Australian Government
Department of Defence
Defence Science and
Technology Organisation

Human Dimensions of Corpora Comparison: An Analysis of Kilgarriff's (2001) Approach

Kathryn Parsons, Agata McCormac and Marcus Butavicius

Command, Control, Communications and Intelligence Division
Defence Science and Technology Organisation

DSTO-TR-2290

ABSTRACT

There is a distinct lack of tools that provide a comprehensive measure of the similarity between corpora. Finding similar corpora is necessary for the design of certain user studies investigating text processing. It is also useful for ensuring comparability between studies on document analysis conducted across classified and unclassified domains. In this study, human judgements of corpora similarity were obtained as a gold standard. These were then compared to the values provided by Kilgarriff's (2001) chi-square (X^2) statistic. The findings indicated a high level of agreement between the participants, with 77% shared variance in overall similarity judgements. The results of the X^2 measure also correlated well with the human results, with a correlation of approximately 0.66. Although there are complexities associated with the X^2 technique that need to be examined in further research, this study provides extremely promising results, suggesting that a statistical technique could provide results that are comparable to human judgements.

RELEASE LIMITATION

Approved for public release

Published by

Command, Control, Communications and Intelligence Division
DSTO Defence Science and Technology Organisation
PO Box 1500
Edinburgh South Australia 5111 Australia

Telephone: (08) 8259 5555
Fax: (08) 8259 6567

AR-014-529
April 2009

Human Dimensions of Corpora Comparison: An Analysis of Kilgarriff's (2001) Approach

Executive Summary

A corpus is a collection of written or spoken material, and in fields such as information retrieval, machine translation and natural language processing, they are a vital resource. Corpora vary considerably, and knowledge regarding their similarities and differences are particularly important. For instance, a measure of similarity is necessary to determine whether the findings of one corpus are applicable to different corpora for the purposes of assessing document processing tools and human-user interaction abilities.

There is a distinct lack of tools that provide corpora comparisons, and the tools that do exist tend to provide a single value, which does not necessarily reflect the complexity associated with a collection of text. For example, corpora could be extremely similar in relation to content, but quite different in regards to structure or language use. Without information regarding the dimension of similarity that is being measured, the value provided by any corpora comparison scores are limited.

Within this study, seventeen corpora were utilised, and two random samples were taken from each corpus. Human corpora comparisons were obtained, which were then compared to the values provided by a statistical technique. The aims of this study were to (1) obtain comprehensive human judgements of corpora similarity to act as a gold standard, (2) compare the judgements obtained by different individuals, and (3) compare human judgements with those provided by a statistical technique.

The human judgements were made on a number of dimensions of similarity. The correlations between the participants' scores were extremely high, with an overall correlation of 0.88, indicating 77% shared variance between the participants. However, when participants' scores were assessed according to the various dimensions and corpora categories, there was far more variation. Hence, this indicates that corpora comparison is influenced by subjectivity and individual differences.

Kilgarriff's (2001)[1] X^2 statistic is a word-frequency based measure, which involves a statistical analysis of the most frequent words in a pair of corpora. The word list is then compared to the most frequent words in both corpora, to examine the discrepancy between the observed frequency of words and the expected frequency if the corpora were derived from the same underlying body of text. This technique was used to determine the similarity between the corpora, and the results were then compared to the human judgements.

[1] Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics 6(1),* 1-37.

The correlations between the chi-square results and the participants' ratings of similarity were high, with an overall correlation of approximately 0.66. This is extremely promising, as it suggests that a statistical technique may provide results that are comparable to human judgements. However, it is necessary to note that there was a large range in the strength of the correlations for the corpora pairs within the various categories. Hence, it is possible that the chi-square technique is more effective for certain types of corpora.

Furthermore, there are a number of complexities associated with the chi-square technique, which could limit the generalisability of these results. For example, it is unclear whether there is an optimal number of words for the word frequency lists or an optimal size for the corpora samples.

Authors

Kathryn Parsons
Command, Control, Communications & Intelligence Division

Kathryn Parsons is a research scientist with the Human Interaction Capabilities Discipline in C3ID where her work focuses on cognitive and perceptual psychology, information visualisation and interface design. She obtained a Graduate Industry Linked Entrepreneurial Scheme (GILES) Scholarship in 2005, with LOD, where she was involved in human factors research, in the Human Sciences Discipline, specifically in the area of Infantry Situation Awareness. She completed a Master of Psychology (Organisational and Human Factors) at the University of Adelaide in 2005.

Agata McCormac
Command, Control, Communications & Intelligence Division

Agata McCormac joined DSTO in 2006. She is a research scientist with the Human Interaction Capabilities Discipline in C3ID where her work focuses on cognitive and perceptual psychology, information visualisation and interface design. She was awarded a Master of Psychology (Organisational and Human Factors) at the University of Adelaide in 2005.

Marcus Butavicius
Command, Control, Communications & Intelligence Division

Marcus Butavicius is a research scientist with the Human Interaction Capabilities Discipline in C3ID. He joined LOD in 2001 where he investigated the role of simulation in training, theories of human reasoning and the analysis of biometric technologies. In 2002, he completed a PhD in Psychology at the University of Adelaide on mechanisms of visual object recognition. In 2003 he joined ISRD where his work focuses on data visualisation, decision-making and interface design. He is also a Visiting Research Fellow in the Psychology Department at the University of Adelaide.

Contents

1. Introduction

A corpus is, simply put, a collection of written or spoken material. Corpora are vital for numerous fields, including machine translation, natural language processing, information retrieval and linguistics. The scope of available corpora is extremely large, ranging from unstructured and informal corpora, such as transcriptions of conversations or emails, to very structured and formal corpora, such as newspaper articles and abstracts.

Although such corpora are widely used, there is very little research examining their similarities (or differences). Without such information it is difficult to ascertain whether the findings of a study using one corpus can be generalised to other areas. Furthermore, of the few tools that do provide a measure of corpora similarity, little effort has gone into comparing their performance against a human baseline.

Without such a gold standard, it is difficult to ascertain the relevance or value of any corpora similarity score. In essence, it is unclear exactly which aspect of the various corpora is being used to determine 'similarity'. For example, two corpora could be extremely similar in terms of content, but quite dissimilar in relation to structure or style. Essentially, more research is required to ascertain (1) which aspect of similarity is being measured in the various corpora comparisons, and (2) whether this score reflects the sort of judgments that humans make.

This study sought to address this deficiency. Participants were provided with samples from a variety of corpora, and were required to provide a measure of overall corpora similarity. Participants were also asked to make similarity ratings on various measures, such as content, structure and language use. The aim was to assess the way that humans rate the similarity of various corpora. This study also provides a comparison between the human ratings and the similarity measures obtained using a statistical technique.

1.1 Quantifying Corpus Similarity

Questions regarding the similarity between various corpora are of vital importance. With the increased prevalence of communication modes such as email and short message service (SMS), the ability to quantify the similarity between corpora is crucial. Essentially, the scope of available corpora differs so widely in relation to formality and structure that the findings of one corpus may not be transferable to different corpora.

1.1.1 Written Versus Spoken Language

A number of authors have examined the differences between written and spoken language. For instance, Blankenship (1962) examined a sample of articles and speeches by public figures and concluded that syntactic differences are more influenced by differences in individuals rather than differences in the mode of communication. In contrast, Poole & Field (1976) affirm that there are consistent differences between oral and written language. For instance, speech often contains more personal reference, less elaboration, less verb complexity and greater structural complexity.

This idea is supported by Chafe (1979), who suggests that one of the main differences between the communication modes is the type of relation that they imply with their audience. In general, written language is more detached, consistent and more defensible over time. In contrast, spoken language tends to be more involved and fragmented, with speakers far more concerned about the richness of their communication. Drieman (1962) examined texts and transcriptions from graduate students and also found consistent differences between written and spoken language. Speech was more likely to consist of shorter words, more words with one syllable, longer texts and a less varied vocabulary.

In addition to the differences between spoken and written language highlighted above, there are also numerous differences within spoken and written corpora. This is particularly true given the prevalence of email communication, which, despite being a written communication mode, is not limited to formal writing styles but is often written in an extremely colloquial and informal manner. Essentially, both written and spoken communication can differ widely in regards to formality and structure. Hence, classifications such as 'spoken' and 'written' communication do not expose the complexity of the problem, and it is necessary to examine these categories in more detail.

1.1.2 Statistical Measurement Techniques

There are numerous statistical techniques that attempt to quantify corpora similarity. Information such as the number of characters, paragraphs or lines can be used to reveal some information about the corpus. Other available measures of the similarity between various corpora include n-gram or word frequencies, which aim to determine the words that are particularly characteristic of a corpus (Kilgarriff, 2001).

Although there are various techniques available, there is very little information regarding which of the techniques is the most effective (Kilgarriff, 2001). In order to evaluate numerous methods, Kilgarriff (2001) developed a technique referred to as "Known-Similarity Corpora" (KSC). This method involves using two distinct corpora to build a set of known similarities (see also Butavicius, Ferguson & Mullen, 2009).

Several new corpora are created using different percentages of the two distinct corpora. For example, one would consist of 30% of one of the distinct corpora and 70% of the other; another corpus would consist of 40% of one of the distinct corpora and 60% of the other. Based on the different percentages, the similarities between these developed corpora are known, and can be used as a standard to evaluate corpora comparison techniques.

Kilgarriff (2001) used KSC to empirically assess numerous methods, including chi-square (X^2), Spearman's rank correlation coefficient and three cross-entropy measures. The most successful of the approaches was the X^2 technique.

1.1.2.1 The X^2 Statistic

The X^2 method is a word-frequency based measure, in which the occurrence of the most common words in two corpora is compared to the expected frequency if the two corpora were random samples from the same population. In other words, the chi-square technique measures the discrepancy between the observed frequency of words and the expected frequency if the corpora were both drawn from the same larger corpus. This means that the technique not only uses words that are likely to be indicative of a topic, but also uses words that might suggest a certain linguistic style.

1.2 A ‘Human Perspective’ on Corpus Similarity

Most of the available measures for quantifying corpus similarity do not capture the multidimensional nature of the problem. Instead, the available measures often provide a single score, which would only reveal an element of the similarity between corpora.

This study investigates more sophisticated measures of similarity, utilising numerous dimensions, such as content, language use and structure. For example, a corpus of business emails written in a professional manner could be very similar to a corpus of personal emails in relation to the structure, but quite different in terms of language use and content. In fact, when analysed according to language use and content, business emails might have more in common with newspaper articles. This therefore emphasises the importance of obtaining a measure of how humans assess corpora similarity, how this differs between individuals, and how the human judgements compare to those produced by automated techniques.

However, in order to obtain these human judgements, it is necessary for individuals to make judgements of every corpus in relation to every other corpus. This is a very time consuming and demanding process and, due to the considerable effort involved, it is necessary to limit the number of participants. It would be possible for a larger number of participants to judge only a subset of the comparisons. However, evidence suggests that, due to individual differences, averaging across individuals can produce results that do not faithfully represent human similarity judgements (Ashby, Maddox & Lee, 1994; Lee & Pope, 2003). Therefore, for this study, the complete set of judgements were made by only two participants, and the results of both participants were separately compared to the similarity measures obtained using the statistical technique.

2. Methodology

2.1 Participants

The participants were two research scientists from The Defence Science and Technology Organisation (DSTO).

2.2 Materials

2.2.1 Corpora

Seventeen corpora were assessed. These ranged widely in terms of structure, formality and content. As shown in Table 1, the corpora were divided into a number of different categories, namely:

- Academic corpora
- Newspaper corpora
- Speech
- Emails
- Others (i.e. abstracts and SMS)

A number of the corpora were also divided into subcategories. For example, previous experiments on the Enron corpus (Parsons, McCormac & Butavicius, 2009) found that participants' ability to locate facts from within emails was significantly higher for questions that required access to non-work related emails than for the questions that involved the work related messages. Hence, for this experiment, the Enron corpus was divided into subcategories, with work related emails examined separately to the non-work related emails.

Furthermore, the Michigan Corpus of Academic English (MICASE) includes a number of different discourse modes, from a number of different academic disciplines. For this experiment, the corpus was divided by discourse mode, creating a 'dialogue' based subcategory, a 'lecture' based subcategory and a 'panel' based subcategory. For the British Academic Spoken English (BASE) Corpus, only the 'seminar' based documents were used, which provides a good comparison to MICASE.

The LA Times and Foreign Broadcast Information Service (FBIS) articles were taken from the TREC-8 corpus (Voorhees & Harman, 2000).

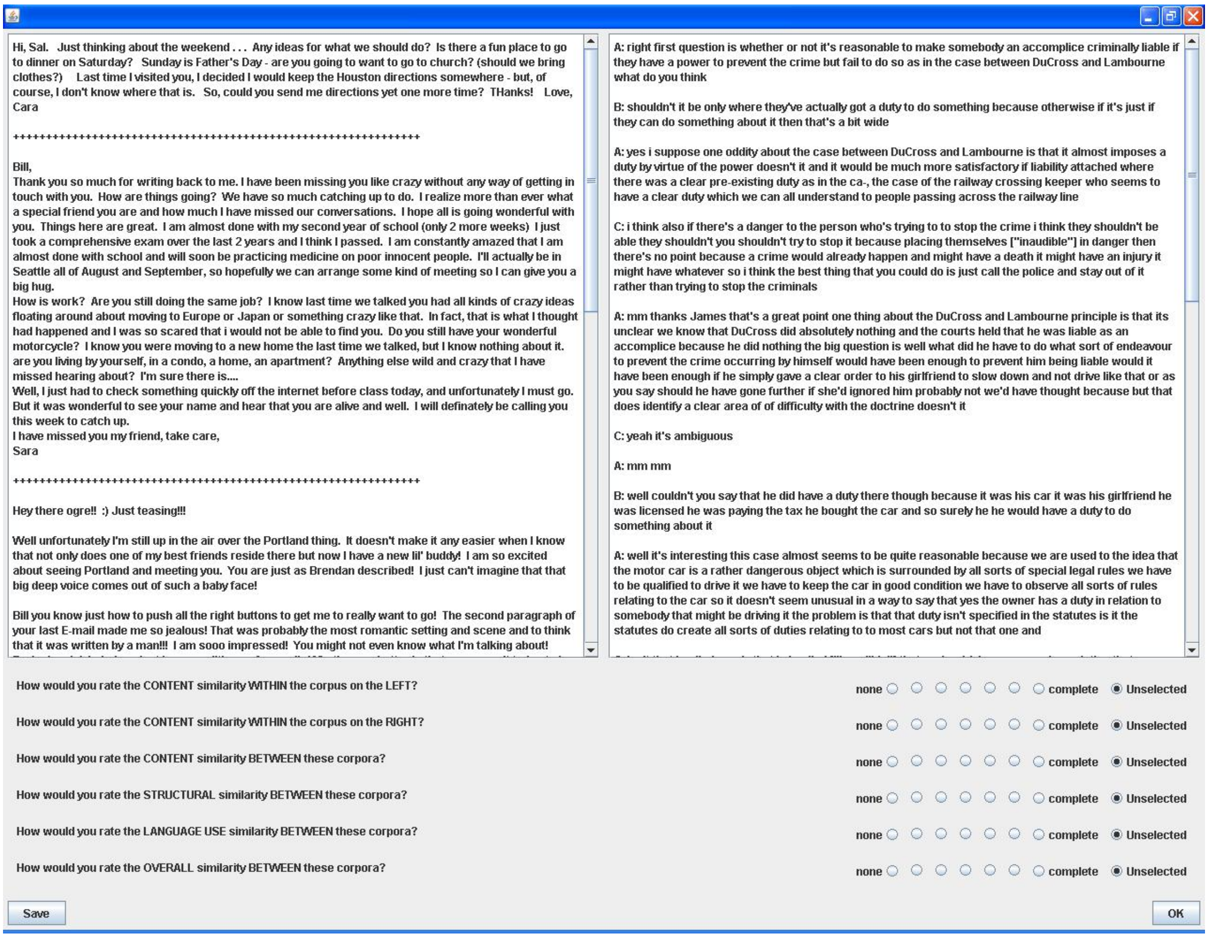
Table 1: The specific corpora used by category of corpora

	Corpora			
Academic	MICASE (dialogue)	MICASE (lecture)	MICASE (panel)	BASE (seminar)
Newspaper	LA Times	FBIS	Reuters	Newsmail
Speech	LDC	BNC (meeting)	BNC (conversation)	-
Email	Enron (non-work)	Enron (work)	SPAM	Newsgroup
Other	Abstracts	SMS (Singapore Corpora)	-	-

Since it was unpractical for judgements to be made on the entire corpora, two random samples, of approximately 1000 words each, were taken from the 17 corpora. The inclusion of two random samples of each corpus allows a measure of split-half reliability, which is determined by correlating the scores for the two samples. This can be used to ensure consistency. The samples began from the start of a document and, where the sample included more than one document, a line was used to separate the documents. Where a single document had more than 1000 words '*[……]*' was used to indicate that the document continued. Appendix A contains more details of the samples, such as the number of words, characters, paragraphs, lines and documents in each sample.

2.2.2 Interface

The experiment was completed using a simple interface, consisting of two document windows, a question box, and a rating scale for the answer. An example of the interface can be seen in Figure 1, below.



Hi, Sal.   Just thinking about the weekend . . . Any ideas for what we should do? Is there a fun place to go to dinner on Saturday?   Sunday is Father's Day - are you going to want to go to church? (should we bring clothes?)   Last time I visited you, I decided I would keep the Houston directions somewhere - but, of course, I don't know where that is.   So, could you send me directions yet one more time? THanks!   Love, Cara

++

Bill,
Thank you so much for writing back to me. I have been missing you like crazy without any way of getting in touch with you. How are things going? We have so much catching up to do. I realize more than ever what a special friend you are and how much I have missed our conversations. I hope all is going wonderful with you. Things here are great. I am almost done with my second year of school (only 2 more weeks) I just took a comprehensive exam over the last 2 years and I think I passed. I am constantly amazed that I am almost done with school and will soon be practicing medicine on poor innocent people. I'll actually be in Seattle all of August and September, so hopefully we can arrange some kind of meeting so I can give you a big hug.
How is work? Are you still doing the same job? I know last time we talked you had all kinds of crazy ideas floating around about moving to Europe or Japan or something crazy like that. In fact, that is what I thought had happened and I was so scared that i would not be able to find you. Do you still have your wonderful motorcycle? I know you were moving to a new home the last time we talked, but I know nothing about it. are you living by yourself, in a condo, a home, an apartment? Anything else wild and crazy that I have missed hearing about? I'm sure there is....
Well, I just had to check something quickly off the internet before class today, and unfortunately I must go. But it was wonderful to see your name and hear that you are alive and well. I will definately be calling you this week to catch up.
I have missed you my friend, take care,
Sara

++

Hey there ogre!! :) Just teasing!!!

Well unfortunately I'm still up in the air over the Portland thing. It doesn't make it any easier when I know that not only does one of my best friends reside there but now I have a new lil' buddy! I am so excited about seeing Portland and meeting you. You are just as Brendan described! I just can't imagine that that big deep voice comes out of such a baby face!

Bill you know just how to push all the right buttons to get me to really want to go! The second paragraph of your last E-mail made me so jealous! That was probably the most romantic setting and scene and to think that it was written by a man!!! I am sooo impressed! You might not even know what I'm talking about!

