Analysis Report

AntMover, AWA2 and AWA3

Table of Contents

1.0 Requirements

1.1 Test Data

The corpus (OA-STM) that being used in this statistic analysis of AntMover and AWA is from Elsevier repository. Elsevier is providing a selection of articles from 10 different STM (Scientific, Technical, and Medical) domains as a freely-redistributable corpus. The articles were selected from our Open Access content and have a Creative Commons CC-BY license so they are free to redistribute and use. The domains are agriculture, astronomy, biology, chemistry, computer science, earth science, engineering, materials science, math, and medicine. Currently we provide 11 articles in each of the 10 domains. Another 2 categories are added from PMC Corpus. These corpora are 'Philosopy' and 'Sociology' categories with each having 11 articles too. So, the total articles in the corpus is 132 with 40991 sentences.

10 categories (Elsevier): Agriculture, Astronomy, Biology, Chemistry, Computer Science, Earth Science, Engineering, Material Science, Mathematics and Medicine

2 categories (PMC): Philosophy and Sociology

1.2 Database

The database used in this research is MySQL database (***will be migrated to postgresSQL in Amazon Web Service***). All the sentences in the 110 articles are pre-processed and stored in the database. The table structures are designed to store all the documents, sentences, annotations and tools in order.

1.3 Statistical and programming language

The programming language used in this research is Python. Meanwhile, the statistic computing used R programming language.

2.0 AntMover

2.1 Background

AntMover 1.0 is a prototype version of a general learning environment that can be applied to the analysis of text structure in any field or discipline, and to any text type. It is a freeware text structure (moves) analysis program.

2.2 Drawback

After some testing on this software, I found these are the drawback of this system:

1. The function to separate sentences are not accurate and efficient

→ not able to identify the end of any sentence that end with a character or abbreviation. As a result, the sentence will not be annotated.

Example:

The result shows that robots can be constructed with the method shown in Figure A. Therefore, that method is efficient. (these two sentences are not broken into two sentences)

2.3 Advantage

Antmover annotates every sentence. However, the test for accuracy on how it annotates a sentence should be conducted by comparing to annotations made my human.

2.4 Annotation Scheme/ Moves

Annotation Id	Annotation Name
1	Claiming centrality
2	Making topic generalization
3	Indicating a gap
4	Announcing present research
5	Announcing principal findings
6	Evaluation of research

