxon test, generated by the *ctree* function in R statistical package. The diagram shows significant splits of the data on different levels by different features (such as whether evaluation is done on the word level vs. overall sentence level ('w' or 'o' -- the top-most split in Figure 2), what was the Source and Target language, whether we evaluated the baseline or the CC-enhanced system, etc.) in the evaluation experiment. Figure 2 shows these splits:

[1] For details and experimental results of dependence of BLEU on target language see:

Bogdan Babych, Anthony Hartley & Debbie Elliott (2005) Estimating the predictive power of n-gram MT evaluation metrics across language and text types . MT Summit X, Phuket, Thailand, September 13-15, 2005, Conference Proceedings: the tenth Machine Translation Summit; pp.412-418., available at http://www.mt-archive.info/MTS-2005-Babych.pdf

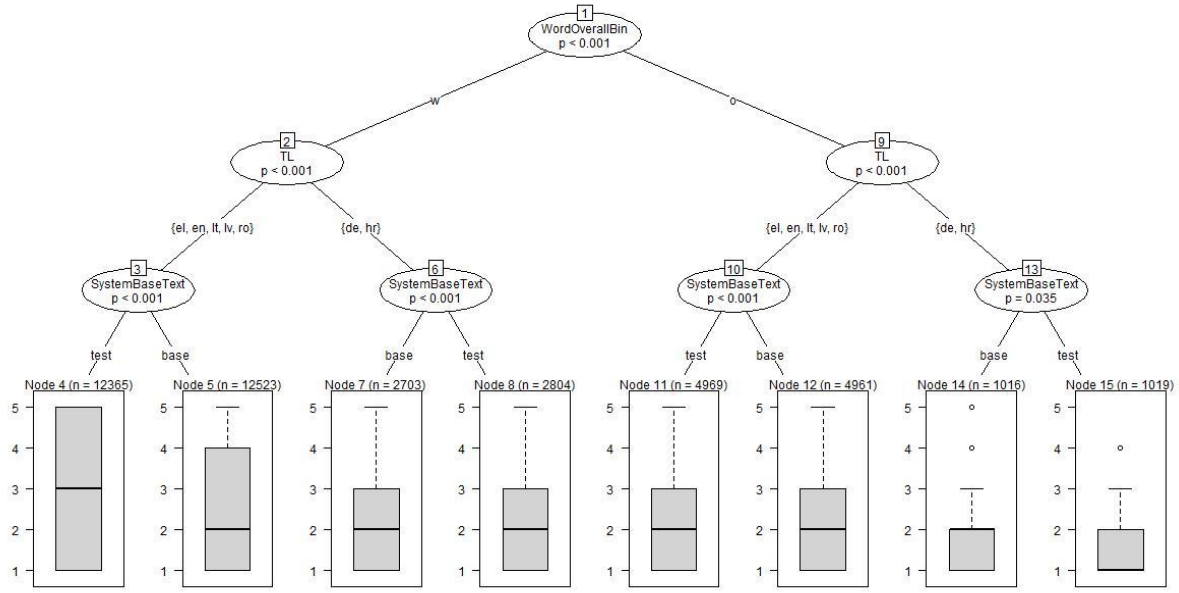

Figure 2 Significant splits of evaluation data.

It can be seen from the figure that the main split in evaluation results is between the lexical (word) level ('w') and overall quality ('o') evaluation: scores are in general higher for word-level evaluation.

At the second level the data are split by the target language: there are different ranges of scores for each of the target languages.

Further down the data is split by whether the baseline (base) or the CC-enhanced (test) systems were evaluated. Manipulation of this parameter is significant at the level of $p<0.001$

This test shows how the data obtained from our experiment can be automatically structured according to the most significant differences between features. The test tells us the following:

1. The most significant difference in the experiment is the difference between lexical word based evaluation vs. traditional overall sentence-based evaluation methodology, which confirms our assumption that for evaluating improvements for ACCURAT methodology we needed to introduce lexical evaluation: it gives very important additional information about improvements.
2. The second most significant difference is between different target languages: normally the degree of improvement depends on morphological complexity of the target language
3. There are statistically significant improvements within the CC-enhanced systems for a larger group of target languages, such as en, el, lt, lv, ro, which proves the point that the methodology of enhancing MT with the data derived from comparable corpora has the potential to improve translation quality, if the morphological complexity of the target language is not too high.

4. Conclusions

We can draw the following conclusion from the presented results:

1. The overall baseline translation quality is very low - 1 or 2 on the 5-point translation quality scale on average. The quality of lexical translation choices is higher and greater improvement for it is achieved for CC-enhanced systems. Therefore our evaluation methodology of focussing on lexical differences is more appropriate to the task of measuring improvements with CC-based data.
2. On average across all translation directions there is improvement in all four areas: for *Overall translation quality* in *Broader domain* the improvement was the smallest: 4.92% over the baseline; then *Lexical* improvement in *Broader* domain was 11.1% on average.
3. In *narrow domain* there is a much higher and consistent improvement for both evaluated systems and in both aspects: *overall* and *lexical* quality, compared for the broad domain. The improvement for *narrow* domains was: 23.14% and 29.72% for *Overall* and *Lexical* translation quality respectively.
4. For the broad news domain improvement or deterioration depends on the translation direction. Translation into English is always improved. All cases of degradation are for translation into more morphologically complex languages, such as Croatian. The mechanism for this fact is not known, and requires further investigation.