A: right first question is whether or not it's reasonable to make somebody an accomplice criminally liable if they have a power to prevent the crime but fail to do so as in the case between DuCross and Lambourne what do you think

B: shouldn't it be only where they've actually got a duty to do something because otherwise if it's just if they can do something about it then that's a bit wide

A: yes i suppose one oddity about the case between DuCross and Lambourne is that it almost imposes a duty by virtue of the power doesn't it and it would be much more satisfactory if liability attached where there was a clear pre-existing duty as in the ca-, the case of the railway crossing keeper who seems to have a clear duty which we can all understand to people passing across the railway line

C: i think also if there's a danger to the person who's trying to to stop the crime i think they shouldn't be able they shouldn't you shouldn't try to stop it because placing themselves ["inaudible"] in danger then there's no point because a crime would already happen and might have a death it might have an injury it might have whatever so i think the best thing that you could do is just call the police and stay out of it rather than trying to stop the criminals

A: mm thanks James that's a great point one thing about the DuCross and Lambourne principle is that its unclear we know that DuCross did absolutely nothing and the courts held that he was liable as an accomplice because he did nothing the big question is well what did he have to do what sort of endeavour to prevent the crime occurring by himself would have been enough to prevent him being liable would it have been enough if he simply gave a clear order to his girlfriend to slow down and not drive like that or as you say should he have gone further if she'd ignored him probably not we'd have thought because but that does identify a clear area of of difficulty with the doctrine doesn't it

C: yeah it's ambiguous

A: mm mm

B: well couldn't you say that he did have a duty there though because it was his car it was his girlfriend he was licensed he was paying the tax he bought the car and so surely he he would have a duty to do something about it

A: well it's interesting this case almost seems to be quite reasonable because we are used to the idea that the motor car is a rather dangerous object which is surrounded by all sorts of special legal rules we have to be qualified to drive it we have to keep the car in good condition we have to observe all sorts of rules relating to the car so it doesn't seem unusual in a way to say that yes the owner has a duty in relation to somebody that might be driving it the problem is that that duty isn't specified in the statutes is it the statutes do create all sorts of duties relating to to most cars but not that one and

How would you rate the CONTENT similarity WITHIN the corpus on the LEFT? none ○ ○ ○ ○ ○ ○ ○ complete ● Unselected

How would you rate the CONTENT similarity WITHIN the corpus on the RIGHT? none ○ ○ ○ ○ ○ ○ ○ complete ● Unselected

How would you rate the CONTENT similarity BETWEEN these corpora? none ○ ○ ○ ○ ○ ○ ○ complete ● Unselected

How would you rate the STRUCTURAL similarity BETWEEN these corpora? none ○ ○ ○ ○ ○ ○ ○ complete ● Unselected

How would you rate the LANGUAGE USE similarity BETWEEN these corpora? none ○ ○ ○ ○ ○ ○ ○ complete ● Unselected

How would you rate the OVERALL similarity BETWEEN these corpora? none ○ ○ ○ ○ ○ ○ ○ complete ● Unselected

Figure 1: A screenshot of the interface

2.3 Method

Using the provided interface, participants were asked to view samples of two corpora simultaneously, and then provide a rating of their similarity on a number of specified dimensions including content, structure and language use. Participants judged the similarity of the content within each corpus. This is necessary as it is possible that a particular corpus could have a lot of variability, which could influence the significance of the similarity judgements between corpora. The questions used are shown below:

- How would you rate the CONTENT similarity WITHIN the document on the LEFT?
- How would you rate the CONTENT similarity WITHIN the document on the RIGHT?
- How would you rate the CONTENT similarity BETWEEN these documents?
- How would you rate the STRUCTURAL similarity BETWEEN these documents?
- How would you rate the LANGUAGE USE similarity BETWEEN these documents?
- How would you rate the OVERALL similarity BETWEEN these documents?

Participants responded on a seven point similarity scale (from 'none' to 'complete') for all responses. A seven point scale was used because research indicates that this is the optimal number of rating scale categories (Tang, Shaw & Vevea, 1999). Participants provided ratings for each of the samples in comparison to every other sample on each of the variables. Hence, participants made 561 comparisons, creating a total of 3366 judgements.

The participants used a score of one for a corpora pair with no similarity, and a score of seven for a corpora pair with complete similarity. However, a subsequent section of this report involves a comparison between the human ratings and the ratings provided by a statistical technique. Since the statistical technique provides a scoring where a smaller score corresponds to greater similarity, the participants' scores were reversed. Hence, a score of one represents a high level of similarity, and a score of seven represents very little similarity.

Before commencement of the judgements, participants were provided with definitions for each of the dimensions. These are shown in Table 2, below.

Table 2: Definitions for the different dimensions

Dimension	Definition
Content	What it is about; the subjects and topics covered; information provided.
Structure	The appearance and arrangement of the document (design and layout). The relationships between fields, entities, language, page and paragraph breaks, length and other editorial devices.
Language Use	Choice of words, grammatical structure, sentence type and language, punctuation, formal/informal, long/short.
Overall	An overall measure that takes into account all other dimensions, including any unspecified dimensions.

To prevent fatigue, participants were able to save their position and complete the judgements in more than one sitting. The corpora were shown in the same order to both participants.

3. Results – Part 1: The Human Comparisons

3.1 Summary of Results

Participants made similarity judgements for content, structure, language use and overall similarity for 34 corpora samples, comparing every sample with every other sample. The comparisons took approximately five hours for each participant. As shown in Table 3, the level of agreement in the similarity judgements assigned by the two participants was extremely high. Content similarity had the lowest correlation, particularly when the corpora in a judged pair belonged to different categories. However, when the corpora in a judged pair belonged to the same category, the agreement between participants was far higher.

The opposite was true for structural similarity; the agreement was extremely high when the corpora were from different categories, and far lower when the judged pair belonged to the same category. For overall similarity, participants were more likely to agree in their judgements when pairs of corpora were from different categories than when the pairs were from the same category.

Although there are differences in the magnitude of these correlations, it is important to note that all correlations were significant to the 0.01 alpha level, suggesting a very high level of agreement, regardless of the corpora pair. This should, however, be interpreted with caution, as the number of corpora pairs was very large, and with large sample sizes, very small correlations can be statistically significant.

Table 3: Correlations between Participant A and Participant B

	Different Categories	**Same Categories**	**All Results**
Content Similarity	0.45	0.82	**0.79**
Structural Similarity	0.82	0.41	**0.90**
Language Use Similarity	0.67	0.61	**0.82**
Overall Similarity	**0.74**	**0.69**	**0.88**

Multiple regression analyses were conducted to assess whether the different measures of similarity (content, structure and language use) were predictive of the participants' overall similarity ratings. For both participants, content similarity, structural similarity and similarity in regards to language use explained a significant 97% of the variance in overall similarity scores (Participant A: $R^2 = 0.970$, $F(3, 557) = 6077.31$, $p < 0.001$ | Participant B: $R^2 = 0.968$, $F(3, 557) = 5633.39$, $p < 0.001$). The Beta values for content, structure and language use were all significant, but interestingly, the variable that was most significant differed for the two participants.

For Participant A, similarity in language use had the most influence on overall similarity, predicting 48% of the variance ($\beta = 0.481$). This was followed by structural similarity, $\beta = 0.397$, and content similarity was the least predictive, $\beta = 0.166$. For Participant B, content similarity was also least predictive, $\beta = 0.161$. However, language use was far less predictive for Participant B, accounting for only 36% of the variance ($\beta = 0.363$). Instead, structural similarity was the most predictive dimension, $\beta = 0.511$. This suggests that participants may differ in the dimensions that they use to judge overall similarity, with language use explaining 48% of Participant A's overall ratings, and structural similarity explaining 51% of Participant B's overall ratings.

The results for each of the assessed dimensions and the results for the corpora categories will now be analysed in more detail.

3.2 Overall Similarity Judgements

The agreement between participants was very high, with a correlation for the overall similarity judgements of 0.88 (*N* = 561), which accounts for 77% shared variance. Generally, the ratings assigned for overall similarity had a high proportion of dissimilar ratings, and the average overall ratings for Participant B (*M* = 5.43, *SD* =1.58) included more dissimilar ratings than the judgements provided by Participant A (*M* = 5.09, *SD* = 1.68). An examination of the mean difference in scores between participants shows that the vast majority of participants' scores were extremely close (*M* = 0.34, *SD* = 0.81).

In fact, 281 of the 561 corpora pairs (50%) were given identical overall similarity ratings by both participants. These ratings are shown in the dark grey squares in Table 4. For example, 164 of the 561 ratings were given a score of six by both participants. Participants were more likely to choose rating six than any other rating, with this response given for 268 (48%) of the ratings for Participant B, and 230 (41%) of Participant A's responses. These response frequencies are depicted in Figure 2.

Table 4: Frequency table comparing overall similarity ratings[2]

					'B'				
		1	**2**	**3**	**4**	**5**	**6**	**7**	**TOTAL**
	1	**12**	5	0	0	0	0	0	**17**
	2	6	**14**	14	11	1	0	0	**46**
	3	0	4	**14**	33	12	2	0	**65**
'A'	**4**	0	2	5	**18**	5	4	1	**35**
	5	0	0	1	2	**1**	62	7	**73**
	6	0	0	0	3	2	**164**	61	**230**
	7	0	0	0	0	1	36	**58**	**95**
	TOTAL	**18**	**25**	**34**	**67**	**22**	**268**	**127**	**561**

The results also indicate that the participants differed in the range of scores that they tended to give, with Participant B less likely to give a score below six. Participant B responded with a score of five or below on 166 occasions (or 30% of responses) in contrast to Participant A, who responded with a score of five or below on 236 occasions (or 42% of responses). However, as shown in the comparative matrix for the overall similarity judgements in Figure 3, participants' scores have a very positive distribution, indicating a strong relationship between the participants' similarity judgements. In this figure, each cell represents the frequency of correspondence between participants' judgments for each possible judgment pairing. For example, the cell at x,y value {4,3} represents the frequency of judgments where Participant A

[2] As mentioned in the Method section, the participants' scoring was reversed. This means that a score of one indicates that a corpora pair was extremely similar, and a score of seven indicates that pair was very dissimilar.

responded with a rating of '4' and Participant B responded with '3' for the same corpora pairing(s). The cells are coloured according to the relative frequency, e.g., black represents a frequency of correspondence of approximately 25% while white represents no corresponding scores for that value.

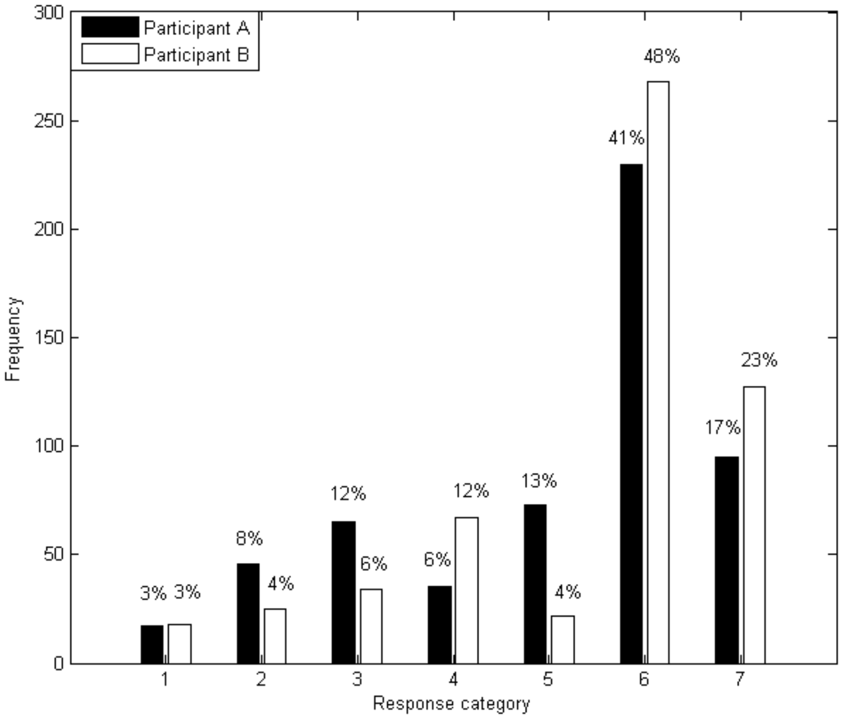


Figure 2: Response frequency distribution for the overall similarity judgements for 'A' and 'B'

Both Figure 3 and Figure 4 indicate that, where the responses between participants differed, in the vast majority of cases, the scores differed by only one point, indicating a very high level of agreement for almost all responses. Figure 4 also shows a negative skew in the distribution, with Participant A frequently assigning a rating one point lower than Participant B.

When the results were assessed to include the identical responses and the responses where the participants' ratings were within one point of each other, the agreement was extremely high. Of the 561 overall similarity judgements, in 92% of cases (516 judgements) participants' responses were identical or within one point of each other.

Hence, although agreement was high, the frequency of responses that differed by one point could indicate a difference in participants' interpretation of the rating scale. Essentially, similarity is a subjective concept, and the participants may differ in their opinion of what constitutes 'similar'. Furthermore, as indicated in the multiple regression analysis in the previous section, the participants differed in the dimension that was most predictive of overall similarity. Therefore, since participants' overall ratings were often measuring different aspects of similarity, it is logical that the scores will have some differences.

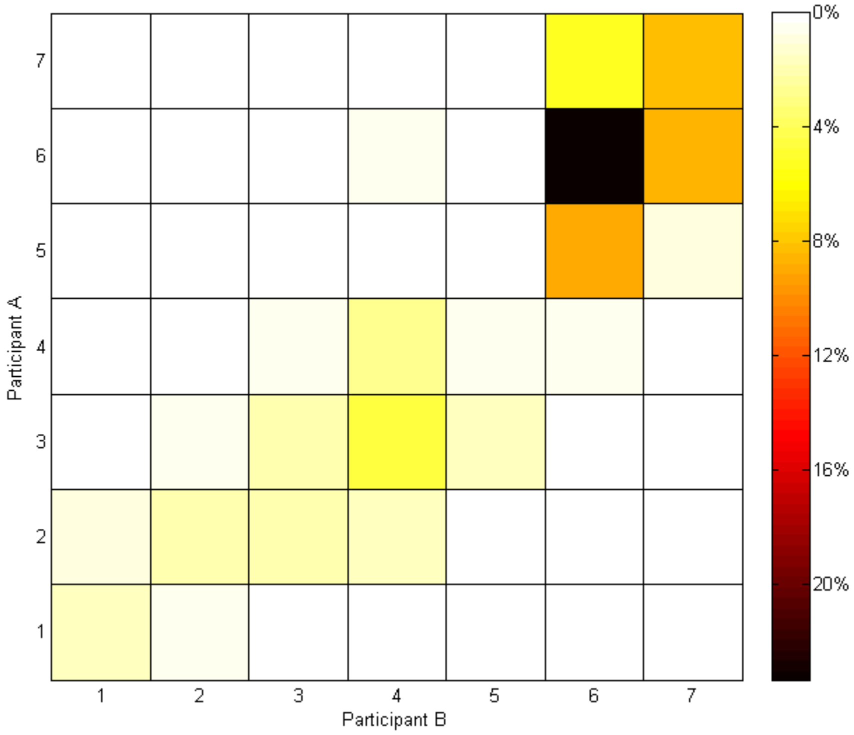


Figure 3: Comparative matrix for the overall similarity judgements for 'A' vs. 'B'

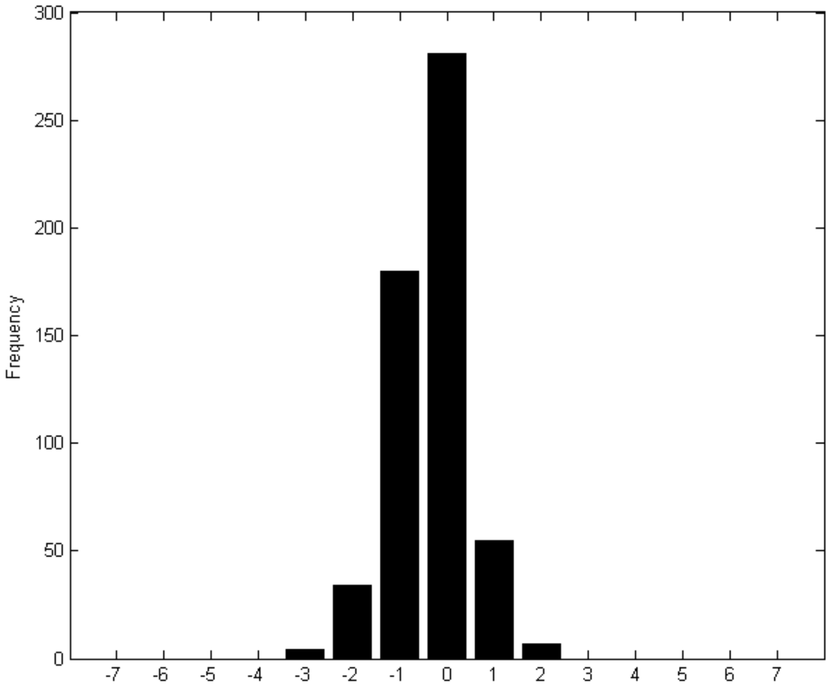


Figure 4: Histogram of the differences in overall similarity for 'A' vs. 'B'

3.3 Content Similarity Judgements

Participants provided similarity ratings for the content within and between all corpora. The content similarity within corpora was judged to be very high, with an average rating of approximately two (which represents high similarity). There was also a high level of agreement between participants, with 83% of responses identical for the corpora of the left, and 79% the same for the corpora on the right of the interface. All other responses were within one point of each other, and the mean differences were extremely small, with a mean difference of only 0.07 (*SD* = 0.40) for the similarity within the corpora on the left of screen and a mean difference of only 0.19 (*SD* = 0.42) for the similarity within the corpora on the right of screen.

The correlation for content similarity ratings (between corpora) was lower than the correlations for the other dimensions. However, it was still very high, with a correlation of 0.78 (*N* = 561), which indicates approximately 61% shared variance. There was a mean difference in content scores between participants of 0.48 (*SD* = 0.03), and as shown in the histogram in Figure 6, there was a negative skew in the distribution, with Participant A frequently giving a rating one below the rating provided by Participant B. This is reflected in the mean scores, with a mean for Participant A of 5.53 (*SD* = 1.17) compared to a mean of 6.01 (*SD* = 1.19) for Participant B.

However, there was still a high level of similarity in all responses, with 275 of the 561 corpora pairs (49%) given identical content similarity ratings by both participants. These ratings are shown in the dark grey squares in Table 5, indicating that 145 of the 561 judgements were given a rating of six by both participants. As found in the scores for overall similarity, participants were more likely to choose rating six than any other rating, with this response given for 269 (48%) of the ratings for Participant B, and 239 (42%) of Participant A's responses.

Table 5: Frequency table comparing content similarity ratings

					'B'				
		1	**2**	**3**	**4**	**5**	**6**	**7**	**TOTAL**
	1	**0**	1	0	0	0	0	0	**1**
	2	0	**22**	3	3	0	0	0	**28**
	3	0	3	**1**	0	2	1	0	**7**
'A'	**4**	0	0	3	**1**	14	9	1	**28**
	5	0	1	2	0	**29**	97	35	**164**
	6	0	0	1	0	5	**145**	88	**239**
	7	0	0	0	0	0	17	**77**	**94**
	TOTAL	**0**	**27**	**10**	**4**	**50**	**269**	**201**	**561**

The results in the comparative matrix in Figure 5 and in Table 5 also indicate that the participants differed in the range of scores that they tended to give, with Participant B less likely to give a score below six. Participant B responded with a score of five or below on 91 occasions (or 16% of responses) in contrast to Participant A, who responded with a score of five or below on 228 occasions (or 41% of responses). Interestingly, Participant B was more than twice as likely to respond with a rating of seven, and Participant A responded with a rating of five more than three times as frequently as Participant B. Despite these differences, participants assigned an identical rating or a rating within one point of each other for 506 out of the 561 cases (90%).