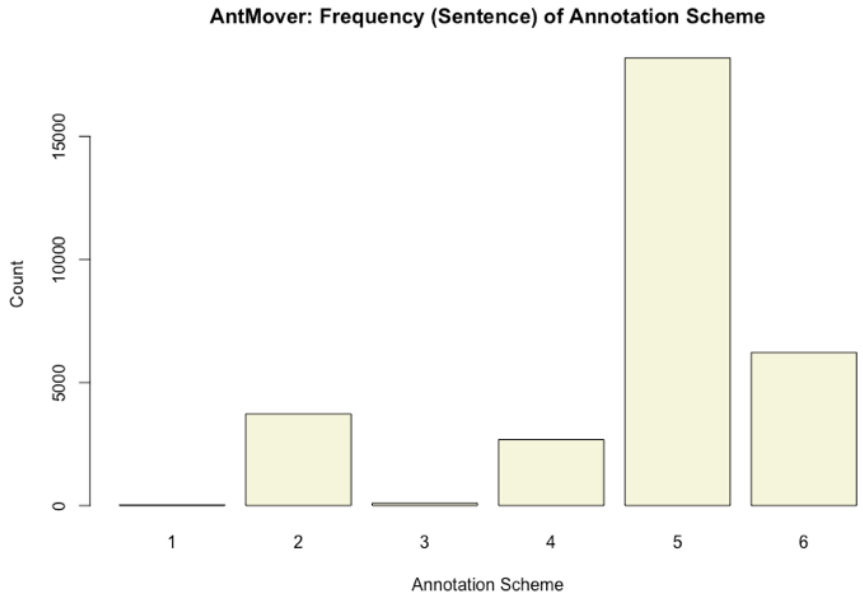
2.5 Test 1: Spread of moves over corpus

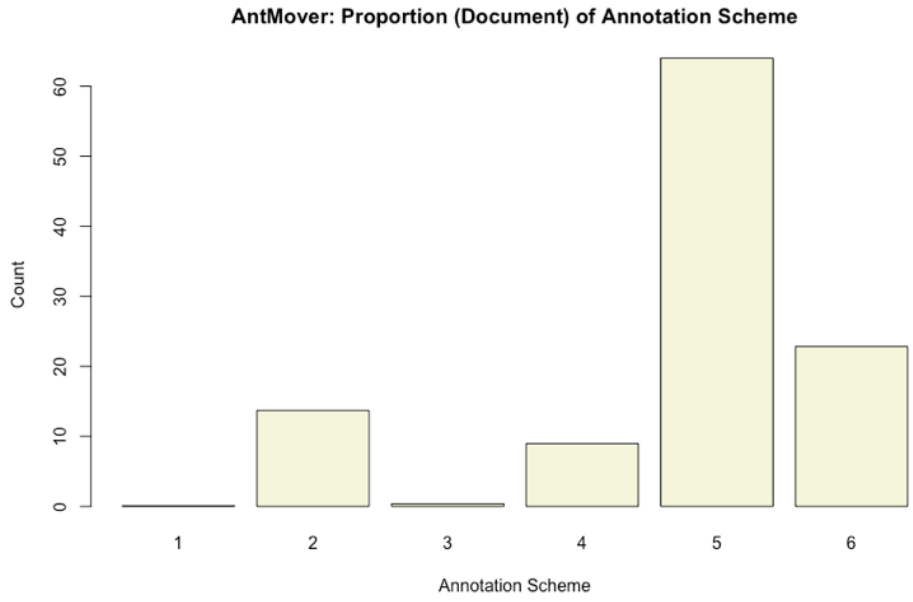
This test shows the frequency of each AntMover's moves in the corpus data.

Annotation Id	1	2	3	4	5	6	Total
Sentence	27	3725	96	2680	18186	6216	30930
Proportion (Document)	0.101	13.707	0.366	8.991	64.016	22.82	110

Proportion (Document) formula

= SUM(total_annotation_id_document/total_sentences_document)