The results point out that the biggest benefit of CC-enhanced data is achieved for narrow domains and for MT into morphologically simpler languages like English.

Appendix 1. Evaluation instructions and examples of evaluated sentences

The following instructions were given to evaluators:

> “”” You will be rating different translation variants of the same text.
>
> For each pair of sentences:
>
> Read carefully the professional reference translation in the left column. Then read the two translation variants (Variant A and Variant B).
> For each variant:
>
> A) On the scale 1 to 5, please rate their **overall translation quality**.
> Please use the following scale:
> **1=Translation is not at all good ... 5=Translation is very good**.
> Select your scores from drop-down menus in two central columns of the tables (Overall A and Overall B).
>
> B) Then look into highlighted differences in each variant.
> On the scale 1 to 5, please rate **translation quality** of the highlighted words or phrases (**1=Very bad translation choice... 5=Very good translation choice**)
> Select your scores from the drop-down menus next to the highlighted words.
>
> After selecting all scores on the page please click 'Submit' button at the bottom. “”””

Following is an example of sentences which have been evaluated for de-en translation direction

Reference:

The aim of the European Union is double: to preserve and support this diversity and to help make it accessible to others.

Cultural industries in the EU - cinema and audiovisual, publishing, music and crafts - are important sources of revenue and of jobs, employing about seven million people.

…

Baseline SMT

The objective of the EU is a twofold: To preserve and support this diversity and help them to other accessible.

The cultural industries in the EU – cinema and audiovisual media, publishing, music and handicrafts – are important sources of income and jobs, but they employ approximately seven million people.

…

CC-enhanced SMT

The objective of the EU is a twofold: To preserve and support this diversity and help them to other accessible.

The cultural industries in the EU – cinema and audiovisual media, publishing, music and handicrafts – are important sources of income and jobs, they are about seven million people.

Appendix 2. Examples of evaluation results for a translation direction

The following is an example of data for Broad domain for de-en translation direction. The columns represent the sentence number, average evaluation score, standard deviation of scores and the number of evaluators who did the evaluation.

Sentence	Ave	Stdev	NofE		Ave	Stdev	NofE	
s1000	base	2,333	0,577	3	test	3	0	3
s1002	base	3,333	0,577	3	test	2	0	3
s1004	base	5	0	3	test	2,667	0,577	3
s1005	base	2	1	3	test	2	1	3
s1006	base	2,333	0,577	3	test	2	1	3
s1007	base	1,5	0,707	2	test	1,5	0,707	2
s1008	base	4	0	2	test	3	0	2
s1010	base	1	0	2	test	1	0	2
s1012	base	2	0	1	test	2	0	1
s1013	base	4	0	1	test	5	0	1
s1014	base	2,5	0,707	2	test	2	0	2
s1015	base	2,5	0,707	2	test	2,5	0,707	2
s1017	base	2,5	0,707	2	test	3	1,414	2
s1018	base	2,5	0,707	2	test	2,5	0,707	2
s1019	base	2,5	0,707	2	test	2,5	0,707	2
s1020	base	2	1	3	test	2	0	3
s1021	base	2,667	0,577	3	test	2	1	3
s1022	base	2	1	3	test	1,667	1,155	3
s1023	base	1,667	0,577	3	test	2,667	0,577	3
s1024	base	2,333	1,528	3	test	2,333	1,528	3
s1025	base	2,333	0,577	3	test	3,667	0,577	3
s1026	base	2	0	3	test	2,333	0,577	3
s1027	base	3,667	0,577	3	test	2,333	0,577	3
s1028	base	2,5	0,707	2	test	1,5	0,707	2
s1029	base	2,333	1,155	3	test	1,333	0,577	3
s1030	base	1,333	0,577	3	test	1,333	0,577	3
s1032	base	2	1	3	test	2	0	3
s1035	base	2	0	3	test	2,333	0,577	3
s1036	base	3	0	3	test	2	1	3
s1037	base	3	1	3	test	3,333	1,528	3
s1039	base	1,333	0,577	3	test	1,333	0,577	3
s1040	base	2,333	0,577	3	test	1,333	0,577	3
s1041	base	2	1	3	test	2,333	1,155	3
s1042	base	2,667	1,155	3	test	1,333	0,577	3
s1043	base	2,667	1,528	3	test	2,333	1,528	3
s1044	base	3,333	1,528	3	test	2,667	0,577	3
s1046	base	2,667	0,577	3	test	1,333	0,577	3
s1047	base	2,333	1,155	3	test	2,667	0,577	3
s1048	base	2,333	0,577	3	test	1,667	0,577	3
s1049	base	3,333	0,577	3	test	2,333	0,577	3
s1052	base	3	1	3	test	2,333	0,577	3
s1053	base	2	1,414	2	test	2	1,414	2
s1054	base	2,333	1,155	3	test	1,333	0,577	3
s1056	base	2	1	3	test	1,667	1,155	3
s1057	base	1	0	3	test	1,333	0,577	3
s1060	base	3	1	3	test	2,667	0,577	3
s1061	base	2	1	3	test	2,333	1,155	3
s1062	base	2,333	0,577	3	test	2,667	1,155	3