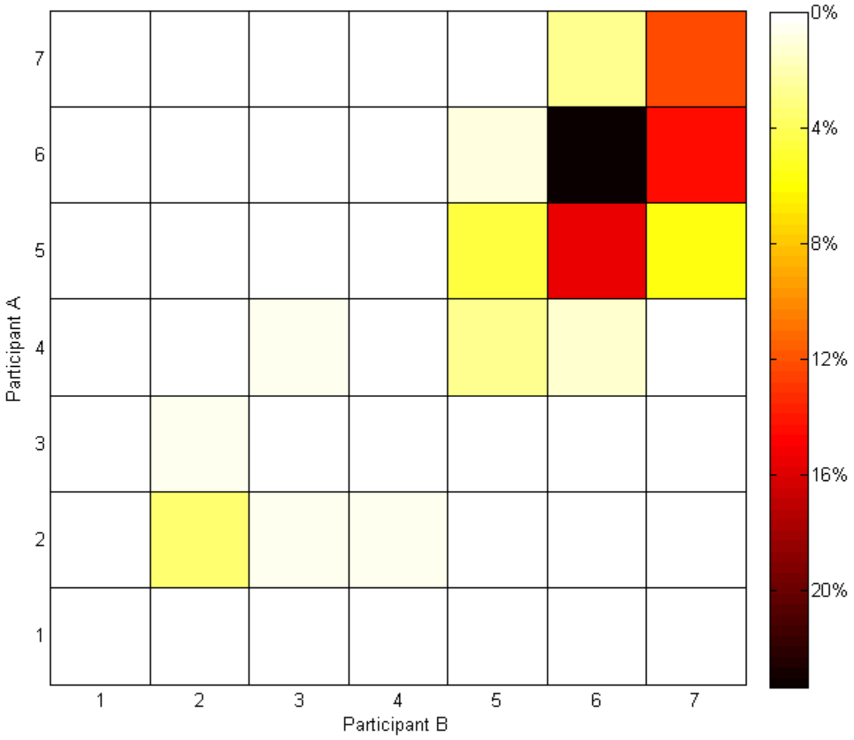


Figure 5: Comparative matrix for the content similarity judgements for 'A' vs. 'B'

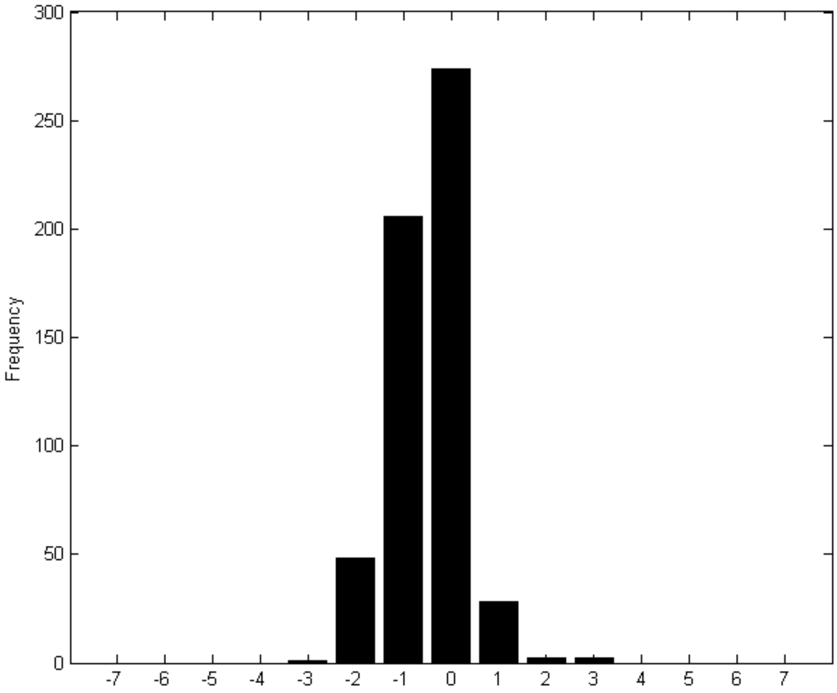


Figure 6: Histogram of the differences in content similarity for 'A' vs. 'B'

3.4 Structural Similarity Judgements

The participants' ratings for structural similarity were the most similar, with an extremely high correlation of 0.90 (*N* = 561), and a very small mean difference in scores of only 0.08 (*SD* = 0.82). As found for the other dimensions, the mean structural similarity ratings indicated that a high proportion of the corpora were judged to be dissimilar (Participant A: *M* = 5.12, *SD* = 1.86 | Participant B: *M* = 5.21, *SD* = 1.84).

Table 6: Frequency table comparing structural similarity ratings

					'B'				
		1	**2**	**3**	**4**	**5**	**6**	**7**	**TOTAL**
'A'	**1**	**16**	11	11	1	0	0	0	**39**
	2	4	**11**	17	1	2	0	0	**35**
	3	2	21	**38**	6	4	0	0	**71**
	4	0	9	5	**1**	1	1	1	**18**
	5	0	1	1	0	**1**	14	3	**20**
	6	0	0	2	0	4	**202**	59	**267**
	7	0	0	0	0	1	22	**66**	**111**
	TOTAL	**22**	**53**	**74**	**9**	**13**	**261**	**129**	**561**

For structural similarity, 60% of the corpora pairs (or 335 out of 561) were given identical structural similarity ratings by both participants. These ratings are shown in the dark grey squares in Table 6, indicating that 202 of the 561 judgements were given a rating of six by both participants. Again, participants were more likely to choose rating six than any other rating, with this response given for 261 (47%) of the ratings for Participant B, and 267 (48%) of Participant A's responses.

Interestingly, the scores for participants' structural similarity ratings had less variability than the scores for the other dimensions, with Participant A providing a score of six or seven (indicating very little similarity) in 70% of cases, compared to Participant B, who provided a rating of six or seven in 67% of cases. This high level of similarity is reflected in the histogram in Figure 7, clearly indicating that participants most frequently responded with the same similarity rating. This finding is also emphasised by Figure 8, which shows a positive distribution, where most responses tended towards the upper end of the response range.

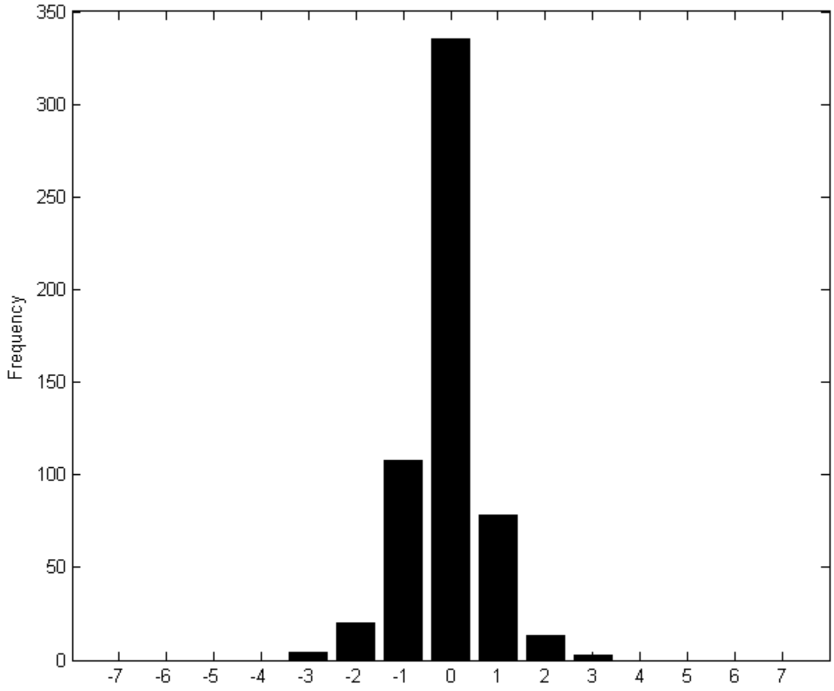


Figure 7: Histogram of the differences in structural similarity for 'A' vs. 'B'

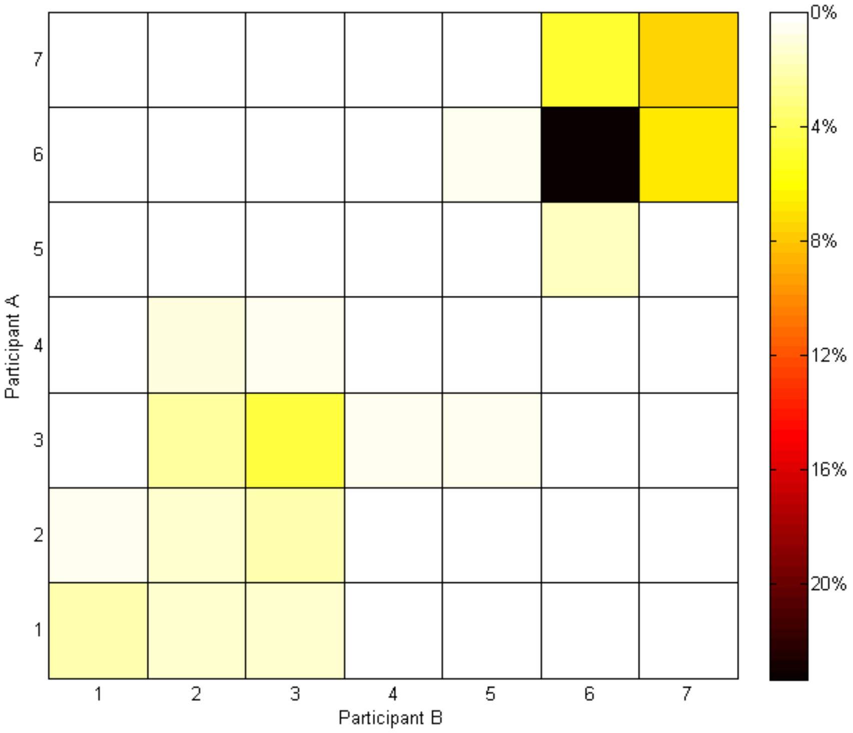


Figure 8: Comparative matrix for the structural similarity judgements for 'A' vs. 'B'

3.5 Language Use Similarity Judgements

Participants had a very high level of agreement in their ratings of language use similarity, with a correlation of 0.82 (*N* = 561), and a mean difference of only 0.47 (*SD* = 1.10). The average rating for Participant A was 4.82 (*SD* = 1.93), which again indicates more similar judgements than Participant B's average rating of 5.30 (*SD* = 1.62).

As shown in Table 7, participants gave identical language use similarity ratings for 43% of responses, or 239 of the 561 corpora pairs. Once again, participants were more likely to choose rating six than any other rating, with this response given for 242 (43%) of the ratings for Participant B. Interestingly, Participant A only responded with a rating of six in 33% of cases (184 ratings), which is far lower than the number of 'six' responses for the other dimensions. In 120 cases both participants assigned a rating of six.

Table 7: Frequency table comparing language use similarity ratings

					'B'				
		1	2	3	4	5	6	7	TOTAL
	1	**24**	14	11	8	3	0	0	**60**
	2	2	**6**	15	11	9	2	0	**45**
	3	0	3	**9**	3	14	5	0	**34**
'A'	4	0	1	5	**10**	9	7	1	**33**
	5	0	0	1	2	**18**	75	11	**107**
	6	0	0	0	2	15	**120**	47	**184**
	7	0	0	0	1	12	33	**52**	**98**
	TOTAL	**26**	**24**	**41**	**37**	**80**	**242**	**111**	**561**

The results also indicate that (as shown for the other dimensions) Participant A commonly responded with a score one rating point below Participant B's response. This is also shown in the comparative matrix in Figure 9. Furthermore, despite the fact that participants had a high level of agreement, as shown in the histogram in Figure 10, there were a small number of cases where the responses differed by four points. For example, as shown in Table 7, there were three occasions where Participant A assigned a language use similarity rating of one (indicating complete similarity) and Participant B assigned a rating of five. This is likely to be a reflection of the subjectivity associated with similarity ratings.

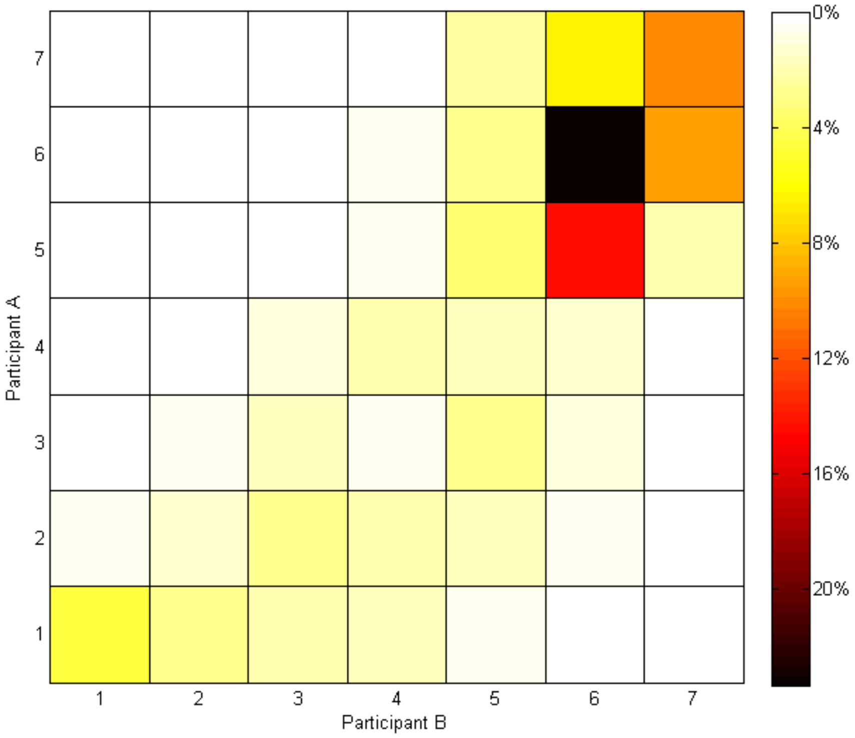


Figure 9: Comparative matrix for the language use similarity judgements for 'A' vs. 'B'

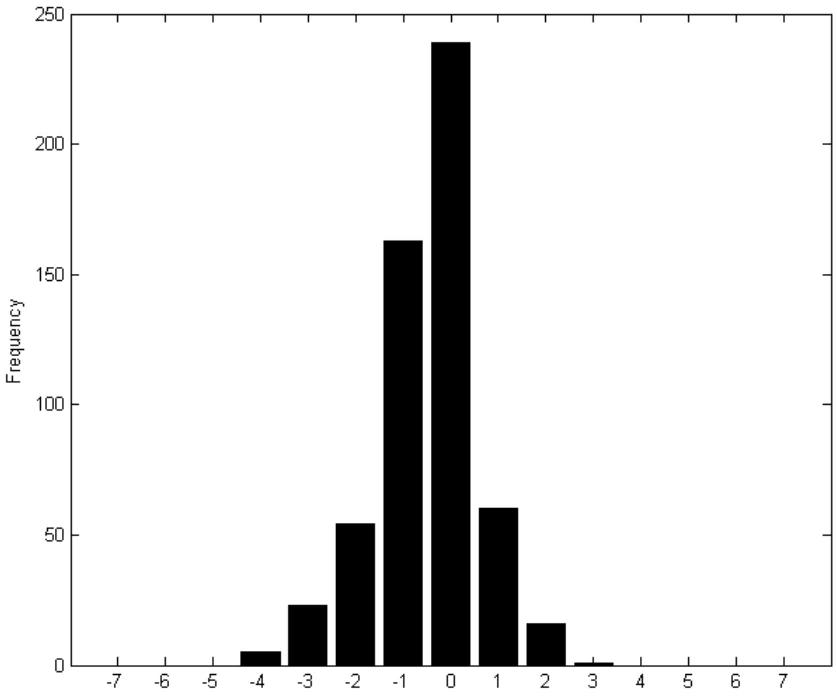


Figure 10: Histogram of the differences in language use similarity for 'A' vs. 'B'

3.6 Split-Half Reliability and Validity

As well as ensuring that there was a high level of agreement between participants' ratings, it was also necessary to attempt to validate the results. As mentioned previously, two samples were taken from each of the 17 corpora. If both the KSC approach was correct and the judgments of the participants were valid, then the participants should have assigned lower scores (representing higher similarity) to the two halves of the same corpus, and higher scores (representing lower similarity) to the other comparisons. Furthermore, if the KSC approach was valid and the participants' ratings were reliable, then the scores assigned to the different halves of the same corpus should be consistent across the two samples.

These within and between corpora judgements for overall similarity are displayed in Figure 11, clearly indicating that the participants tended to view within-corpora judgements as more similar (Participant A: *M* = 1.59, *SD* = 0.87 | Participant B: *M* = 1.59, *SD* = 0.94) and between-corpora judgements as less similar (Participant A: *M* = 5.20, *SD* = 1.58 | Participant B: *M* = 5.55, *SD* = 1.43). These results validate the use of the KSC approach in analysing the corpora judgements as well as the quality of the human judgments.

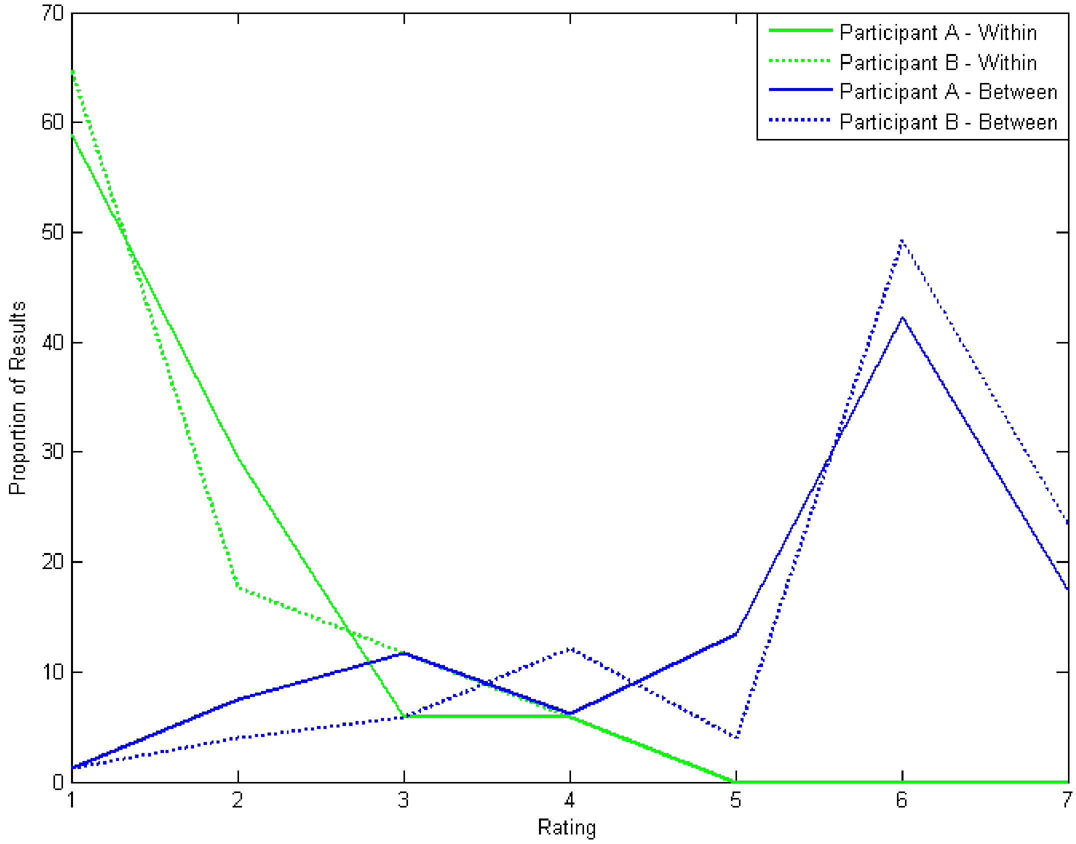


Figure 11: Histogram for within-corpora and between-corpora judgements for 'A' and 'B'

As shown in Table 1, the corpora were divided into different categories such as academic, newspaper, speech and email. The consistency between the participants' ratings for each of the categories will now be analysed in detail.

3.6.1 Academic Corpora

Participants made similarity judgements on eight academic corpora. When the corpora were assessed according to the halves of the same corpus, the average ratings for both participants indicated very high similarity [Participant A: *M* = 2.25, *SD* = 0.43 and Participant B: *M* = 2.75, *SD* = 1.09]. Although the averages were high, the ratings for content similarity were generally far lower, with a mean of 4.88 (*SD* = 1.13). However, since the academic corpora were divided by discourse mode rather than academic division, these differences in similarity are not surprising. For example, although Corpora 4 and 7 were both student presentations from the panel section of MICASE, Corpus 4 was from Humanities and Arts, whereas Corpus 7 was from Biological and Health Sciences. Hence, the observed variation in content similarity is logical.

3.6.2 Newspaper Corpora

When just the results from the newspaper corpora were examined there was extremely high agreement between participants, with only two judgements where the ratings differed (and in those two cases the ratings only differed by one point). There was also extremely high similarity between the halves of the corpora, with an average for overall similarity of 1.13 (*SD* = 0.35), suggesting that newspaper corpora were judged to be very similar.

3.6.3 Speech Corpora

Interestingly, Participant B rated the similarity for the different halves of the speech corpora to be higher than Participant A. The average overall similarity rating for Participant B was 1.33 (*SD* = 0.58) in contrast to an average rating of 2.67 (*SD* = 1.15) for Participant A. The highest variation was for content similarity ratings, where Participant A tended to judge the corpora as less similar, which suggests more individual differences or subjectivity in ratings of content similarity.

3.6.4 Email Corpora

Participants had a high level of agreement for the email corpora, with only four different judgements, which differed by only one point. For overall similarity, Participant A assigned an average rating of 1 (*SD* = 0) and Participant B assigned an average rating of 1.25 (*SD* = 0.5), which suggests that both participants agreed that there was very little variation in the email corpora.

3.6.5 Other Corpora

Participants gave identical similarity ratings for structure, language use and overall similarity for the SMS corpora. The content similarity judgements both within and between the SMS corpora were lower, but given that each corpus had approximately 80 messages, it was expected that there would be variation in content.

There was very little variation in the abstract corpora, with Participant A assigning scores of one for all dimensions, and Participant B assigning scores of one for all except for content (which was assigned a score of two).

3.7 The Influence of Category on Similarity Ratings

The results were examined according to whether the corpora in an examined pair were from different categories (e.g., one email corpus and one newspaper corpus) or from the same category (e.g., both emails).