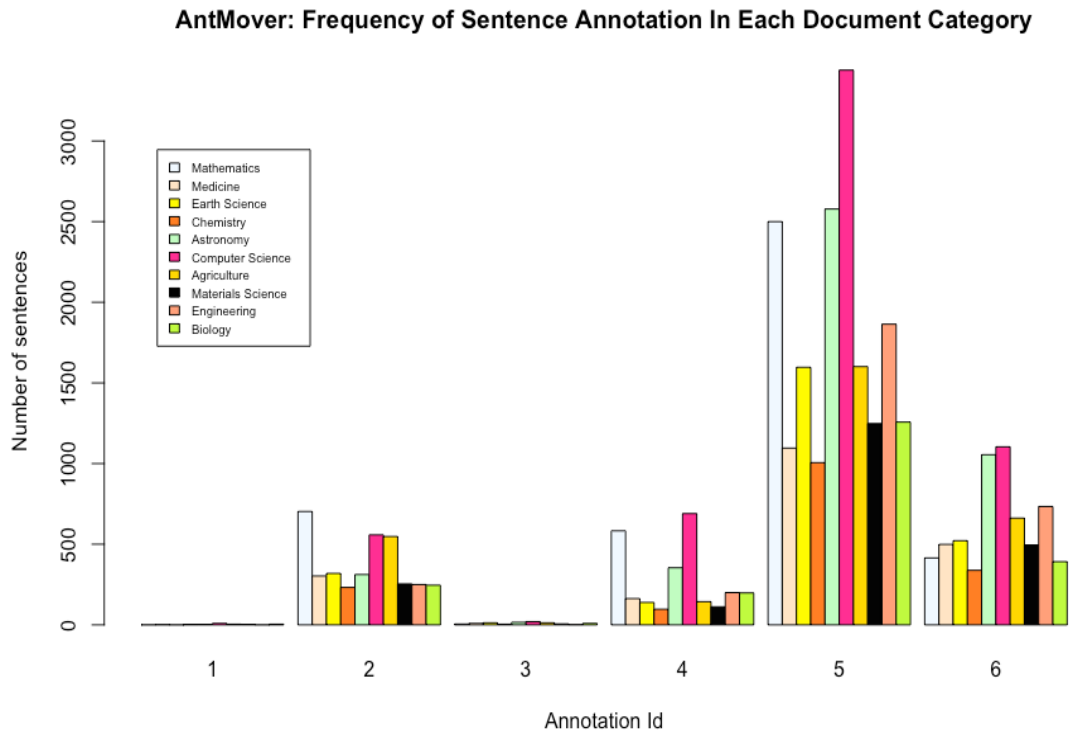


Result: It appears that most of the sentences are annotated with annotation 5. Meanwhile, annotations '1' and '3' are very less likely appear in the corpus. It also shows that not all annotations ('1' to '6') can be found in every document in the corpus (110 documents). As an example, only 27 annotation '1' in the whole corpus even though there are 110 documents. One more finding is that each sentence only annotated with at most one move as the total number of moves is 30930 which is equal to the total number of sentences in the corpus.

2.6 Test 2: Spread of moves over sub-corpora (document category)

This test shows the frequency of occurrences of each annotation scheme in each category of documents of the corpus data.

Category	1	2	3	4	5	6	Total
Mathematics	0	703	5	583	2500	415	4206
Medicine	1	303	9	163	1096	499	2071
Earth Science	0	319	13	139	1597	522	2590
Chemistry	3	232	4	97	1006	338	1680
Astronomy	3	312	16	354	2578	1055	4318
Computer Science	9	558	20	690	3439	1104	5820
Agriculture	4	548	13	144	1601	662	2972
Materials Science	3	254	6	112	1248	495	2118
Engineering	0	250	1	200	1864	734	3049
Biology	4	246	9	198	1257	392	2106
Total	27	3725	96	2680	18186	6216	30930



Result: Since AntMover annotates every sentence, there is nothing to derived from the above charts. It just appears that most of the annotations fall into annotation '5'. It also shows that some categories of documents do not contain certain annotations at all. Example, documents under category Mathematics do not contain any annotation '1'.

2.7 Test 3: Is every sentence annotate by AntMover?

By using the below SQL query, it shows that all sentences are annotated by AntMover.

Query:

SELECT COUNT(*) FROM SENTENCE WHERE sentence_id

NOT IN(SELECT sentence_id FROM SENTENCE_ANNOTATION WHERE tool_id=1);

Query result: 0

Result: All sentences are annotated by AntMover's moves.

2.8 Test 4: Is AntMover annotation process remains stable (reliable)?

I scheduled two time of annotations with the same corpus (110 articles) in different session. Then I tried to compare the results of the two sessions.

Query:

#to count the number of sentences that ***having same annotation id for two different dates***

SELECT COUNT(*) FROM SENTENCE_ANNOTATION sa1, SENTENCE_ANNOTATION sa2

WHERE sa1.sentence_id=sa2.sentence_id

AND sa1.annotation_id = sa2.annotation_id

AND sa1.tool_id = 1 AND sa2.tool_id=1

AND sa1.sentence_date='2017-10-08'

AND sa2.sentence_date='2017-10-10';

Query Result: 30930 (this is the total number of lines in 110 documents)

#to count number of sentences with **different annotation id for two different dates**

SELECT COUNT(*) FROM SENTENCE_ANNOTATION sa1, SENTENCE_ANNOTATION sa2

WHERE sa1.sentence_id=sa2.sentence_id

AND sa1.annotation_id != sa2.annotation_id

AND sa1.tool_id = 1 AND sa2.tool_id=1

AND sa1.sentence_date='2017-10-08'

AND sa2.sentence_date='2017-10-10';

Query Result: 0

Result: Each session produce the same result (same annotation ids on same sentences). Therefore, AntMover's annotation process is reliable.

3.0 Academic Writing Analytics (AWA2)

3.1 Background

Academic Writing Analytics or AWA (pronounced ay-wah) is a web application designed to provide insights to students on their writing. AWA comes in 2 flavours: An analytic version for analysing normal analytical style academic writing, and a reflection version for analysing reflective writing. Both flavours are actively being developed as part of CIC research projects. However, the focus of this research is on the analytical parser.

