This extract contains a new benchmark dataset, STSS-131, for evaluating Short Text Semantic Similarity (STSS) measurement algorithms. STSS-131 draws on a range of resources from traditional grammar to cognitive neuroscience. The human ratings are obtained from a set of trials using new and improved experimental methods, with validated measures and statistics.

STSS-131 Short Text Semantic Similarity Benchmark Data Set: Extract from "A new benchmark dataset with production methodology for Short Text Semantic Similarity algorithms"

James D. O'Shea, Zuhair A, Bandar, Keeley A. Crockett

Citation Details

Please do NOT cite as the technical report. Cite this work as the ACM Transactions paper from which it is extracted. The paper is "in press" at the time of writing, to appear in

O'Shea, James, Bandar, Zuhair and Crockett, Keeley, *A new benchmark dataset with production methodology for Short Text Semantic Similarity algorithms*, ACM Transactions on Speech and Language Processing , Volume 10, Number 4.

The page numbers and DOI number should be available after publication (target, December 2013).

The paper itself is 24 pages long and the appendix is 44 pages, in the journal format, and it contains a great deal of supporting information including how to reproduce the method for collecting gold standard STSS datasets. It will also contain a performance evaluation of the STASIS and LSA, which shows the dataset to be more demanding that the original STSS-65 dataset, and therefore suitable for evaluating new and improved STSS algorithms.

When the necessary processes have been completed the full paper with appendices will be available for free download under open access.

Contact for correspondence:

Dr. James D. O'Shea,
Intelligent Systems Group Leader,
School of Computing, Mathematics and Digital Technology,
John Dalton Building,
Chester St.,
Manchester,
M1 5GD
United Kingdom

E-mail: j.d.oshea@mmu.ac.uk
Telephone: +44 161 247 1546

A. THE FULL DATASET

The first column is the number of the sentence pair. The second is the two sentences making up the pair. The third is the semantic similarity rating calculated as the average of the human ratings for the sentence pair (0.00 – 4.00). The final column is the standard deviation of the human ratings, which gives a measure of noisiness. The two faint entries are calibration pairs borrowed from STSS-65. These should NOT be used in calculations and are for reference only. For a fuller understanding of how the data set was collected and the method for using it to compare STSS measures please see *A new benchmark dataset with production methodology for Short Text Semantic Similarity algorithms*, also [O'Shea, et al. 2008] & [O'Shea 2010] which you may also wish to cite.

Table 1. Semantic similarity ratings for STSS-131 (on a scale from 0.00 to 4.00)

SP	Sentences	$\overline{X}$ (Human ratings of semantic similarity)	S
66	Would you like to go out to drink with me tonight? I really don't know what to eat tonight so I might go out somewhere.	1.01	0.77
67	I advise you to treat this matter very seriously as it is vital. You must take this most seriously, it will affect you.	3.38	0.69
68	When I was going out to meet my friends there was a delay at the train station. The train operator announced to the passengers that the train would be delayed.	3.13	0.68
69	Does music help you to relax, or does it distract you too much? Does this sponge look wet or dry to you?	0.1	0.29
70	You must realise that you will definitely be punished if you play with the alarm. He will be harshly punished for setting the fire alarm off.	2.84	0.87
71	I will make you laugh so much that your sides ache. When I tell you this you will split your sides laughing.	3.75	0.38
72	You shouldn't be covering what you really feel. There is no point in covering up what you said, we all know.	2.21	0.97
73	Do you want to come with us to the pub behind the hill? We are going out for drinks tonight in Salford Quays if you would like to come.	1.82	1.09
74	This key doesn't seem to be working, could you give me another? I dislike the word quay, it confuses me, I always think of things for locks, there's another one.	0.72	0.87
75	The ghost appeared from nowhere and frightened the old man. The ghost of Queen Victoria appears to me every night, I don't know why, I don't even like the royals.	1.45	0.75
76	You're not a good friend if you're not prepared to be present when I need you. A good friend always seems to be present when you need them.	3.14	0.94
77	The children crossed the road very safely thanks to the help of the lollipop lady. It was feared that the child might not recover, because he was seriously ill.	0.13	0.29
78	I have invited a variety of people to my party so it should be interesting. A number of invitations were given out to a variety of people inviting them down the pub.	2.18	0.88
79	I offer my condolences to the parents of John Smith, who was unfortunately murdered. I express my sympathy to John Smith's parents following his murder.	3.91	0.23
80	Boats come in all shapes and sizes but they all do the same thing. Chairs can be comfy and not comfy, depending on the chair.	0.5	0.69
81	If you continuously use these products, I guarantee you will look very young. I assure you that, by using these products consistently over a long period of time, you will appear really young.	3.58	0.57
82	We ran farther than the other children that day. You ran farther than anyone today.	2.43	1.06
83	I always like to have a slice of lemon in my drink especially if it's Coke. I like to put a wedge of lemon in my drinks, especially cola.	3.81	0.55
84	It seems like I've got eczema on my ear doctor, can you recommend something for me?	2.05	0.9