3.7.1 Validation Using Multidimensional Scaling Displays

As mentioned previously, if the participants' similarity judgements were reliable and valid and the KSC approach correct, the scores assigned to the two halves of the same corpus should be consistent. In addition, the corpora from the same category should tend to be judged as more similar than the corpora from different categories. In order to visually inspect the relationship between the categories and represent the participants' judgements in two dimensions, Multidimensional Scaling (MDS)[3] was applied to the human ratings (Cox & Cox, 1994). Using MDS, each corpora was assigned *x*, *y* coordinate pairs, so that the more similar corpora were placed closer together in the display. The validity and utility of MDS for presenting similarity data in such a manner has been verified in a number of empirical studies (e.g., Butavicius & Lee, 2007; Lee, Butavicius & Reilly, 2003).

Appendix B shows MDS displays for the similarity judgements of Participant A and Participant B. As shown in these displays, the different halves of the same corpus were placed extremely close together, indicating a high level of similarity. The corpora were also clearly clustered according to categories. This level of clustering was, however, far weaker for the judgements of content similarity. This suggests that language use and structural similarity were generally more consistent across corpora, and across different corpora from the same category.

Interestingly, for all dimensions of similarity, Participant A tended to group the corpora from the academic and speech categories very closely together. Since the academic corpora were also transcriptions of speech, it is logical for these to be clustered together. In the majority of

[3] The Euclidean distance metric (with an additive constant to ensure that the triangular inequality axiom was met) was used to generate the MDS displays. Non-linear least squares optimisation based on the Levenberg-Marquardt optimisation approach was used (Lee, 1999). The solution was optimised with respect to the Variance Account For (VAF) in the two-dimensional solution in comparison to the empirical space. The MDS algorithm was tested on 100 iterations for each display and the solution selected was that which minimised the VAF. For more details see Butavicius, Lee, Pincombe, & Mullen (2006).

cases, the two halves of the same corpus were reciprocal nearest neighbours; that is, the closest point to one half of the corpora was usually the other half of the same corpus. These findings suggest that the participants' similarity judgements were both reliable and valid and that the KSC approach was successful.

3.7.2 Similarity Judgements for Corpora from the Same Category

Of the 561 corpora comparisons, 101 judgements were made on category pairs, where both corpora were from the same category. However, within some categories there was still substantial variation in the types of corpora. For instance, within the email category there were emails from a SPAM corpus, messages from newsgroups, and emails from the Enron corpus, with separate corpora for work and non-work emails. Hence, although these are one category, some variance was still anticipated.

As expected, the response frequency distribution for the corpora comparisons from the same category (see Figure 12) differs quite significantly to the overall response frequency distribution (see Figure 4). For the overall response distribution, rating six and seven were extremely common. In contrast, for the response distribution for pairs from the same category, ratings of one and two were most common, and there were very few responses for ratings six and seven.

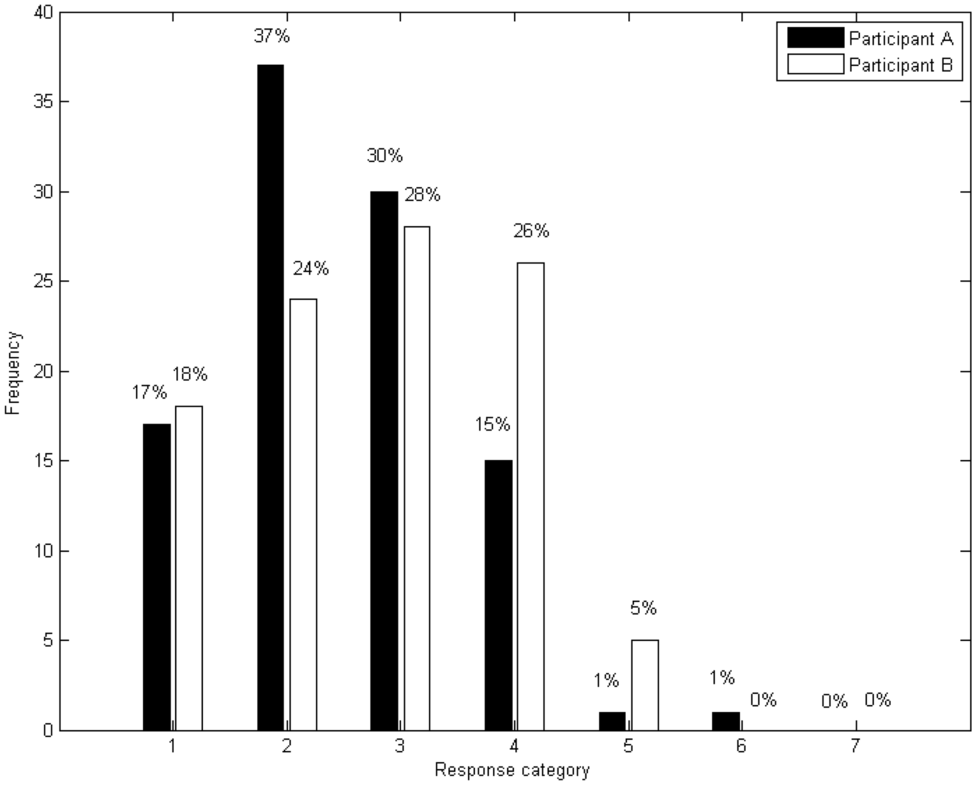


Figure 12: Response frequency distribution for overall similarity for category pairs

Unsurprisingly, the mean similarity ratings indicated far greater similarity for the comparisons involving corpora from the same category. For example, the overall rating for Participant A for the category pairs was 2.50 (*SD* = 1.04), compared to a mean of 5.53 (*SD* = 1.17) when all comparisons were included. Participant B had a similar increase in similarity ratings when only the same category comparisons were examined, with the mean score for all comparisons of 6.01 (*SD* = 1.19) and a score for category pairs of 2.76 (*SD* = 1.17). This high level of agreement is also shown in Table 8, which indicates that the participants' scores were identical in 46 of the 101 judgements, and where identical or within one point of each other in 91 of the 101 judgements (90% of cases).

Table 8: Frequency table comparing both overall similarity ratings for the category pairs

					'B'				
		1	**2**	**3**	**4**	**5**	**6**	**7**	**TOTAL**
'A'	**1**	**12**	5	0	0	0	0	0	**17**
	2	6	**14**	12	5	0	0	0	**37**
	3	0	4	**12**	12	2	0	0	**30**
	4	0	1	3	**8**	3	0	0	**15**
	5	0	0	1	0	**0**	0	0	**1**
	6	0	0	0	1	0	**0**	0	**1**
	7	0	0	0	0	0	0	**0**	**0**
	TOTAL	**18**	**24**	**28**	**26**	**5**	**0**	**0**	**101**

Figure 13 shows the similarity scores in regards to structural, language use, content and overall similarity for each participant by the different categories. These graphs indicate a high level of agreement in the participants' judgements. Also, Figure 13 indicates that (particularly for the academic, speech and email corpora) there was relatively more variation in the content covered in these corpora.

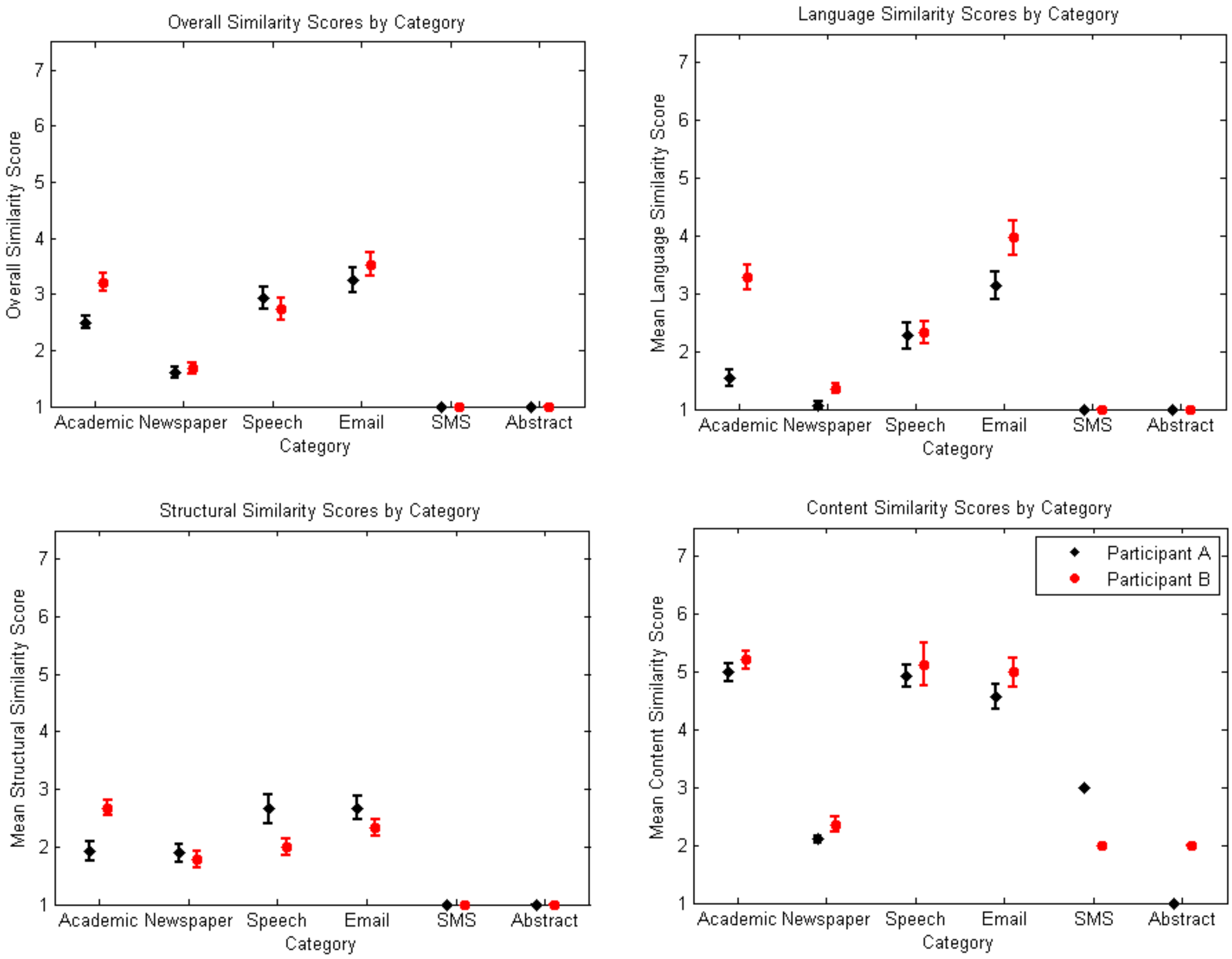


Figure 13: Mean scores for each of the categories for each of the similarity dimensions

3.7.3 Similarity Judgements for Corpora from Different Categories

Of the 561 corpora comparisons, 460 of the judgements were made on corpora pairs where the corpora were from different categories (e.g., one from a newspaper corpus and one from an email corpus). Figure 14, which shows the response frequency distribution of the corpora from different categories, closely resembles Figure 4, which shows the response frequency distribution of all corpora pairs.

When only the comparisons from different categories were assessed, the correlation for the overall similarity ratings dropped from 0.88 (N = 561) to 0.74 (N = 460). There was a similar decrease for the language use similarity correlations, decreasing from 0.82 (N = 561) for all comparisons, to 0.67 (N = 460) when only the comparisons from different categories were included. Interestingly, the largest difference was in the ratings for content similarity, which dropped from 0.78 (N = 561) to only 0.45 (N = 460) when just the comparisons from different categories were assessed. In contrast, there was very little change in the correlation for structural similarity.

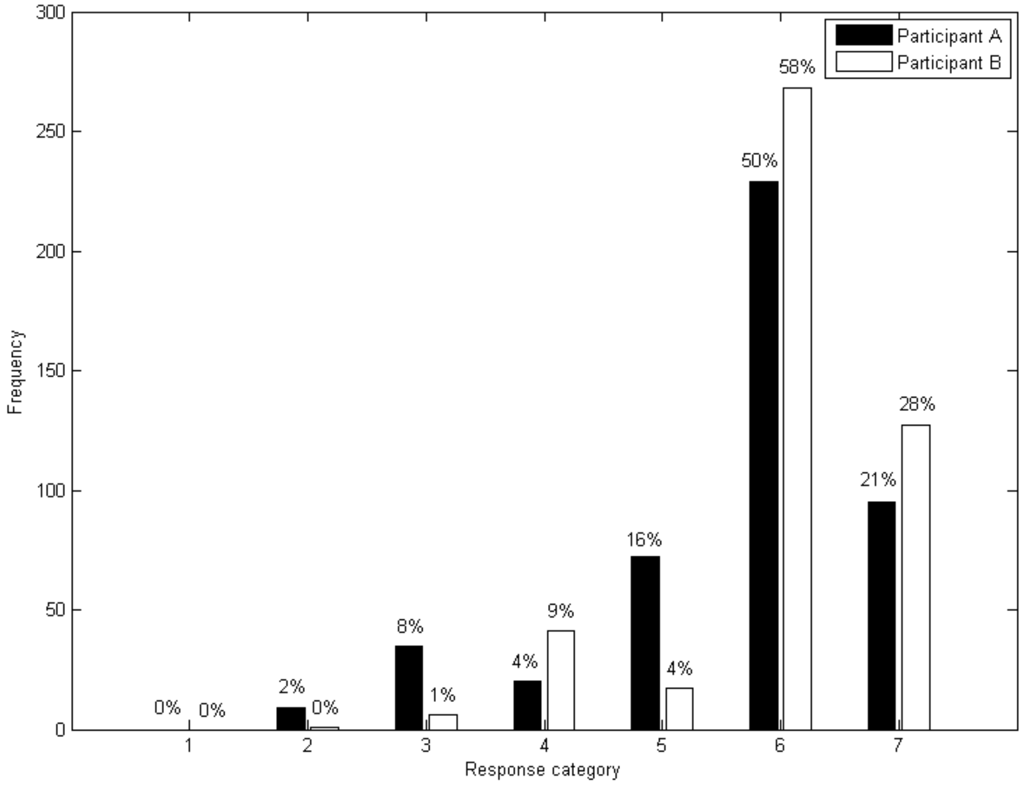


Figure 14: Response frequency distribution for overall similarity judgements for different categories

This seems to suggest that, when there is less similarity between the corpora, individual differences have a greater influence on the assessment of content similarity. This highlights the subjective nature of similarity judgements, and the fact that different aspects or dimensions of the similarity ratings are influenced in a different manner. It is also important to note that these 'unlike' judgements are arguably less useful from a practical perspective than the similarity judgements for corpora from the same category. Essentially, people are more likely to be interested in finding two corpora that are highly similar than two corpora that are different.

Although there appears to be a large amount of variation between the correlations for the different category pairs, as shown in the comparative matrixes in Figure 15, the relationship between the participants' responses were quite similar for all dimensions. Essentially, there was a high level of agreement between participants. This is also shown in Table 9, which indicates that participants had identical responses for 235 of the 460 responses (51%). When the responses within one point of each other were also included, the agreement rose to 92%.

Table 9: Frequency table comparing overall similarity ratings for the different category pairs

					'B'				
		1	**2**	**3**	**4**	**5**	**6**	**7**	**TOTAL**
	1	**0**	0	0	0	0	0	0	**0**
	2	0	**0**	2	6	1	0	0	**9**
	3	0	0	**2**	21	10	2	0	**35**
'A'	**4**	0	1	2	**10**	2	4	1	**20**
	5	0	0	0	2	**1**	62	7	**72**
	6	0	0	0	2	2	**164**	61	**229**
	7	0	0	0	0	1	36	**58**	**95**
	TOTAL	**0**	**1**	**6**	**41**	**17**	**268**	**127**	**460**

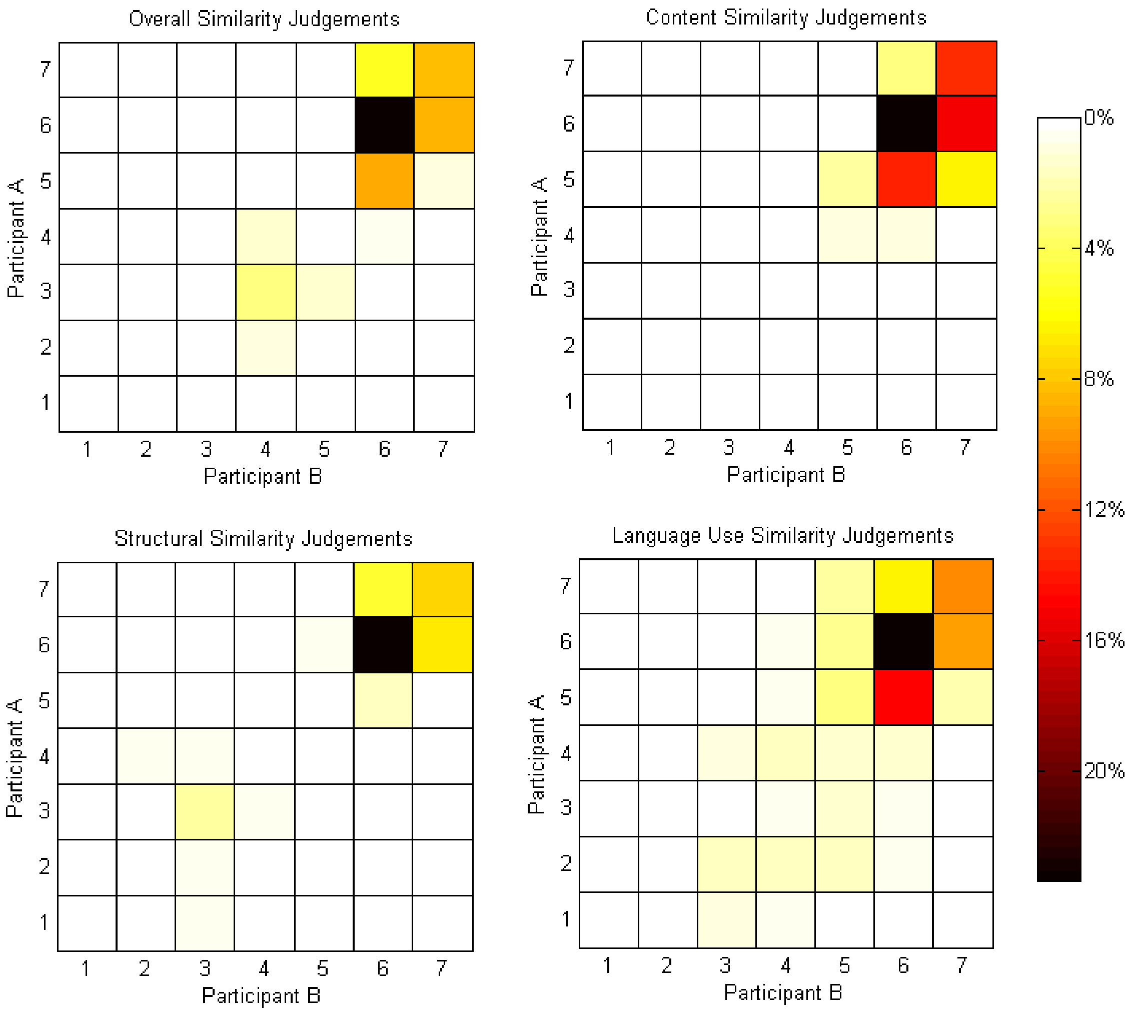


Figure 15: Comparative matrixes for different category similarity judgements for participants 'A' vs. 'B'

4. Results – Part 2: The X^2 Statistics

4.1 Document Pre-processing

Since it is extremely time consuming for humans to make comparisons between every corpus, a statistical technique that provides a valid measure of corpus similarity is invaluable. Kilgarriff's (2001) X^2 method of corpus similarity was used for this study and a script was developed to run the technique. Essentially, the chi-square technique involves the creation of a word list containing the top *n* words that exist in two corpora. Since this list is an integral part of the technique, it is important that the number of words used is large enough to adequately represent the corpora, but this must also be balanced to ensure that the technique does not use too high a proportion of words (or all unique words in a corpus). Essentially, the measure is more likely to be accurate if it is based on the words that are most characteristic or representative of a corpus (Kilgarriff, 2001).

Prior to running the script, unnecessary characters and strings (such as punctuation) were removed, and the file was concatenated, with a single file for all of the different documents from the same corpus. Next, the text was converted into word frequency lists, which consists of a list of unique words along with the number of times that each word appeared in a file.