3.2 Drawback

I found that AWA does sentence correction. It may be helpful but it changes the original sentence without notify the users. In this project, it poses a challenge because the uploaded sentences from the corpus are edited and cannot be matched to the original sentences in the database. Therefore, manual matchings have been done to find out which original sentences are being annotated.

3.3 Annotation Scheme

Annotation Id	Category	Annotation Name
27	Main category	Important
28		Summary
29		Important&Summary
30	Sub category	Background
31		Contrast
32		Emphasis
33		Novelty
34		Position
35		Question
36		Surprise
37		Trend

The annotation scheme of AWA is divided into two categories which are ‘main category’ and ‘sub category’. The ‘main category’ contains ‘important’, ’summary’ and ‘important &

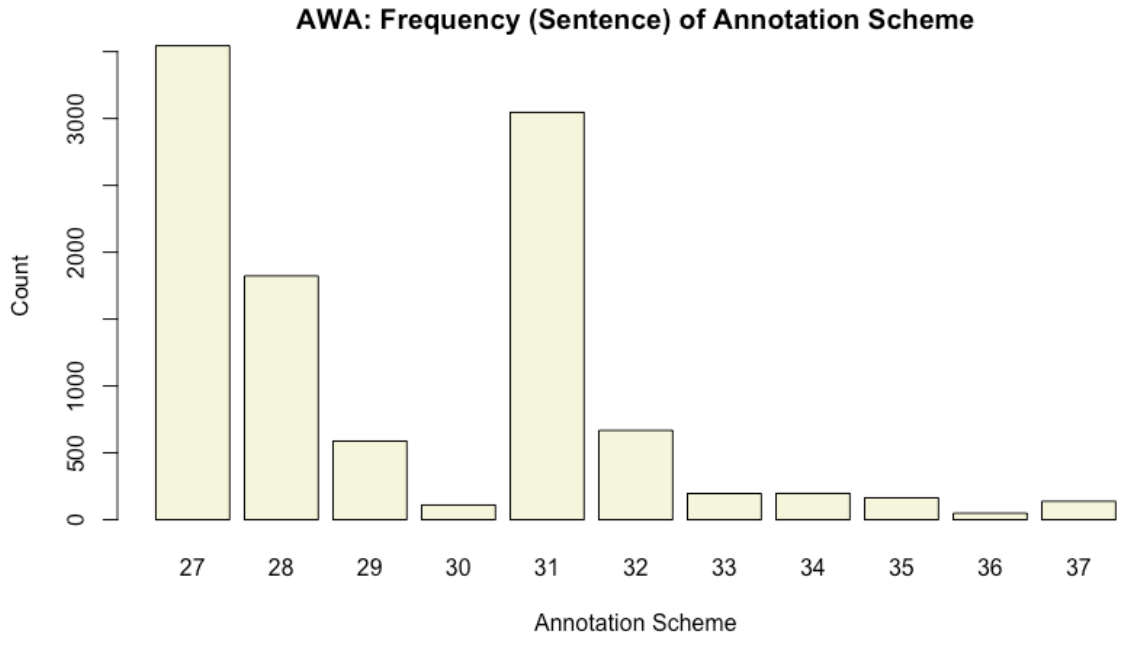
summary’. Meanwhile the ‘sub category’ contains ‘background’, ‘contrast’, ‘emphasis’, ‘novelty’, ‘position’, ‘question’, ‘surprise’ and ‘trend’.