	I had to go to a chemist for a special rash cream for my ear.		
85	I am proud of our nation, well, most of it. I think of myself as being part of a nation.	1.71	1.03
86	There was a heap of rubble left by the builders outside my house this morning. Sometimes in a large crowd accidents may happen, which can cause deadly injuries.	0.09	0.27
87	Water freezes at a certain temperature, which is zero degrees Celsius. The temperature of boiling water is 100 C and the temperature of ice is 0 C.	3.08	0.98
88	We got home safely in the end, although it was a long journey. Though it took many hours travel, we finally reached our house safely.	3.06	0.95
89	A man called Dave gave his fiancée a large diamond ring for their engagement. The man presented a diamond to the woman and asked her to marry him.	3.22	0.73
90	I used to run quite a lot, in fact once I ran for North Tyneside. I used to climb lots at school as we had a new climbing wall put in the gym.	0.74	0.75
91	I love to laugh as it makes me happy as well as those around me. I thought we bargained that it would only cost me a pound.	0.08	0.32
92	Because I am the eldest one I should be more responsible. Just because of my age, people shouldn't think I'm a responsible adult, but they do?	2.23	0.79
93	I need to dash into the kitchen because I think my chip pan is on fire. In the event of a chip pan fire follow the instructions on the safety note.	1.7	1.03
94	Peter was a very large youth, whose size intimidated most people, much to his delight. Now I wouldn't say he was fat, but I'd certainly say he was one of the larger boys.	1.96	0.95
95	I'm going to buy a grey jumper today, in half an hour. That's a nice grey top, where did you get it from?	1.25	0.98
96	We got soaked in the rain today, but now we are nice and dry. I was absolutely soaking wet last night, I drove my bike through the worst weather.	1.68	0.75
97	Global warming is what everyone is worrying about today. The problem of global warming is a concern to every country in the world at the moment.	3.14	0.84
98	He was harshly punished for setting the fire alarms off. He delayed his response, in order to create a tense atmosphere.	0.22	0.6
99	Midday is 12 o'clock in the middle of the day. Noon is 12 o'clock in the middle of the day.	3.96	0.16
100	That's not a very good car, on the other hand mine is great. This is a terrible noise level for a new car.	1.05	0.95
101	There was a terrible accident, a pileup, on the M16 today. It was a terrible accident, no one believed it was possible.	2.33	0.93
102	After hours of getting lost we eventually arrived at the hotel. After walking against the strong wind for hours he finally returned home safely.	1.09	0.91
103	The first thing I do in a morning is make myself a cup of coffee. The first thing I do in the morning is have a cup of coffee.	3.85	0.39
104	Someone spilt a drink accidentally on my shirt, so I changed it. It appears to have shrunk, it wasn't that size before I washed it.	0.48	0.72
105	I'm worried most seriously about the presentation, not the essay. It is mostly very difficult to gain full marks in today's exam.	0.77	0.82
106	It is mostly very difficult to gain full marks in today's exam. The exam was really difficult, I've got no idea if I'm going to pass.	2.54	0.98
107	Meet me on the hill behind the church in half an hour. Join me on the hill at the back of the church in thirty minutes time.	3.93	0.25
108	If you don't console with a friend, there is a chance you may hurt their feelings. One of the qualities of a good friend is the ability to console.	3.01	0.85
109	We tried to bargain with him but it made no difference, he still didn't change his mind. I tried bargaining with him, but he just wouldn't listen.	3.43	0.54
110	It gives me great pleasure to announce the winner of this year's beauty pageant. It’s a real pleasure to tell you who has won our annual beauty parade.	3.88	0.24