Since it was difficult to ascertain the ideal number of words to use in the corpora comparison a priori, a number of different alternatives were assessed. Pilot testing revealed that humans were able to make conclusions regarding the entire corpus after seeing only a small sample, and therefore participants analysed only 1000 word corpora samples. In contrast, since the nature of the statistical techniques relies on the number of times a word occurs, it is likely that the small samples may not provide an adequate representation of the whole corpus. Hence, in addition to the 1000 word samples, larger samples, containing approximately 10000 words, were also obtained for each of the 34 corpora. This was done in order to empirically assess the influence of sample size on the chi-square approach.

Kilgarriff's (2001) original testing of the chi-square technique involved analysis of all the words in the word frequency lists. However, in this study our testing was done under two conditions; one with all words included and the other with stop words removed. Stop words are words that occur very frequently in normal language use and are considered to be words that carry no semantic or content information. Examples include words such as 'the', 'and', 'if' and 'but'. The removal of stop words is usually undertaken to improve the performance of a text processing algorithm and there is recent empirical evidence supportive of this practice (Lee, Pincombe, & Welsh, 2005). However, although stop words do not generally add any further information about content, they may provide important information about the structure or style of a corpus. For example, stop words may have meaning if you are looking at a corpus that is based on transcriptions of speech as opposed to a highly-structured news text corpus. Perhaps including stop words may actually tell you a lot about the structure of a corpus. The corpora with the stop words removed will be referred to as the *stopped* corpora while the corpora containing all the original words will be referred to as the *all words* corpora. The list of stop words is included in Appendix C.

It was also difficult to determine the appropriate size for the word frequency lists. Since the statistic does not dictate how many words to use, Kilgarriff (2001) tested the technique on word frequency lists of various sizes, ranging from 10 words to 5120 words. Kilgarriff (2001) achieved the best results with 320 or 640 words. However, since it was unknown whether that result would also apply for the corpora used in this study, it was necessary to thoroughly test the appropriate number of words. The script was therefore run through intervals of 50 words, ranging from 50 to 700 words.

4.2 Interpreting the X^2 Statistic

4.2.1 Dimensionality of the Chi-Square Measure

One limitation of the chi-square technique that was not discussed in Kilgarriff (2001) is the dimensionality of the measure as detailed by Butavicius et al. (2009). This means that a similarity score of Corpora 1 to Corpora 2 is not necessarily the same as that comparing Corpora 2 to Corpora 1. Because of this directionality, the chi-square value cannot be considered a metric. In Butavicius et al.'s (2009) study, the directional sensitivity of the measure was related to the magnitude of the averaged chi-square value across the two comparisons. Specifically, the sensitivity was reduced for corpora halves that, on average, were judged to be either very similar or very dissimilar. In contrast, the participants' task was to compare the two documents without reference to directionality.

The chi-square scores obtained in both directions were examined, and the difference between these measures is shown in Table 10, which shows the proportion of change from the average of the two scores. The differences between the directional measures were also examined based on whether they were *within corpora* results or *between corpora* results, where the *within corpora* results include the comparisons of the two halves of the same corpora, and *between corpora* results include all other findings. Results support the findings of Butavicius et al. (2009), suggesting that the difference between the directional measures tended to be higher for the between corpora results than the within corpora results. Furthermore, as found by Butavicius and colleagues (2009), there was less difference for highly similar versus highly dissimilar scores. Overall, these findings (see Table 10) indicate that there was not a large difference between the scores, and therefore, for the purposes of this study, the average of the chi-square scores was used.

Table 10: Proportion of change from the average of the two scores for 1000 and 10000 word samples

Words Used	1000 stopped	1000 all words	10000 stopped	10000 all words
50	1%	6%	7%	8%
100	<1%	5%	5%	6%
150	0%	4%	4%	5%
200	<1%	4%	3%	5%
250	2%	3%	3%	4%
300	4%	3%	2%	3%
350	6%	3%	2%	3%
400	6%	1%	1%	3%
450	6%	1%	1%	3%
500	6%	1%	1%	2%
550	6%	1%	1%	2%
600	6%	1%	<1%	2%
650	6%	1%	<1%	2%
700	6%	1%	<1%	2%

4.2.2 Differences in Chi-Square Distributions

One of the most important methods of determining the accuracy of the chi-square method involves comparing the within and between corpora results. As explained, the *within corpora* results include the comparisons of the two halves of the same corpora, and *between corpora* results include all other findings. Figure 16 shows the within and between comparisons for 1000 word sample corpora with 50, 350 and 700 most common words, and Figure 17 shows the within and between comparisons for 10000 word sample corpora with 50, 350 and 700 most common words. These graphs allow the comparison of distributions of the averaged X^2 scores across the *all words* condition and the *stopped* condition, along with the type of comparisons, these being within and between corpora.

An examination of the averaged chi-square values generally reveals overlap in the between and within corpora comparisons for both the *all words* and *stopped* conditions. A high level of overlap essentially reveals that the technique cannot differentiate which corpora are the same and which corpora are different. This means that using the chi-square score between a pair of corpora to determine whether they were from the same underlying corpus would be an unreliable technique.

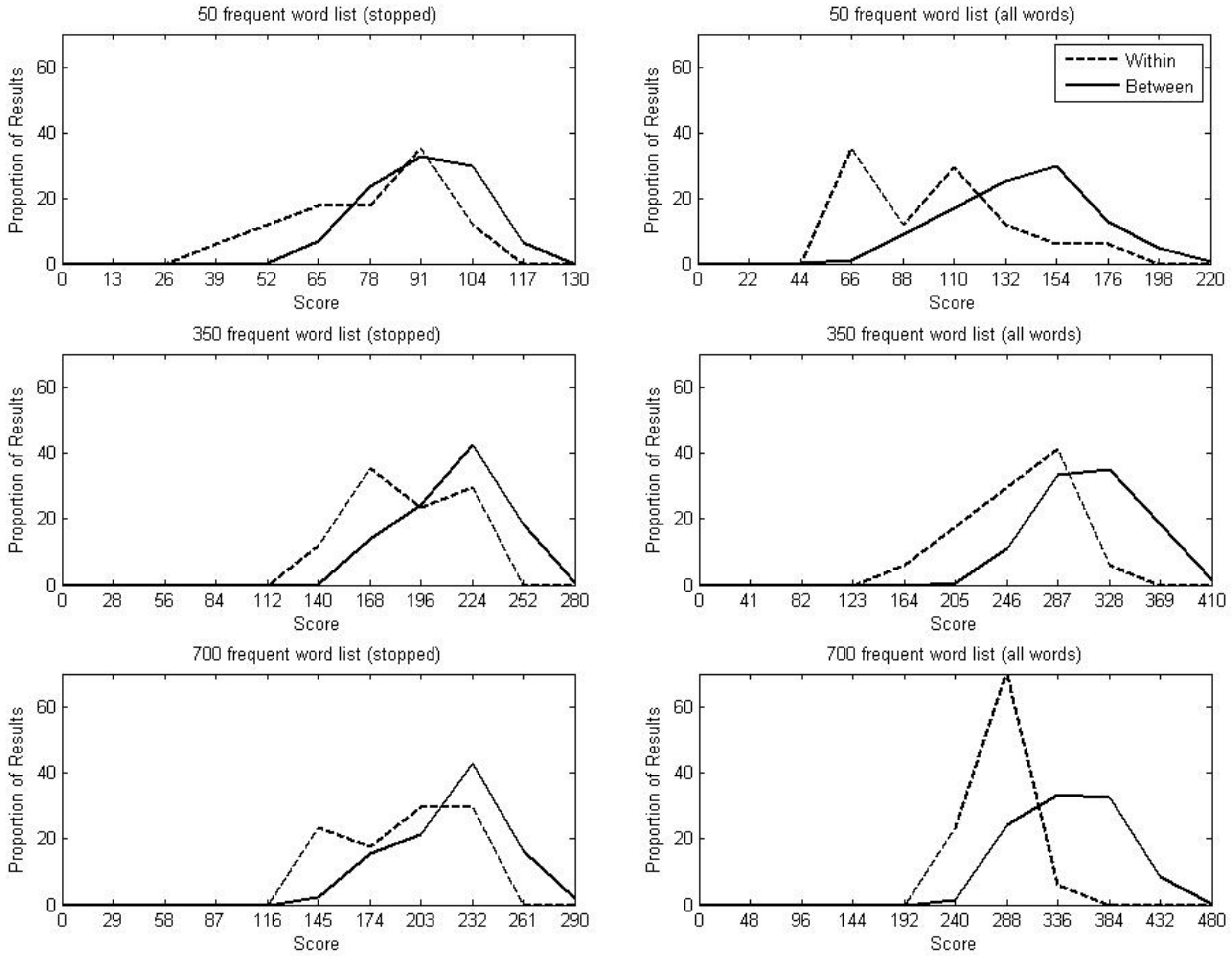


Figure 16: Within and between comparisons for 1000 word sample corpora with 50, 350 and 700 most common words

There is far less overlap for the condition that utilises the 10000 word sample, with the 700 most common words, and *all words* included (see Figure 17). This indicates better discrimination for the between and within corpus values, meaning that the value provided should enable a user to judge (with reasonable accuracy) whether an unknown corpus is from the same or a different corpora. In the graph below, a score over 2400 is most likely to represent a corpus from different corpora, and a score below 1000 would represent the same corpora.

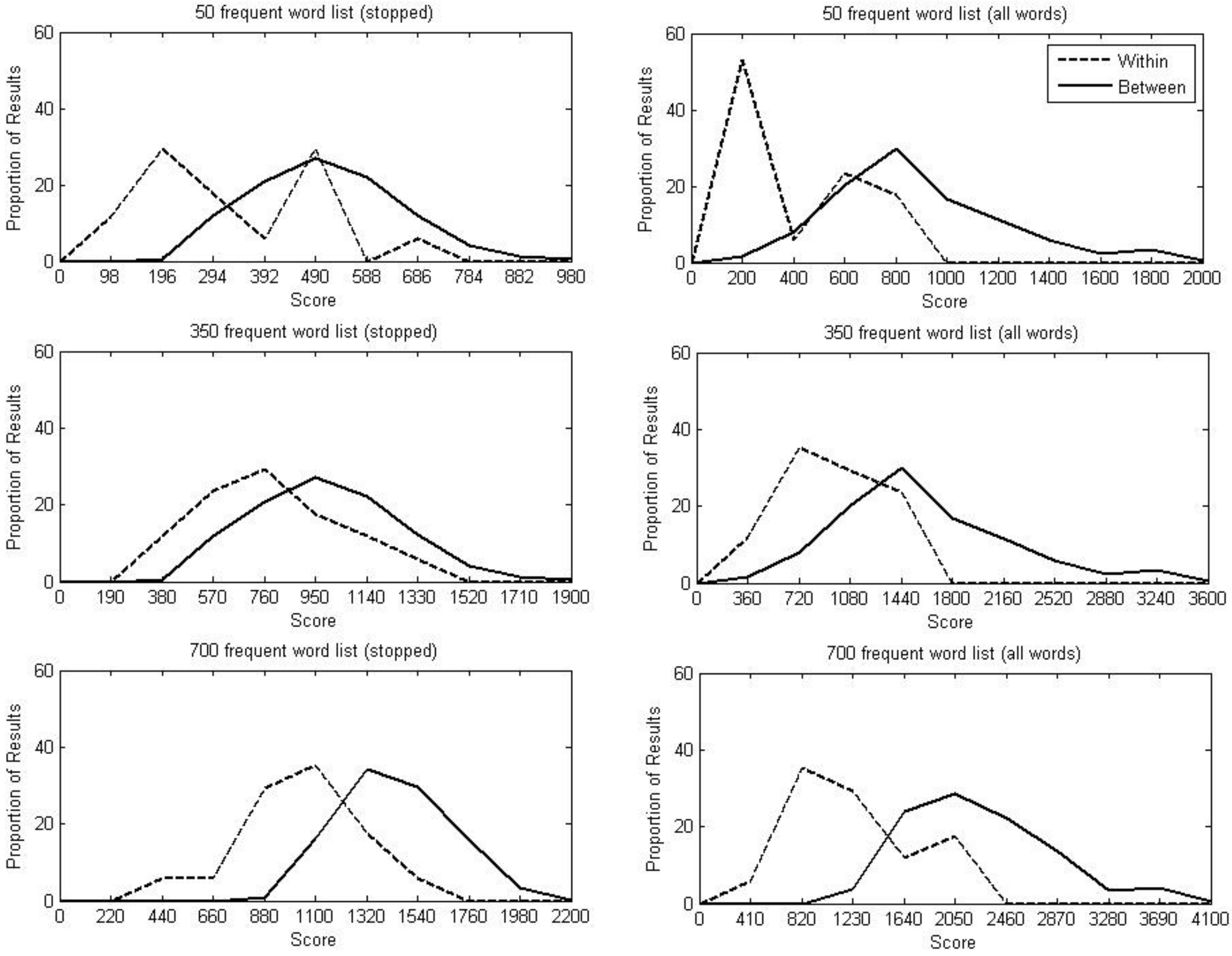


Figure 17: Within and between comparisons for 10000 word sample corpora with 50, 350 and 700 most common words

Figure 18 shows the correlations between the participants' overall similarity scores and the chi-square results. A general pattern was observed for both participants. Correlations were highest for the 10000 *all words* sample, and these correlations also improved as the number of words in the word frequency list increased. Interestingly, once the number of words in the word frequency list exceeded 450 words, the 1000 *all words* sample had similar correlations to the 10000 *all words* sample.

Further statistical analyses were conducted for content, structure and language for Participant A and B (See Appendix D). These results were generally consistent with those observed above; the 10000 *all words* sample yielded the highest correlations, followed by the 1000 *all words* sample. The exception to this was observed with the structure and language score of Participant A, in which the correlation for 1000 *all words* was slightly higher than the correlation for 10000 *all words,* but this was only the case when there were more than 450 words in the word frequency lists.

Given these findings and observations it was deemed appropriate to use the 10000 word sample, with the 700 most common words, and *all words* included, as a comparison between statistical and human corpora comparisons.

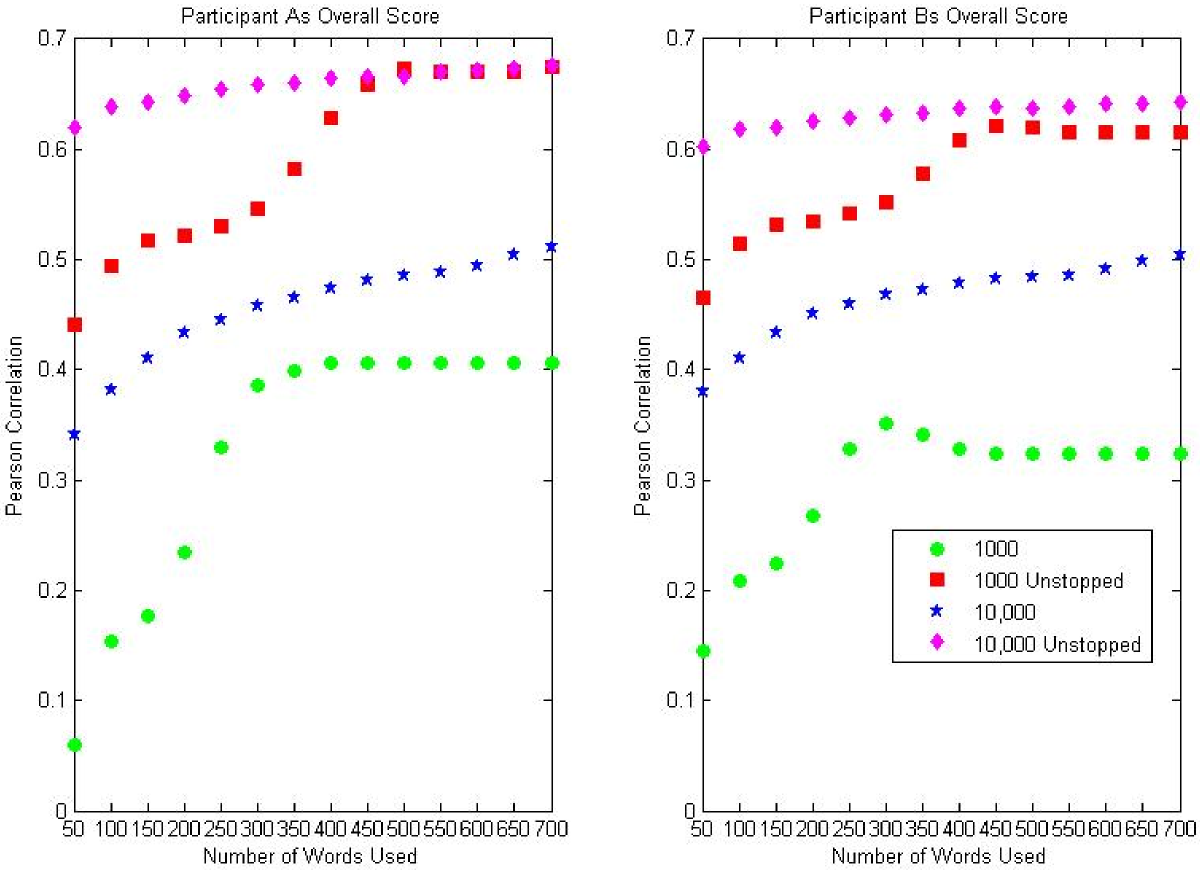


Figure 18: Graph showing correlations between the participants' similarity scores and the chi square results

4.3 The Chi-Square Results

In order to further examine the results provided by the chi-square statistic, MDS was applied to the KSC values in order to visualise the data. This approach is detailed in Butavicius et al. (2009). As shown in Section 3.7.1, this displays the results in two dimensions, providing a visual inspection of the relationship between the categories, and between the two halves of the same corpus. The colour and shape combination indicates the corpus half and the colour indicates the category. In this display, the distance between the symbols indicates similarity with more similar corpus halves (as judged by the chi-square technique) placed nearer to each other in space.

The MDS display in Figure 19 clearly indicates that the different categories tended to be faithfully clustered. Like Participant A, the chi-square value tended to group the academic and speech categories together. The results of the statistical technique also replicated the human ratings in regards to the SMS corpus, which was judged to be quite different to all other corpora. Furthermore, for both the human and chi-square results, the newspaper corpora were all very closely grouped, suggesting less variation in the corpora within that category. Furthermore, in the vast majority of cases, the different halves of the same corpus were very closely grouped. Taken together, these results suggest that the chi-square values can indicate with a high level of accuracy whether an unknown corpus is similar or not to any previously assessed corpora. This could also be extremely useful for determining the applicability of previous studies on different corpora. For example, the MDS display indicates a high level of similarity between the BNC conversation corpus and the other speech and academic corpora. Hence, the findings of a study utilising the BNC conversation corpus could be generalised to the other assessed speech and academic corpora with reasonable confidence. In contrast, since the SMS corpus was highly dissimilar to the other corpora, the findings would be unlikely to be generalisable to other corpora. This is consistent with a priori expectations about the unique nature of linguistic use in SMS communications.

These findings suggest that the chi-square results were both reliable and valid. The results provided by the X^2 statistic will now be compared to the human results in more detail.