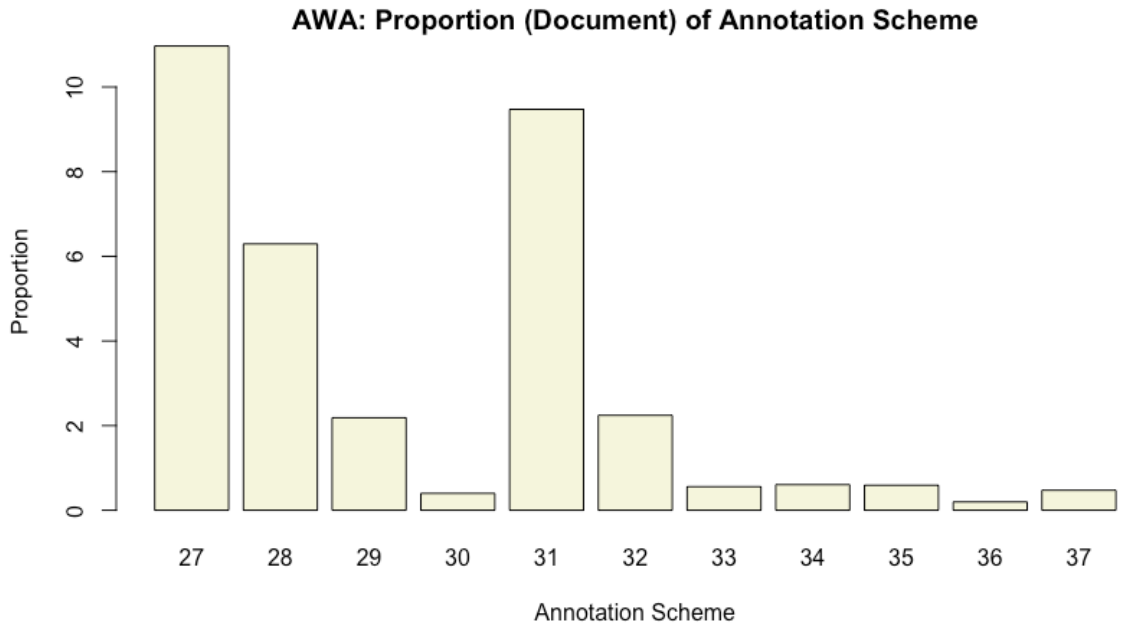
3.4 Test 1: Spread of annotations over corpus

This test shows the frequency of occurrences of each annotation scheme in the corpus data.

	Main Category			Sub Category								
Annotation Id	27	28	29	30	31	32	33	34	35	36	37	Total
Sentence	3544	1823	588	110	3046	668	196	197	164	49	138	10523
Proportion (Document)	10.968	6.293	2.186	0.399	9.473	2.243	0.557	0.605	0.591	0.197	0.473	33.985

Proportion = (total sentence of an annotation id)/ (total sentence in a document)





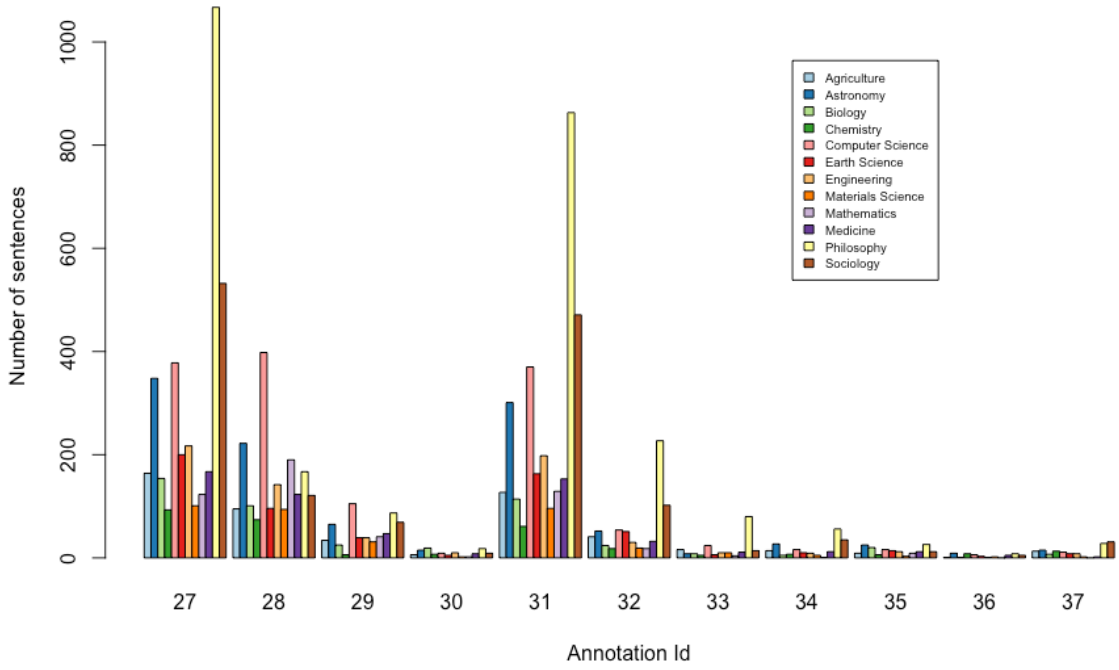
Result: In main category, most sentences fall into 'Important' category. Meanwhile, 'Contrast' category has the highest number of sentences in sub category. Total number of annotated sentences is 6495. It means that AWA is not annotating every sentence as total number of lines is 30930. This is contradicted to AntMover that annotates every sentence with moves.