111	They said they were hoping to go to America on holiday. I like to cover myself up in lots of layers, I don't like the cold.	0.16	0.5
112	Will I have to drive far to get to the nearest petrol station? Is it much farther for me to drive to the next gas station?	3.84	0.37
113	I think I know her from somewhere because she has a familiar face. You have a very familiar face, where do I know you from?	3.36	0.8
114	I am sorry but I can't go out as I have a heap of work to do. I've a heap of things to finish so I can't go out I'm afraid.	3.6	0.72
115	The responsible man felt very guilty when he crashed into the back of someone's car. A slow driver can be annoying even though they are driving safely.	0.88	0.75
116	Get that wet dog off my brand new white sofa. Make that wet hound get off my white couch – I only just bought it.	3.59	0.86
117	He fought in the war in Iraq before being killed in a car crash. The prejudice I suffered whilst on holiday in Iraq was quite alarming.	0.55	0.65
118	The cat was hungry so he went into the back garden to find lunch. The hen walked about in the yard eating tasty grain.	1.2	0.82
119	My bedroom wall is lemon coloured but my mother says it is yellow. Roses can be different colours, it has to be said red is the best though.	0.68	0.77
120	Would you like to drink this wine with your meal? Will you drink a glass of wine while you eat?	3.56	0.65
121	Roses can be different colours, it has to be said red is the best though. Roses come in many varieties and colours, but yellow is my favourite.	2.83	0.9
122	Flies can also carry a lot of disease and cause maggots. I dry my hair after I wash it or I will get ill.	0.12	0.28
123	Could you climb up the tree and save my cat from jumping please? Can you get up that tree and rescue my cat otherwise it might jump?	3.83	0.34
124	The pleasure that I get from studying, is that I learn new things. I have a doubt about this exam, we never got to study for it.	0.74	0.76
125	The perpetrators of war crimes are rotten to the core. There are many global issues that everybody should be aware of, such as the threat of terrorism.	0.95	0.91
126	The damp was mostly in the very corner of the room. The young lady was somewhat partially burnt from the sun.	0.11	0.31
127	We often ran to school because we were always late. I knew I was late for my class so I ran all the way to school.	3.1	0.85
128	I hope you're taking this seriously, if not you can get out of here. The difficult course meant that only the strong would survive.	0.5	0.87
129	The shores or shore of a sea, lake or wide river is the land along the edge of it. An autograph is the signature of someone famous which is specially written for a fan to keep.	0.11	0.43
130	I bought a new guitar today, do you like it? The weapon choice reflects the personality of the carrier.	0.16	0.34
131	I am so hungry I could eat a whole horse plus dessert. I could have eaten another meal, I'm still starving.	3.06	0.85

The following guidance is intended to help make benchmark tests performed with the data set comparable.

1. An STSS measure can be validated by comparing its performance with human ratings, in particular the ratings that a "typical" human might give.

2. The ratings in the table follow the practice used in word similarity studies [Miller and Charles 1991]. The "typical" human rating is the mean of those given by a set of participants. The measure of agreement is the Pearson product-moment correlation coefficient (*r*) quoted with statistical significance. The final column, *S*, is the corresponding standard deviation for each mean, a measure of noisiness or lack of precision of the ratings.

3. The ratings are the numbers in the $\overline{X}$ column; they are from a rating scale running from 0.00 to 4.00. The simplest procedure is to calculate the correlation coefficient between a new measure and the human ratings in the original range (0.00 – 4.00). Linear transformations are permissible, e.g. dividing by 4 to re-scale them to

run from 0.00 to +1.00. Re-scaling should not lead to a different correlation co-efficient (however, see below on rounding noise).

4. Consistency with other studies. Most STSS algorithms produce measures in the range from 0 to +1. Applying different rounding procedures can introduce noise and lead to variations in the least significant digit of *r*. For consistency with other studies, round the ratings from the STSS algorithm to 3 decimal places. Then calculate *r*, and round *r* to 3 decimal places. Common sense dictates that as the least significant digit of the 3 is based on the estimated digit, the importance of differences between measures based on this digit alone should not be exaggerated.

5. Those familiar with measurement theory may argue that mean and *r* are unsuitable statistics for data collected on this measurement scale. We are aware of the argument; however we have used the techniques because they are well-established and understood in the field of word similarity. Furthermore in the data collection process and the steps taken to improve ratio scale properties are described in detail in [O'Shea, et al. 2008] and [O'Shea 2010] as well as the current paper.

6. The calibration sentence pairs (SP99 and SP129) are taken from STSS-65 and **should not be used as part of this dataset**.

B. OTHER STATISTICAL TESTS.

Various tests for significance are appropriate in different circumstances. If we want to test the statistical significance of the difference between one STSS algorithm and another, these are dependent samples and the appropriate test is Steiger's z-test. This requires the construction of a correlation triangle, described later. The one-sample t-test can be used to compare a single correlation coefficient with an average of correlation coefficients (e.g. STASIS with the STSS-131 average from human raters). The 2-sample t-test can be used to compare averages from independent samples (e.g. to find a significant difference between the STSS-65 and STSS-131 datasets). Finally Fisher's r-to-z test can be used to compare correlation coefficients for the same algorithm across two different datasets (e.g. difference between LSA on STSS-65 vs. STSS-131).