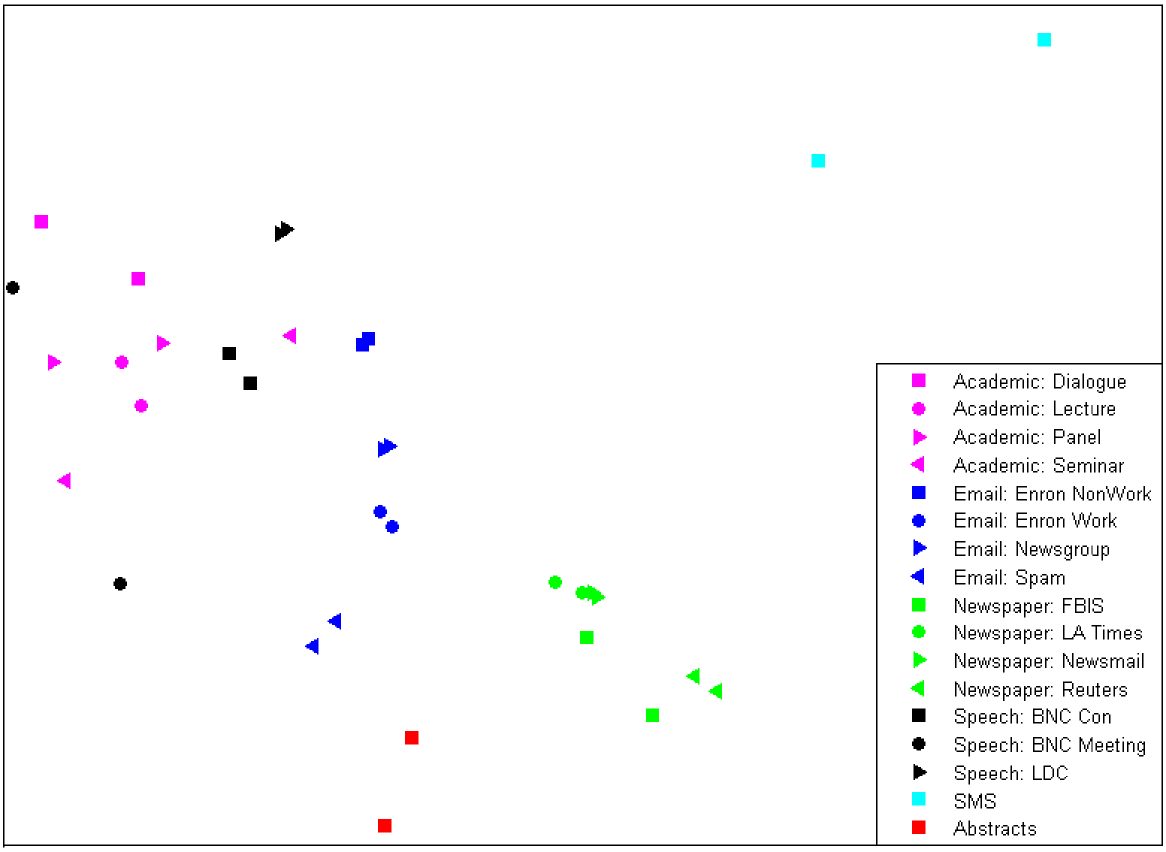


Figure 19: MDS display for 10000 word corpora sample (all words), 700 frequent words

5. Results – Part 3: Comparisons Between the X^2 Statistic and Human Corpora Similarity Judgements

5.1 Summary of Results

The similarity judgements provided by the two participants were compared to the chi-square results using the 10000 word corpora sample, with 700 words in the frequency lists, and all unique words included. Participants' judgements in relation to content similarity, structural similarity, language use similarity and overall similarity were all compared to the chi-square result, with the aim of determining whether the results provided by the statistical technique were more representative of a certain aspect of the participants' ratings.

Part 1 of the Results Section indicated very high agreement between the human judges. However, due to the individual differences that exist between people, averaging between individuals can result in values that do not capture a faithful human representation of the space (Ashby, Maddox & Lee, 1994; Lee & Pope, 2003). Hence, rather than averaging the results, instead, both participants' scores were independently compared to the results provided by the chi-square technique.

As shown in Table 11, the level of agreement between both participants' judgements and the chi-square results was quite high. Interestingly, there was much more agreement between the chi-square results and the content similarity scores when corpora were from the same category than from a different category. In contrast, particularly for Participant A, the correlation between language use similarity and the chi-square results was far higher for the corpora from different categories, and far lower for the corpora from the same category.

Table 11: Correlations between Participant A and Participant B

	Different Categories		**Same Categories**		**All Results**	
	A	**B**	**A**	**B**	**A**	**B**
Content Similarity	0.46	0.38	0.57	0.58	0.62	0.57
Structural Similarity	0.44	0.42	0.35	0.47	0.62	0.63
Language Use Similarity	0.51	0.50	0.26	0.45	0.66	0.65
Overall Similarity	**0.52**	**0.47**	**0.48**	**0.49**	**0.68**	**0.64**

5.2 Overall Similarity Judgements

The agreement between the participants' overall similarity scores and the statistical technique was high, with a correlation of 0.68 for Participant A, and 0.64 for Participant B (N = 561).

As mentioned previously, participants made similarity judgements on a seven point scale. To allow a direct comparison between the range of scores provided by the participants and the range of scores given by the chi-square technique, the chi-square scores were transferred into a seven point scale. The maximum score was divided by seven to obtain the appropriate intervals, and the data was then transformed, so that results within the first interval were assigned a score of one, results in the second interval were assigned a score of two, and so on, till the scores in the highest interval, which were assigned a score of seven. This allows a comparison between the response frequencies assigned by human judgements and the scores obtained by the chi-square technique.

Of the 561 comparisons, Participant A had identical ratings to the chi-square technique on 86 occasions, and 56% of the ratings (335) were identical or within one point. 105 of Participant B's responses were identical, and 303 (54%) ratings were identical or within one point. These scores can be seen in the frequency table in Appendix E (Section E1), and are graphically displayed in the response frequency distribution in Figure 20. These results indicate that participants were more likely to view corpora pairs as dissimilar, and there were also more occasions when the participants judged corpora pairs to be highly similar. In contrast, the chi-square values had a more normal distribution, with very few extremely similar or extremely dissimilar results. Hence, this indicates that the chi-square is more conservative, and is less likely to output values at extremes of similarity.

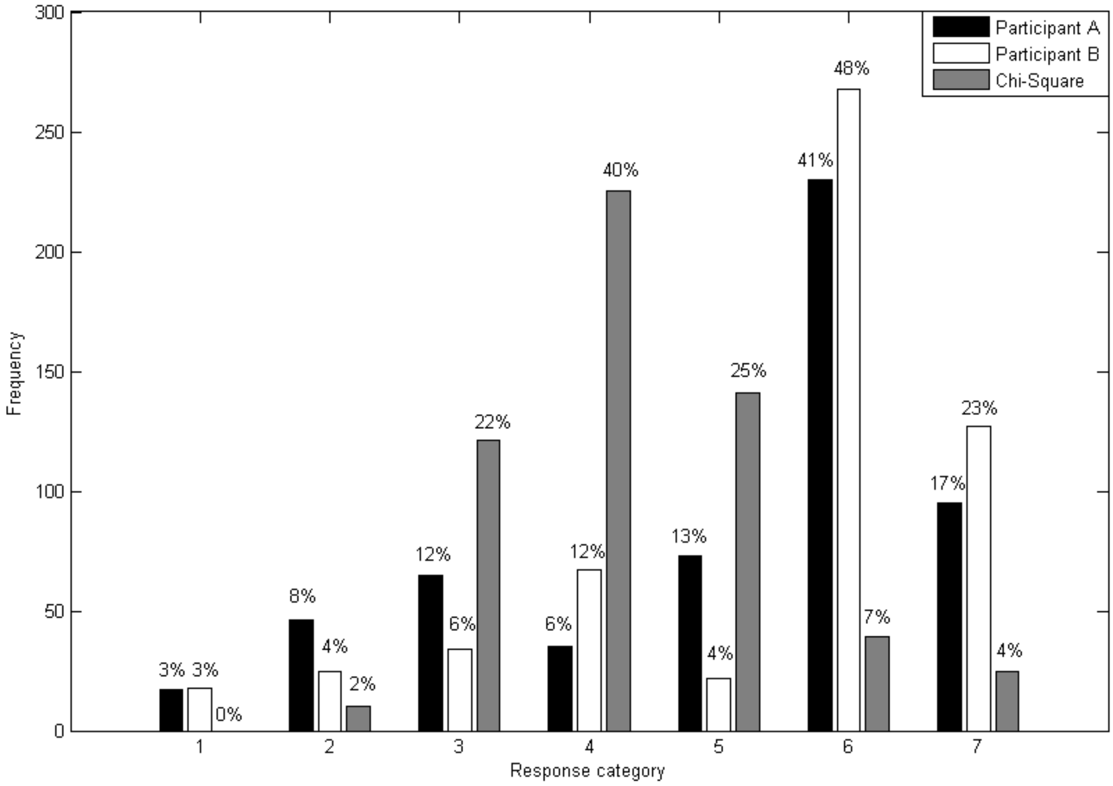


Figure 20: Response frequency distribution for Participant A and B's overall similarity ratings

This is also supported by the histogram in Appendix E (Section E1), which shows how much participants' scores differ from the chi-square values. The histogram clearly indicates a positive skew in the distribution, with the participants' responses more likely to be at the extremes of the response range, and the chi-square values more likely to fall within the midpoint of the response range. Despite that, the comparative matrixes in Figure 20 still indicate a positive distribution, suggesting a strong relationship between participants' overall similarity ratings and the values obtained using the chi-square technique.

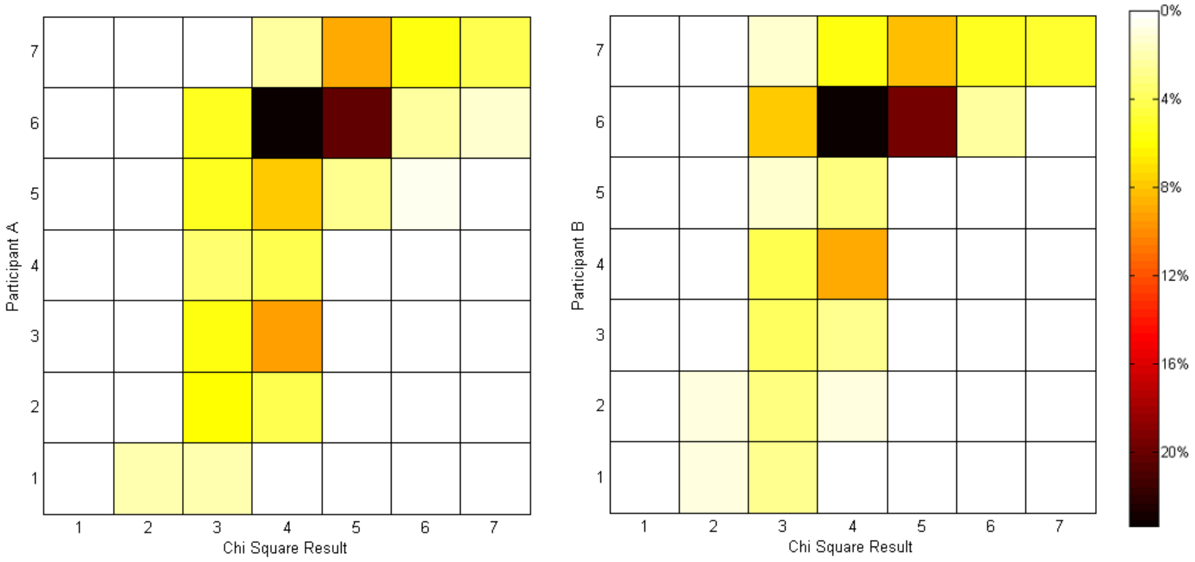


Figure 21: Comparative matrix for overall similarity for Participant A (left) and Participant B (right) with the chi-square results

5.3 Content Similarity Judgements

The correlation between the values provided by the chi-square technique and Participant A's content similarity judgements were high, with a correlation of 0.62 (*N* = 561). As indicated in the frequency table in Appendix E (Section E2), Participant A provided a score identical to that of the chi-square technique on 53 occasions. In addition, there were 324 occasions where the ratings were identical or within one point, which equates to 58% of all ratings.

There was a similar pattern for the chi-square value and Participant B's content similarity score, with a correlation of 0.57 (*N* = 561). Although this is a lower correlation than that of Participant A, the frequency table in Appendix E (Section E2) reveals that there were more occasions when Participant B provided an identical response to the chi-square technique. Participant B's response was identical on 89 occasions (or 16% of responses), but there were only 201 occasions when Participant B's ratings were identical or within one point of the chi-square ratings.

As found for the overall similarity ratings, the histogram in Appendix E (Section E2) showing the differences between the scores provided by the participants and the chi-square values indicates that the participants tended to assign similarity scores that were one or two points higher than the chi-square technique, suggesting that the human ratings were more likely to focus on the dissimilarity between two corpora. Despite this, the comparative matrixes in Figure 22 indicate a strong positive relationship between the participants' content similarity scores and the chi-square results.

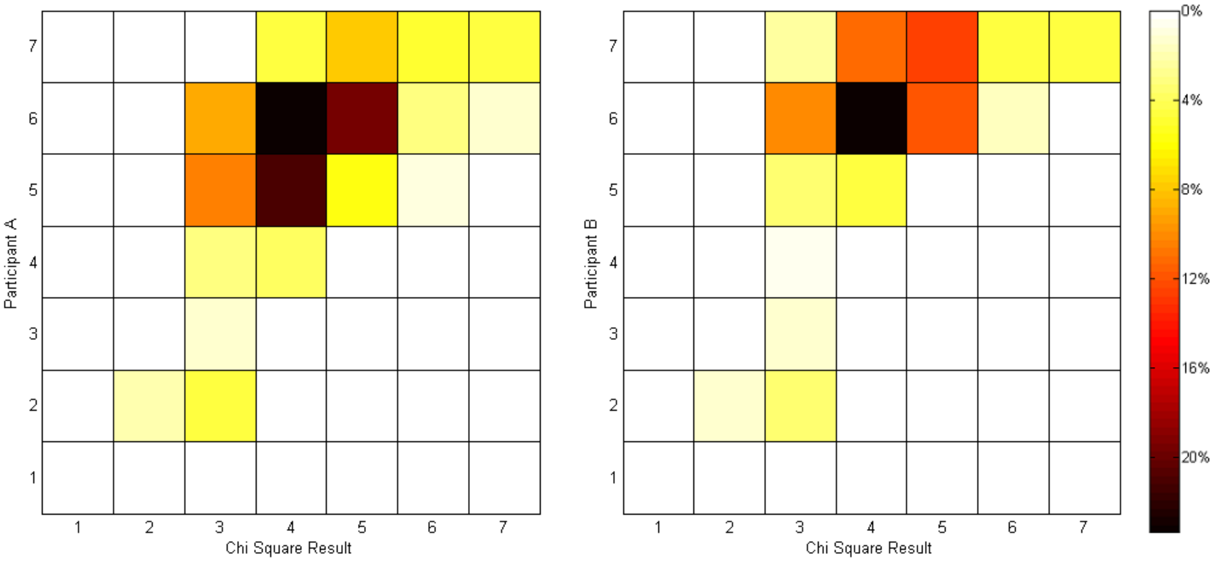


Figure 22: Comparative matrix for content similarity for Participant A (left) and Participant B (right) with the chi-square results

5.4 Structural Similarity Judgements

As shown in the comparative matrixes in Figure 23, there was also a strong positive relationship between the structural similarity scores assigned by both participants and the chi-square values. It is, however, necessary to note that the chi-square values tended to be more conservative, with few scores at the extremes of the distribution, whereas the participants were more likely to provide similarity ratings that were highly similar or highly dissimilar.

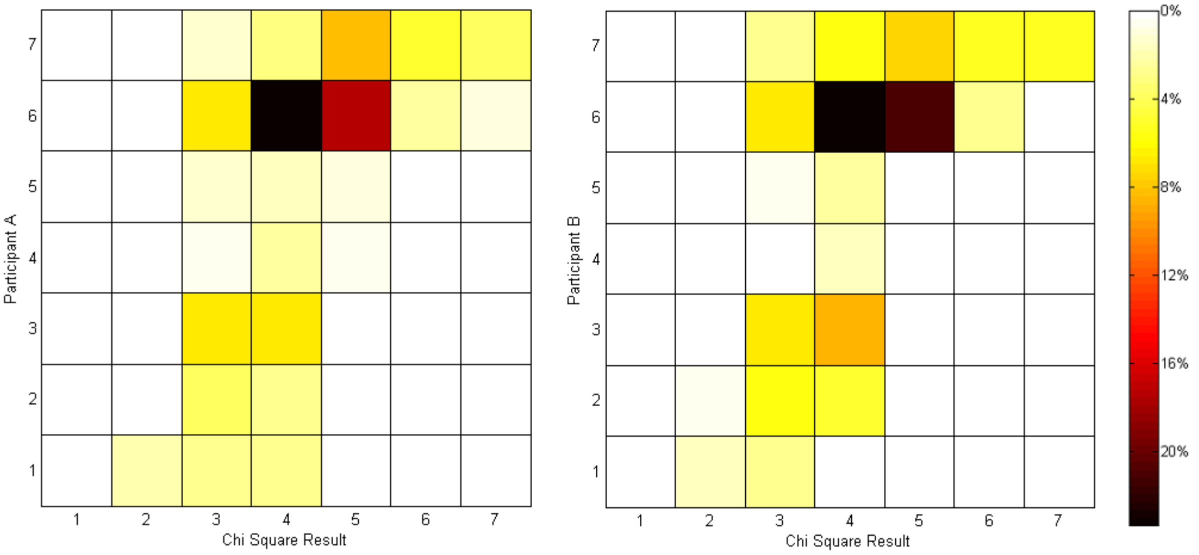


Figure 23: Comparative matrix for structural similarity for Participant A (left) and Participant B (right) with the chi-square results

Participant A's structural similarity judgements and the chi-square values had a correlation of 0.62 ($N = 561$), and there were 85 cases (15.15%) in which the identical response was assigned to corpora pairs. Participant B's structural similarity judgements and the chi-square values had a similar correlation [$r = 0.63$, $N = 561$], and there were 80 cases (14.26%) where the value was identical. When the ratings that were within one point were also included, more than 50% of scores were identical for both participants [Participant A = 50.98% | Participant B = 52.76%]. These findings can be viewed in Appendix E (Section E3).

5.5 Language Use Similarity Judgements

For both participants, the relationship between the chi-square values and the assigned similarity scores was highest for ratings of language use [Participant A: $r = 0.66$ | Participant B: $r = 0.65$]. However, the histogram showing the differences in scores in Appendix E (Section E4) indicates that participants were more likely to view corpora pairs as less similar than the chi-square technique, and the chi-square values were generally more conservative than the participants' ratings. Despite that, the comparative matrixes in Figure 24 clearly show a strong relationship, with a large amount of agreement between the participants' scores and the chi-square values.

The chi-square values were identical to the participants' ratings in 80 cases (14.26%) for Participant A and 101 cases (18%) for Participant B. When the scores within one point were also included, this increased to 301 cases (53.65%) for Participant A and 308 cases (54.90%) for Participant B. These findings can be seen in the frequency tables in Appendix E (Section E4).

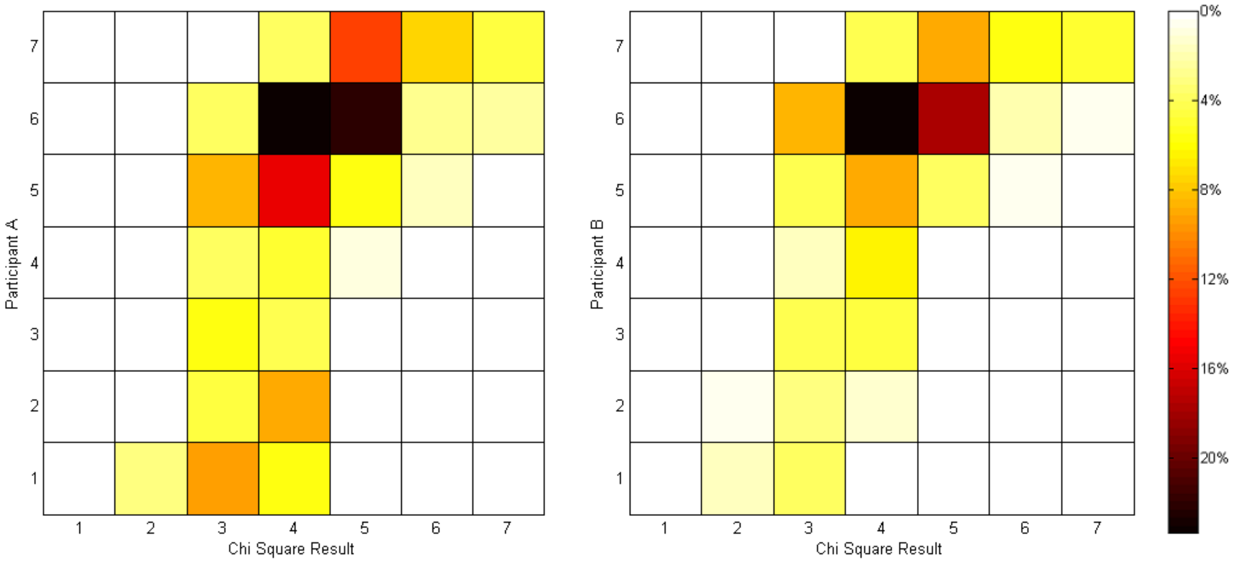


Figure 24: Comparative matrix for language use similarity for Participant A (left) and Participant B (right) with the chi-square results

5.6 The Influence of Category on Similarity Judgements

The results were also analysed according to whether the corpora in an examined pair were from the same or different categories. As shown in Table 12, there was a great deal of variation in the strength of the correlations between the human ratings and the chi-square values. This variation was inconsistent across the different categories and the different aspects of the humans' similarity judgements.

For example, for the academic corpora, the correlation between the chi-square values and Participant A's language use ratings was only 0.22 (which indicates a weak relationship), whereas the correlation with Participant B's language use ratings was highly significant, at 0.60 (which indicates a strong relationship). In contrast, Participant A's overall similarity score for the newspaper corpora was highly significant, at 0.65, whereas Participant B's overall similarity score was only 0.18. For the email corpora, the correlations for both participants for all dimensions of similarity were highly statistically significant. This suggests that the chi-square technique could be more effective when judging the similarity of certain types of corpora.