Sentence	Ave	Stdev	NofE		Ave	Stdev	NofE	
s1063	base	1	0	3	test	1,333	0,577	3
s1064	base	1	0	3	test	1,333	0,577	3
s1065	base	2,333	0,577	3	test	2	0	3
s1067	base	1,667	0,577	3	test	1,667	0,577	3
s1068	base	1	0	3	test	1	0	3
s1069	base	1,667	1,155	3	test	1,333	0,577	3
s1071	base	2,333	0,577	3	test	2	1	3
s1072	base	2,333	0,577	3	test	1,667	0,577	3
s1076	base	1,667	0,577	3	test	1,333	0,577	3
s1077	base	2	1	3	test	2	1	3
s1078	base	1,667	0,577	3	test	1,333	0,577	3
s1079	base	2,333	1,155	3	test	2	1	3
s1080	base	1,667	0,577	3	test	1,333	0,577	3
s1081	base	1,667	0,577	3	test	1,667	0,577	3
s1082	base	2,667	1,155	3	test	2	1	3
s1083	base	3,333	0,577	3	test	2,333	1,155	3
s1084	base	1,5	0,707	2	test	2	1,414	2
s1086	base	2	0	3	test	1,667	0,577	3
s1087	base	2	0	3	test	2,333	0,577	3
s1088	base	1,667	0,577	3	test	2	1	3
s1089	base	1,333	0,577	3	test	1	0	3
s1090	base	1	0	3	test	1,333	0,577	3
s1091	base	1	0	3	test	1,333	0,577	3
s1093	base	1,333	0,577	3	test	1,667	1,155	3
s1095	base	2,333	1,155	3	test	2,333	1,155	3
s1096	base	2,667	0,577	3	test	2	1	3
s1097	base	2,667	0,577	3	test	1,333	0,577	3
s1098	base	3,333	1,155	3	test	2,333	0,577	3
s1099	base	2,667	0,577	3	test	2,667	0,577	3
s1100	base	2,333	1,155	3	test	1,667	0,577	3
s1101	base	2,333	1,155	3	test	2,667	1,528	3
s1103	base	2	1	3	test	2,333	1,155	3
s1104	base	2,333	0,577	3	test	1,667	0,577	3
s1105	base	1,333	0,577	3	test	1,333	0,577	3
s1106	base	1,667	1,155	3	test	1,333	0,577	3
s1107	base	2	1	3	test	2	1	3
s1108	base	2	1	3	test	2,667	0,577	3
s1109	base	1,667	0,577	3	test	1,333	0,577	3
s1110	base	1,333	0,577	3	test	1,333	0,577	3
s1111	base	1,333	0,577	3	test	1,333	0,577	3
s1112	base	1	0	3	test	1	0	3
s1113	base	1,667	0,577	3	test	1	0	3
s1114	base	1	0	2	test	1,333	0,577	3
s1115	base	1,667	0,577	3	test	2	1	3
s1116	base	1,667	0,577	3	test	1,667	0,577	3
s1117	base	2	1	3	test	1	0	3
s1118	base	2,333	0,577	3	test	1,333	0,577	3
s1119	base	3	0	3	test	3	0	3
s1120	base	1,667	1,155	3	test	1,333	0,577	3
s1121	base	2	1	3	test	1,667	0,577	3
s1122	base	2	1	3	test	2,333	0,577	3
s1123	base	2,333	0,577	3	test	3	1	3
s1124	base	1,667	0,577	3	test	1,667	0,577	3
s1125	base	3	1,732	3	test	1,667	1,155	3
s1127	base	3,333	2,082	3	test	3	1,732	3
s1128	base	1,667	0,577	3	test	1,333	0,577	3

Sentence	Ave	Stdev	NofE		Ave	Stdev	NofE	
s1129	base	2,667	0,577	3	test	3	1	3
s1130	base	1	0	3	test	1	0	3
s1132	base	3	1	3	test	3,667	0,577	3
s1133	base	3,667	0,577	3	test	3,333	0,577	3
s1134	base	3,667	1,155	3	test	4,333	0,577	3
s1135	base	2,667	0,577	3	test	2	0	3
s1136	base	4,333	1,155	3	test	2,667	0,577	3
s1137	base	4	1,414	2	test	2,5	2,121	2
s1138	base	3	0	3	test	3,667	0,577	3
s1139	base	2,333	0,577	3	test	2,667	0,577	3
s1141	base	3,333	1,155	3	test	3	1	3
s1143	base	2,333	0,577	3	test	3	1	3
s1144	base	2,667	0,577	3	test	3,667	0,577	3
s1146	base	2,333	0,577	3	test	4	1	3
s1147	base	1,667	0,577	3	test	2,333	1,155	3
s1148	base	4	0	3	test	3,667	0,577	3