USING STEIGER'S Z-TEST TO COMPARE TWO CORRELATION COEFFICIENTS (DEPENDEDNT SAMPLES)

Using Steiger's test to compare 2 correlation coefficients requires the construction of a correlation triangle. For example, consider comparing the correlation between STASIS and STSS-131 human ratings with the correlation between LSA and STSS-131 human ratings. Correlation triangles are formed according to Figure 1 and the specific triangle required for this calculation is shown in Figure 2.

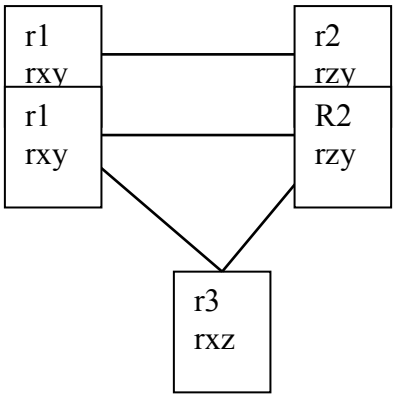

Fig. 1 General form of correlation triangle

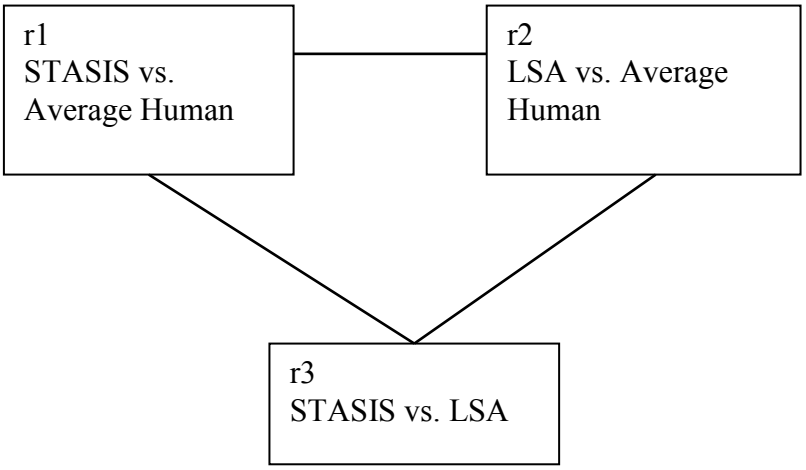

Fig. 2 Specific correlation triangle for STASIS vs. LSA

From Table IV in the main paper:

r1 rxy STASIS vs. Average humans 0.636

r2 rzy LSA vs. Average human 0.693

n=64 (64 Sentence Pairs without the two calibration pairs)

Calculated correlation:

r3 rxz STASIS vs. LSA 0.52

Applying the test gives the following results:

z-values for all differences:
Method Steigers Z

0.636-0.693= -0.057; z= -0.677; p= 0.7507
(left p: 0.2493; two sided: 0.4986)

0.636-0.52= 0.116; z= 1.48; p= 0.0695
(left p: 0.9305; two sided: 0.139)

0.693-0.52= 0.173; z= 2.126; p= 0.0167
(left p: 0.9833; two sided: 0.0334)

At the time of writing, Steiger's test was supported in neither Minitab nor SPSS. However, online calculators were available and two different calculators were used which gave consistent results, with very small

differences attributable to operations such as rounding within floating point calculations. The calculators were available at:

[Uitenbroek 2013] and [Grabin 2013]

Note that the test compares single correlation coefficients, not a correlation against an average of correlation coefficients, so in this case for "Average Human" we used the human with the inter-rater agreement closest to the average inter-rater agreement.

References

O'SHEA, J. D., et al. 2008. *A comparative study of two short text semantic similarity measures*. Lecture Notes in Artificial Intelligence 4953/2008, 172-181.

O'SHEA, J. 2010. A framework for applying short text semantic similarity in goal-oriented conversational agents. Manchester Metropolitan University.

MILLER, G. A. and CHARLES, W. G. 1991. *Contextual correlates of semantic similarity*. Language and Cognitive Processes 6 1-28.

UITENBROEK , D. G., 2013. Simple statistical correlation analysis online. http://www.quantitativeskills.com/sisa/statistics/correl.htm

GRABIN, C., 2013. General statistics. http://psych.unl.edu/psycrs/statpage/regression.html