It is, however, necessary to note that the number of comparisons within the different categories was small, and therefore the results should be interpreted with caution. Furthermore, since the number of comparisons in the '*same category*', '*different category*' and '*between*' groups was very large, the significance of those correlations should also be interpreted cautiously. An analysis of the effect size indicates that the correlations between the chi-square values and the participants' overall similarity scores for the same and different category pairs accounted for between 22% and 27% of the variance. For the between comparisons, the correlation between the chi-square values and the overall similarity scores accounted for 41% of variance for Participant A and 36% of variance for Participant B.

Table 12: Correlations by Category for Participant A and Participant B compared to the Chi-Square Results

		Participant A				*Participant B*			
Category	**Number**	**Content**	**Structure**	**Language Use**	**Overall**	**Content**	**Structure**	**Language Use**	**Overall**
Academic	28	0.476*	0.296	0.219	0.310	0.345	0.241	0.603**	0.375*
Newspaper	28	0.215	0.606**	0.205	0.653**	0.151	0.376*	0.321	0.179
Speech	15	-0.082	0.544*	0.425	0.613*	0.467	0.382	0.381	0.568*
Email	28	0.755**	0.494**	0.702**	0.669**	0.650**	0.534**	0.619**	0.605**
Same Category	101	0.572**	0.348**	0.262**	0.482**	0.581**	0.469**	0.454**	0.487**
Different Category	460	0.462**	0.436**	0.514**	0.522**	0.382**	0.419**	0.503**	0.467**
Within	17	0.496*	0.615**	0.765**	0.678**	0.625**	0.628**	0.532*	0.494*
Between	544	0.587**	0.584**	0.624**	0.640**	0.509**	0.585**	0.608**	0.603**

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

6. Conclusions

This report has highlighted the subjective nature of corpora comparison, indicating that the similarity between corpora can be measured on a number of dimensions. There was a high level of consistency between the participants in this study, with a correlation for overall similarity of 0.88. However, although consistency in relation to overall similarity was very high, further analysis of the results indicated that the participants differed in the manner in which they judged certain corpora. For example, participants had a high level of agreement when judging content similarity for corpora from the same category, and when judging structural similarity for corpora from different categories. In contrast, there was far less agreement when participants were judging content similarity for corpora from different categories and structural similarity for corpora from the same category. Participants also differed in the dimension of similarity that had the greatest influence on their overall score. Participant A's language use similarity score was most predictive of overall similarity, whereas Participant B's overall similarity score was more strongly influenced by judgements of structural similarity.

This report also provides further support to the use of the chi-square statistic as a measure of corpora comparison. The correlations between the participants' ratings of similarity and the chi-square results were high. The chi-square values accounted for between 41 and 46% of shared variance in the overall similarity provided by Participant B and Participant A respectively. However, the results indicate that the chi-square values were more conservative, and were more likely to fall into a normal distribution, with few very similar ratings and few very different ratings. In contrast, participants were more likely to focus on the difference between two corpora, and were also more likely to judge corpora to be highly similar. Since these highly similar and highly dissimilar judgements are likely to be far more useful, there may be limits to the practical value of the chi-square technique.

There are also complexities associated with the chi-square technique. Essentially, the optimal size of a corpora sample and the optimal number of words for the word frequency lists are unknown, and evidence suggests that these configurations can greatly influence the effectiveness of the technique. In this study, the human ratings were used as an objective standard to find the best combination of corpora size and number of words used.

Various sized corpora samples and most frequent word lists were tested empirically, and it was found that the 10000 word sample with the 700 most frequent words was the best performing, and these are recommended for future applications of this corpora comparison technique. However, this finding has the caveat that corpora with radically different word frequency distributions to those tested here may result in variation in the effectiveness of the chi-square technique under these parameter settings. Furthermore, it is necessary to note that the human judgements used in this study may not be representative of all human judgements. Therefore, it is important to repeat this study with a larger sample size to confirm the findings.

7. References

Ashby, F.G., Maddox, W.T., & Lee, W.W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science, 5(3),* 144-151.

Blankenship, J. (1962). A linguistic analysis of oral and written style. *Quarterly Journal of Speech*, *48*, 419-22

Butavicius, M.A., Ferguson, L., & Mullen, L.G. (2009). *An application of Kilgarriff's (2001) corpora comparison approach: classified versus unclassified documents*. INT xx/xxxx

Butavicius, M.A. & Lee, M.D. (2007). An empirical evaluation of four data visualization techniques for displaying short news text similarities, *International Journal of Human–Computer Studies*, *65 (11)*, 931-944.

Butavicius, M.A., Lee, M.D., Pincombe, B.M. & Mullen, L.G. (2006). *An empirical evaluation of four document visualization techniques for displaying spontaneous language extracts*. INT 04/249.10

Chafe, W.L. (1979). Integration and involvement in spoken and written language. *Proceedings of the Second World Congress of the International Association for Semiotic Studies*, Vienna, July 1979.

Cox, T.F. & Cox, M.A.A. (1994). *Multidimensional scaling*. London: Chapman and Hall.

Drieman, G.H.J. (1962). Differences between written and spoken language. *Acta Psychologica*, *20*, 36-58.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics 6(1)*, 1-37.

Lee, M.D., Butavicius, M.A. & Reilly, R.E. (2003). Visualizations of binary data: A comparative evaluation. *International Journal of Human-Computer Studies*, *59*, 569-602.

Lee, M.D., Pincombe, B.M., & Welsh, M.B., (2005). An empirical evaluation of models of text document similarity, In B.G. Bara, L. Barsalou & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Meeting of the Cognitive Science Society (pp. 1254-1259)*. Cognitive Science Society: Austin, TX,.

Lee, M.D., Pope, K.J. (2003). Avoiding the dangers of averaging across subjects when using multidimensional scaling. *Journal of Mathematical Psychology*, *47,* 32-46.

Leech, G. & Fallon, R. (1992). Computer corpora – what do they tell us about culture? *ICAME Journal*, *16*, 29-50.

Parsons, K., McCormac, M. & Butavicius, M. (2009). *The Use of an Email Corpus to Empirically Evaluate Data Visualisation*. Manuscript submitted for publication.

Poole, M.E. & Field, T.W. (1976). A comparison of oral and written code elaboration. *Language and Speech*, *19*, 305-311.

Tang, R., Shaw, W.S. & Vevea, J.L. (1999). Towards the identification of the optimal number of relevance categories, *Journal of the American Society for Information Science*, *50(3)*, 254-264.

Voorhees, E.M. & Harman, D. (2000). Overview of the Eighth Text REtrieval Conference (TREC-8), In E.M. Voorhees & D.K. Harman (Eds.), *NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC-8) (pp.1-24)* National Institute of Standards and Technology: Maryland, USA.

Appendix A: Statistics of the Corpora Samples

#[4]	Corpus	Category	Part	Word Length	Unique Words (stopped)	Unique Words (all words)	Characters (no spaces)	Characters (with spaces)	Docs	Paragraphs	Lines
1	Enron (non-work)	Email	2	1028	276	428	4177	5250	6	27	82
2	LA Times	Newspaper	1	943	396	501	4627	5569	8	7	65
3	MICASE (dialogue)	Academic	1	996	232	369	4579	5573	1	29	98
4	MICASE (panel)	Academic	2	1007	207	327	4588	5590	1	7	62
5	MICASE (dialogue)	Academic	2	1018	215	332	4966	5983	1	19	87
6	Enron (work)	Email	2	973	335	463	4692	5658	6	30	100
7	MICASE (panel)	Academic	1	1034	222	350	4234	5243	1	48	127
8	BASE (seminar)	Academic	2	1016	192	328	4359	5358	1	29	96
9	Reuters	Newspaper	2	970	330	433	4945	5881	5	34	107
10	LDC	Speech	1	1090	219	360	4680	5765	1	76	171
11	Newsmail	Newspaper	1	937	440	533	4949	5877	10	10	69
12	BNC (meeting)	Speech	1	1067	175	302	4173	5238	1	89	196
13	BNC (meeting)	Speech	2	1061	190	335	4309	5367	1	14	71
14	BNC (conversation)	Speech	1	1026	183	322	4286	5312	1	23	85
15	SPAM	Email	2	998	355	475	5102	6108	6	82	137
16	BNC (conversation)	Speech	2	1052	246	385	4428	5478	1	22	83
17	LA Times	Newspaper	2	918	390	476	4689	5568	9	16	74
18	BASE (seminar)	Academic	1	1013	152	291	4048	5024	1	53	136
19	FBIS	Newspaper	1	1002	378	466	5307	6307	7	30	106
20	SMS	Other	2	1009	399	527	3874	4812	78	78	166

[4] The corpora were examined according to this ordering

#[4]	Corpus	Category	Part	Word Length	Unique Words (stopped)	Unique Words (all words)	Characters (no spaces)	Characters (with spaces)	Docs	Paragraphs	Lines
21	Newsmail	Newspaper	2	955	403	520	5019	5964	11	11	72
22	MICASE (lecture)	Academic	2	1064	218	332	4967	6028	1	13	73
23	Reuters	Newspaper	1	1004	363	456	5020	5993	6	31	104
24	Enron (non-work)	Email	1	1020	286	429	4117	5153	6	22	81
25	Enron (work)	Email	1	980	322	426	4901	5885	7	25	94
26	SPAM	Email	1	973	331	450	5035	5975	6	98	152
27	LDC	Speech	2	1012	176	315	4316	5321	1	54	133
28	Abstracts	Other	2	1042	378	472	5997	7047	7	7	77
29	Newsgroup	Other	1	1026	344	498	4829	5829	10	76	130
30	FBIS	Newspaper	2	992	327	413	5329	6318	7	23	94
31	Newsgroup	Other	2	1019	329	472	4857	5894	7	77	133
32	SMS	Other	1	1001	412	536	4001	4923	79	79	170
33	Abstracts	Other	1	1050	362	458	5937	6996	7	7	75
34	MICASE (lecture)	Academic	1	1052	236	378	4279	5331	1	11	65

Appendix B: Multidimensional Scaling Displays for the Human Similarity Judgements

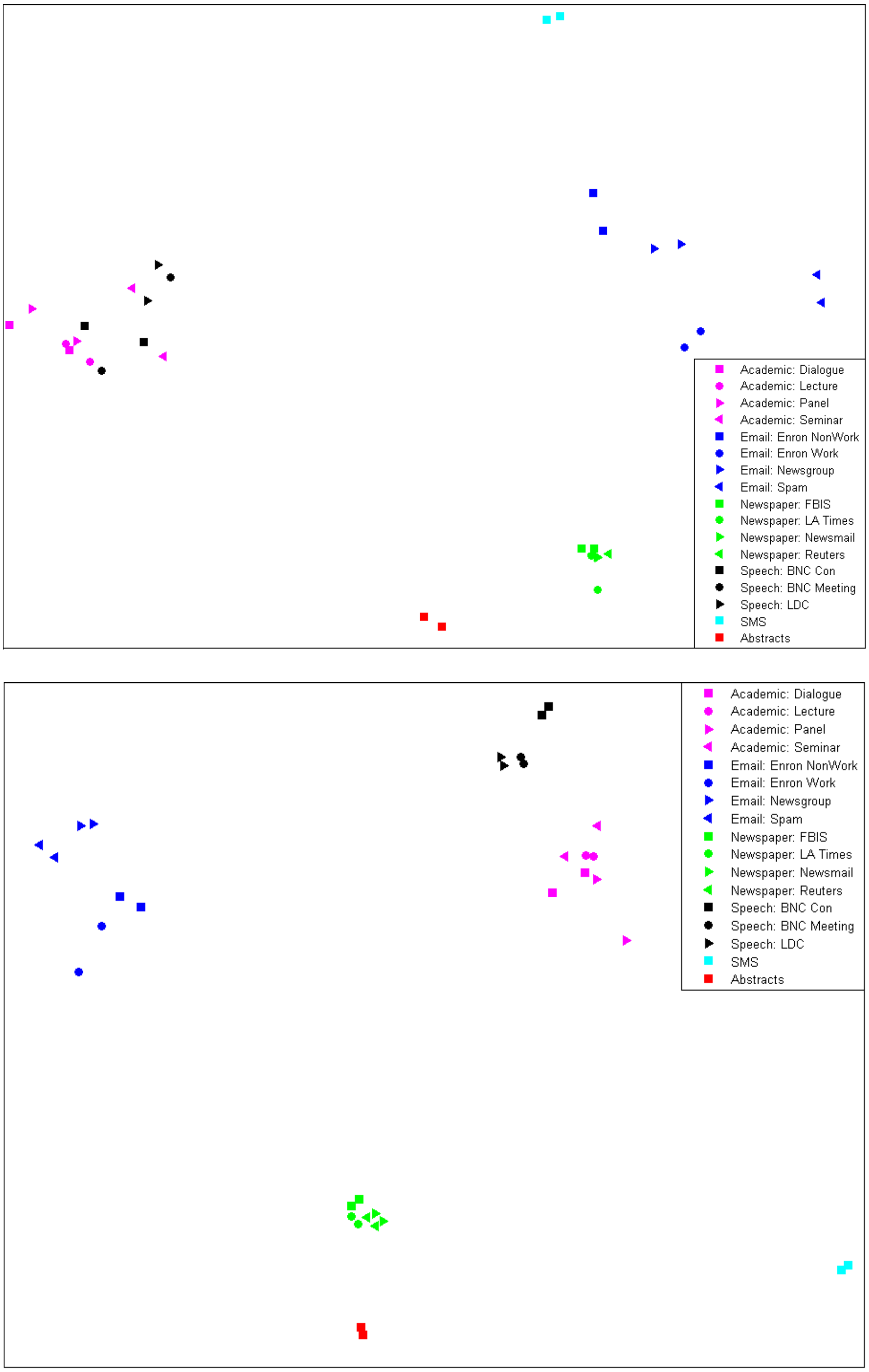
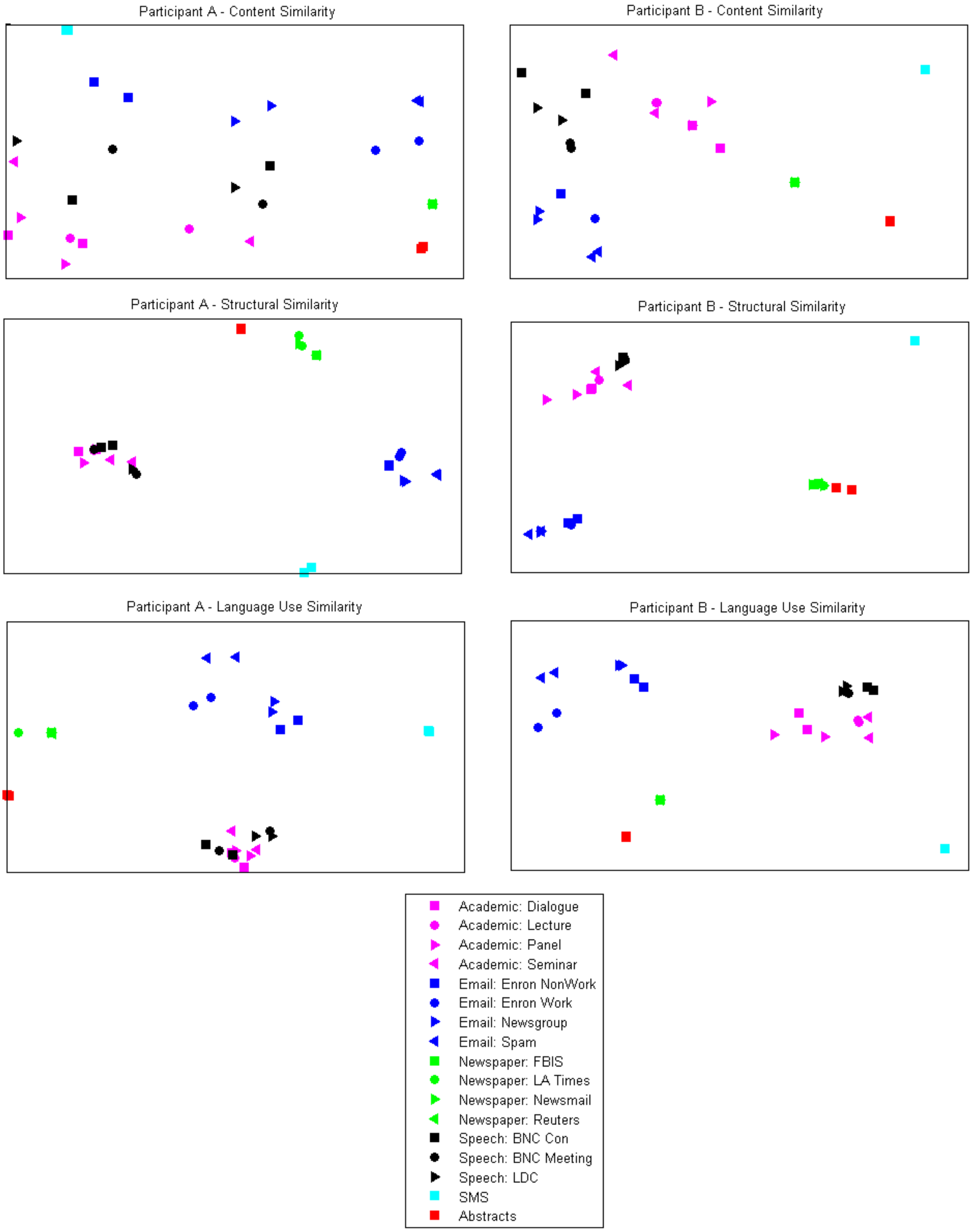


Figure B1: MDS displays for Participant A (top) and Participant B (bottom)



Participant A - Content Similarity

Participant B - Content Similarity

Participant A - Structural Similarity

Participant B - Structural Similarity

Participant A - Language Use Similarity

Participant B - Language Use Similarity

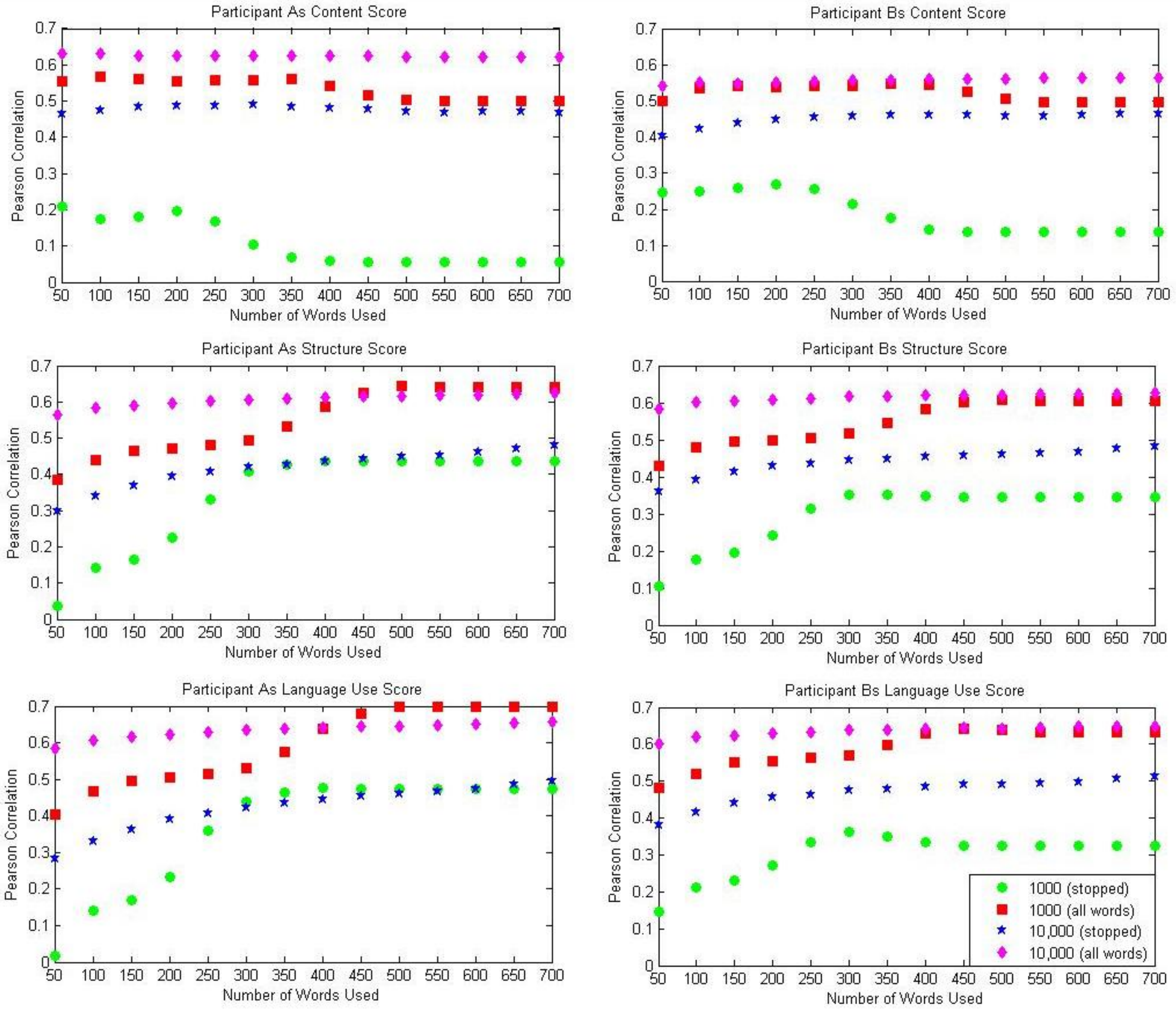
Figure B2: MDS displays for Participant A (left) and Participant B (right)

Appendix C: Stop List

*
0
1
2
3
4
5
6
7
8
9
a
about
above
accordingly
across
after
afterwards
again
against
al
al.
all
allows
almost
alone
along
already
also
although
always
am
among
amongst
an
and
another
any
anybody
anyhow
anyone
anything
anywhere
apart
appear
appropriate
are
around
as
aside
associated
at
available
away
awfully
b
back
be
became
because
become
becomes
becoming
been
before
beforehand
behind
being
below
beside
besides
best
better
between
beyond
both
brief
but
by
c
came
can
cannot
cant
cause
causes
certain
changes
co
come
consequently
contain
containing
contains
corresponding
could
currently
d
day
described
did
different
do
does
doing
done
down
downwards
during
e
each
eg
eq
e.g
eight
either
else
elsewhere
enough
et
etc
even
ever
every
everybody
everyone
everything
everywhere
ex
example
except
f
far
few
fifth
first
five
followed
following
for
former
formerly
forth
four
from
further
furthermore
g
get
gets
given
gives
go
gone
good
got
great
h
had
hardly
has
have
having
he
hence
her
here
hereafter
hereby
herein
hereupon
hers
herself
him
himself
his
hither
how
howbeit
however
i
ie
i.e
if
ignored
immediate
in
inasmuch
inc
indeed
indicate
indicated
indicates
inner
insofar
instead
into
inward
is
it
its
itself
j
just
k
keep
kept
know
l
last
latter
latterly
least
less
lest
let
life
like
little
long
ltd
m
made
make
man
many
may
me
meanwhile
men
might
more
moreover
most
mostly
mr
much
must
my
myself
n
name
namely
near
necessary
neither
never
nevertheless
new
next
nine
no
nobody
none
noone
nor
normally
not
nothing
novel
now
nowhere
o
of
off
often
oh
old
on
once
one
ones
only
onto
or
other
others
otherwise
ought
our
ours
ourselves
out
outside
over
overall
own
p
particular
particularly
people
per
perhaps
placed
please
plus
possible
probably
provides
q
que
quite
r
rather
really
relatively
respectively
right
s
said
same
second
secondly
see
seem
seemed
seeming
seems
self
selves
sensible
sent
serious

seven
several
shall
she
should
since
six
so
some
somebody
somehow
someone
something
sometime
sometimes
somewhat
somewhere
specified
specify
specifying
state
still
sub
such
sup
t
take
taken
than
that
the
their
theirs
them
themselves
then
thence
there
thereafter
thereby
therefore
therein
thereupon
these
they
third
this
thorough
thoroughly
those
though
three
through
throughout
thru
thus
time
to
together
too
toward
towards
twice
two
u
under
unless
until
unto
up
upon
us
use
used
useful
uses
using
usually
v
value
various
very
via
viz
vs
w
was
way
we
well
went
were
what
whatever
when
whence
whenever
where
whereafter
whereas
whereby
wherein
whereupon
wherever
whether
which
while
whither
who
whoever
whole
whom
whose
why
will
with
within
without
work
world
would
x
y
year
years
yet
you
your
yours
yourself
yourselves
z
zero

Appendix D: Correlations between the participants' similarity scores and the chi square results



Participant As Content Score

Participant Bs Content Score

Participant As Structure Score

Participant Bs Structure Score

Participant As Language Use Score

Participant Bs Language Use Score

THIS PAGE HAS BEEN
INTENTIONALLY LEFT BLANK

Appendix E: Frequency Tables & Histograms

E.1. Overall Similarity Ratings

		‘Chi-Square Results’							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**TOTAL**
	1	**0**	9	8	0	0	0	0	**17**
	2	0	**1**	27	18	0	0	0	**46**
	3	0	0	**24**	40	1	0	0	**65**
‘A’	**4**	0	0	15	**19**	1	0	0	**35**
	5	0	0	23	35	**12**	3	0	**73**
	6	0	0	23	102	88	**11**	6	**230**
	7	0	0	1	11	39	25	**19**	**95**
	TOTAL	**0**	**10**	**121**	**225**	**141**	**39**	**25**	**561**

		‘Chi-Square Results’							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**TOTAL**
	1	**0**	5	13	0	0	0	0	**18**
	2	0	**5**	15	5	0	0	0	**25**
	3	0	0	**19**	14	1	0	0	**34**
‘B’	**4**	0	0	21	**45**	1	0	0	**67**
	5	0	0	6	16	**0**	0	0	**22**
	6	0	0	40	117	98	**12**	1	**268**
	7	0	0	7	28	41	27	**24**	**127**
	TOTAL	**0**	**10**	**121**	**225**	**141**	**39**	**25**	**561**

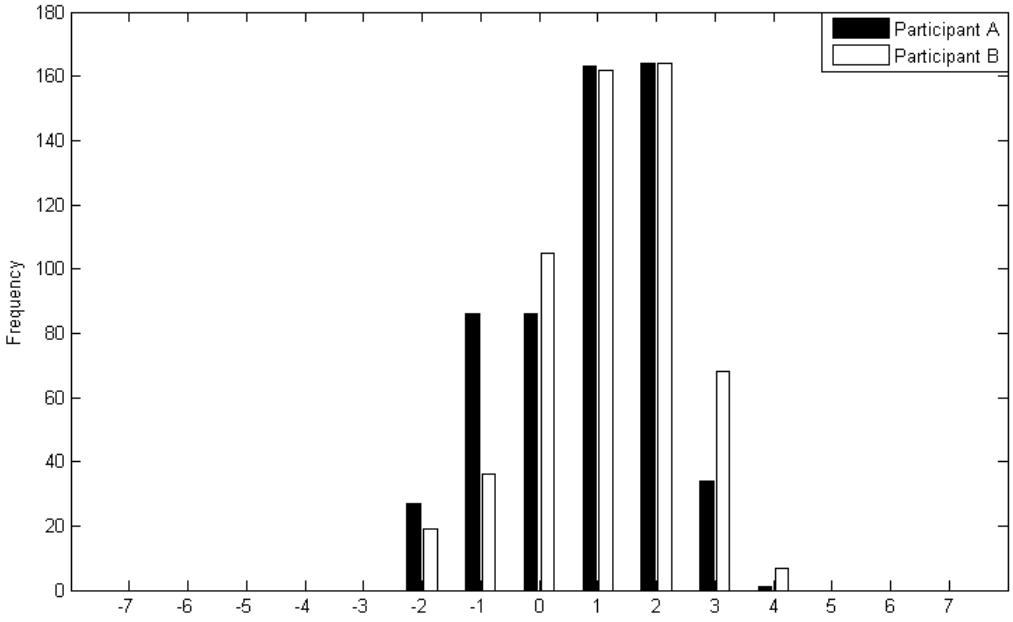


Figure E1: Frequency tables comparing the chi-square result and overall similarity for Participant A (top), and Participant B (middle), and a histogram of the difference for ‘A’, ‘B’ and the chi-square result (bottom)

E.2. Content Similarity Ratings

		'Chi-Square Results'							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**TOTAL**
	1	**0**	0	1	0	0	0	0	**1**
	2	0	**9**	19	0	0	0	0	**28**
	3	0	0	**5**	0	0	0	0	**5**
'A'	**4**	0	0	13	**16**	0	0	0	**29**
	5	0	1	45	90	**25**	4	0	**165**
	6	0	0	38	100	83	**14**	5	**240**
	7	0	0	0	19	33	21	**20**	**93**
	TOTAL	**0**	**10**	**121**	**225**	**141**	**39**	**25**	**561**

		'Chi-Square Results'							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**TOTAL**
	1	**0**	0	0	0	0	0	0	**0**
	2	0	**8**	19	0	0	0	0	**27**
	3	0	2	**7**	1	0	0	0	**10**
'B'	**4**	0	0	3	**1**	0	0	0	**4**
	5	0	0	20	26	**2**	2	0	**50**
	6	0	0	58	133	68	**10**	0	**269**
	7	0	0	14	64	71	27	**25**	**201**
	TOTAL	**0**	**10**	**121**	**225**	**141**	**39**	**25**	**561**

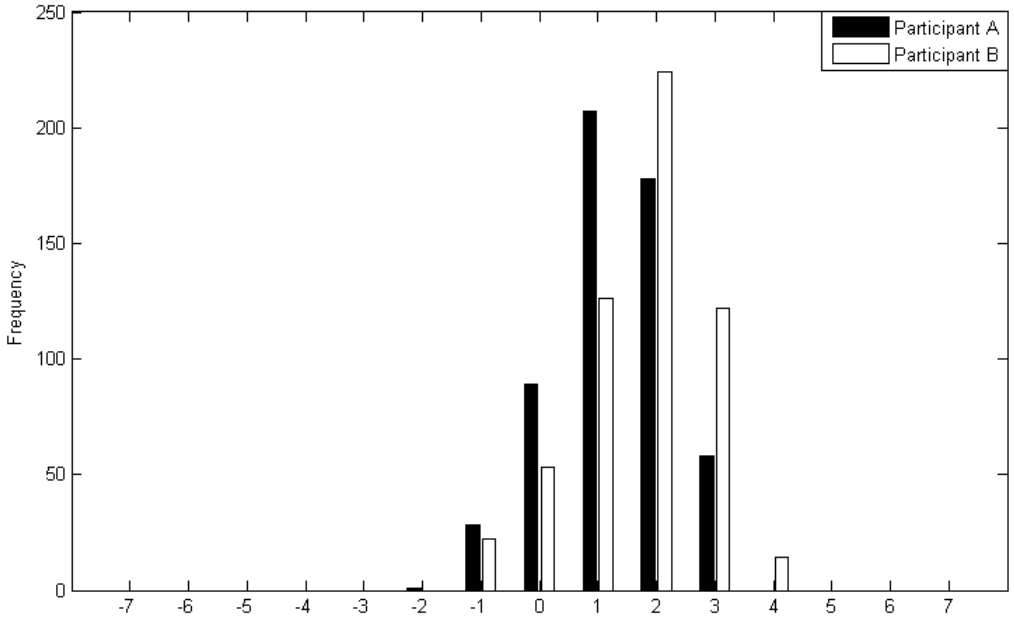


Figure E2: Frequency tables comparing the chi-square result and content similarity for Participant A (top), and Participant B (middle), and a histogram of the difference for 'A', 'B' and the chi-square result (bottom)

E.3. Structural Similarity Ratings

				‘Chi-Square Results’					
		1	**2**	**3**	**4**	**5**	**6**	**7**	**TOTAL**
‘A’	**1**	**0**	10	14	15	0	0	0	**39**
	2	0	**0**	20	15	0	0	0	**35**
	3	0	0	**35**	35	1	0	0	**71**
	4	0	0	3	**13**	2	0	0	**18**
	5	0	0	7	9	**4**	0	0	**20**
	6	0	0	36	122	91	**13**	5	**267**
	7	0	0	6	16	43	26	**20**	**111**
	TOTAL	**0**	**10**	**121**	**225**	**141**	**39**	**25**	**561**

				‘Chi-Square Results’					
		1	**2**	**3**	**4**	**5**	**6**	**7**	**TOTAL**
‘B’	**1**	**0**	8	13	1	0	0	0	**22**
	2	0	**2**	27	23	1	0	0	**53**
	3	0	0	**32**	41	1	0	0	**74**
	4	0	0	1	**8**	0	0	0	**9**
	5	0	0	2	11	**0**	0	0	**13**
	6	0	0	33	113	102	**13**	0	**261**
	7	0	0	13	28	37	26	**25**	**129**
	TOTAL	**0**	**10**	**121**	**225**	**141**	**39**	**25**	**561**

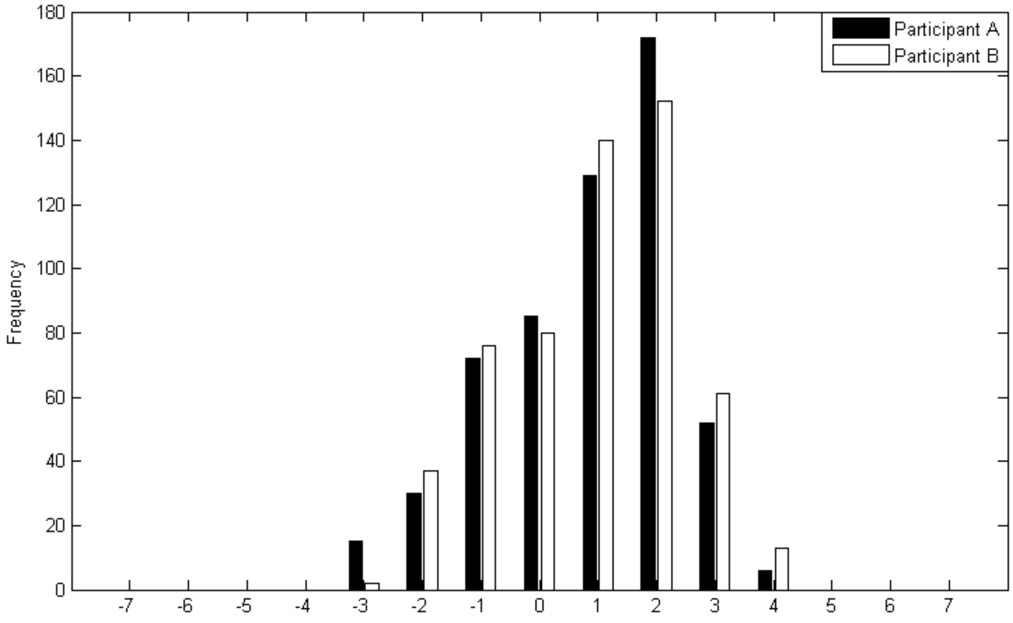


Figure E3: Frequency tables comparing the chi-square result and structural similarity for Participant A (top), and Participant B (middle), and a histogram of the difference for ‘A’, ‘B’ and the chi-square result (bottom)

E.4. Language Use Similarity Ratings

					'Chi-Square Results'				
		1	**2**	**3**	**4**	**5**	**6**	**7**	**TOTAL**
'A'	**1**	**0**	10	31	19	0	0	0	**60**
	2	0	**0**	15	30	0	0	0	**45**
	3	0	0	**19**	14	1	0	0	**34**
	4	0	0	13	**17**	3	0	0	**33**
	5	0	0	29	53	**19**	5	1	**107**
	6	0	0	13	79	75	**9**	8	**184**
	7	0	0	1	13	43	25	**16**	**98**
	TOTAL	**0**	**10**	**121**	**225**	**141**	**39**	**25**	**561**

					'Chi-Square Results'				
		1	**2**	**3**	**4**	**5**	**6**	**7**	**TOTAL**
'B'	**1**	**0**	8	18	0	0	0	0	**26**
	2	0	**2**	15	6	1	0	0	**24**
	3	0	0	**20**	21	0	0	0	**41**
	4	0	0	8	**29**	0	0	0	**37**
	5	0	0	19	42	**17**	2	0	**80**
	6	0	0	40	108	82	**10**	2	**242**
	7	0	0	1	19	41	27	**23**	**111**
	TOTAL	**0**	**10**	**121**	**225**	**141**	**39**	**25**	**561**

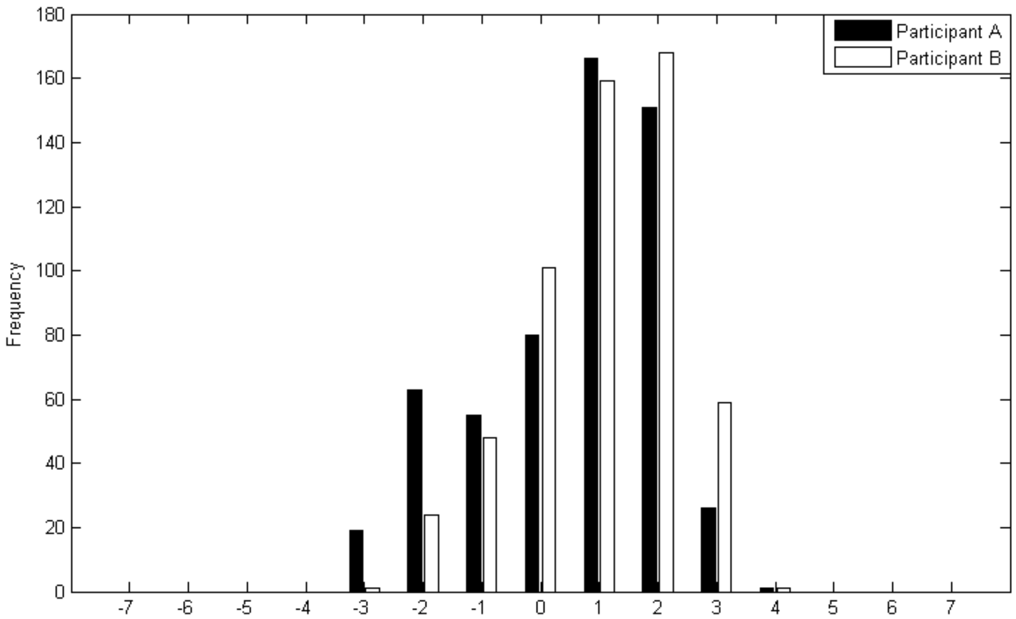


Figure E4: Frequency tables comparing the chi-square result and content similarity for Participant A (top), and Participant B (middle), and a histogram of the difference for 'A', 'B' and the chi-square result (bottom)

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION DOCUMENT CONTROL DATA	1. PRIVACY MARKING/CAVEAT (OF DOCUMENT)

2. TITLE Human Dimensions of Corpora Comparison: An Analysis of Kilgarriff's (2001) Approach	3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED REPORTS THAT ARE LIMITED RELEASE USE (L) NEXT TO DOCUMENT CLASSIFICATION) Document (U) Title (U) Abstract (U)
4. AUTHOR(S) Kathryn Parsons, Agata McCormac and Marcus Butavicius	5. CORPORATE AUTHOR DSTO Defence Science and Technology Organisation PO Box 1500 Edinburgh South Australia 5111 Australia

6a. DSTO NUMBER DSTO-TR-2290	6b. AR NUMBER AR-014-529	6c. TYPE OF REPORT Technical Report	7. DOCUMENT DATE April 2009

8. FILE NUMBER 2008/1147346	9. TASK NUMBER INT 007/020	10. TASK SPONSOR Intelligence	11. NO. OF PAGES 54	12. NO. OF REFERENCES 17

13. DOWNGRADING/DELIMITING INSTRUCTIONS To be reviewed three years after date of publication	14. RELEASE AUTHORITY Chief, Command, Control, Communications and Intelligence Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for public release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SA 5111

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CITATION IN OTHER DOCUMENTS Yes

18. DSTO RESEARCH LIBRARY THESAURUS http://web-vic.dsto.defence.gov.au/workareas/library/resources/dsto_thesaurus.shtml

Corpora Comparison, Human Performance, Algorithms, Empirical methods

19. ABSTRACT

There is a distinct lack of tools that provide a comprehensive measure of the similarity between corpora. Finding similar corpora is necessary for the design of certain user studies investigating text processing. It is also useful for ensuring comparability between studies on document analysis conducted across classified and unclassified domains. In this study, human judgements of corpora similarity were obtained as a gold standard. These were then compared to the values provided by Kilgarriff's (2001) chi-square (X^2) statistic. The findings indicated a high level of agreement between the participants, with 77% shared variance in overall similarity judgements. The results of the X^2 measure also correlated well with the human results, with a correlation of approximately 0.66. Although there are complexities associated with the X^2 technique that need to be examined in further research, this study provides extremely promising results, suggesting that a statistical technique could provide results that are comparable to human judgements.