3.5 Test 2: Spread of annotations over sub-corpora (document category)

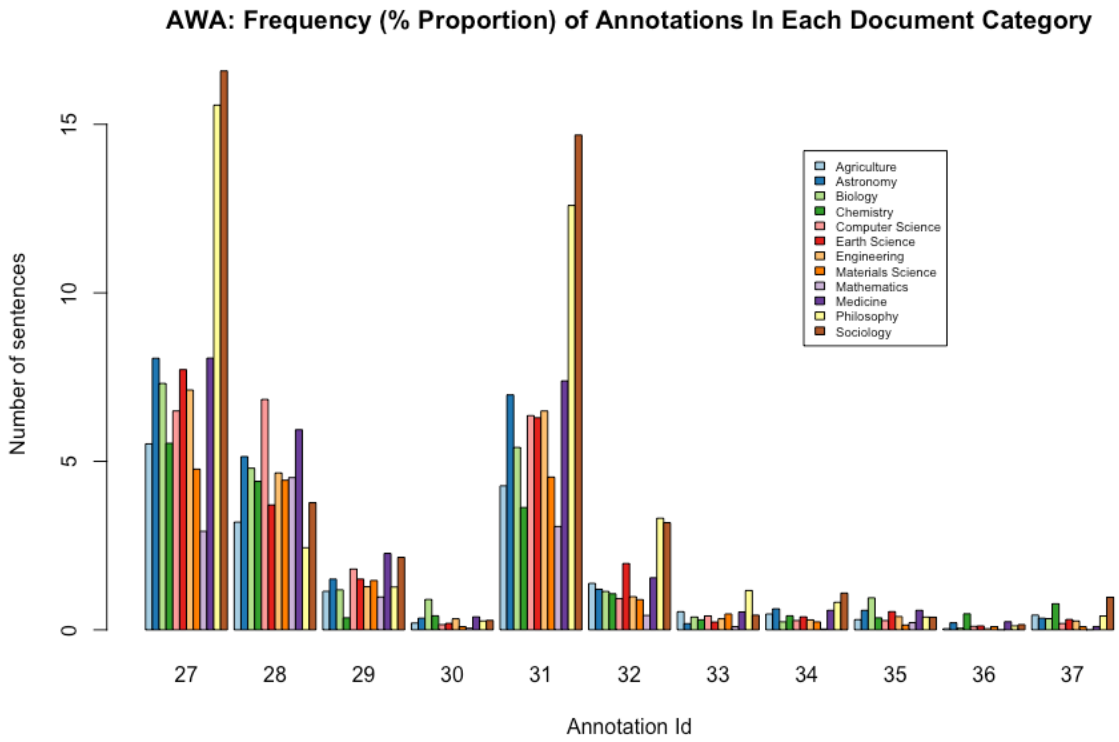
This test shows the frequency of occurrences of each annotation scheme in each category of documents of the corpus data. Notice that one sentence with two annotations will be counted as two as this test focus on number of annotations and not annotated sentences.

Category	27	28	29	30	31	32	33	34	35	36	37	Total
Agriculture	164	95	34	6	127	41	16	14	9	1	13	520
Astronomy	348	222	65	15	301	52	8	27	25	9	15	1087
Biology	154	101	25	19	114	24	8	5	20	1	7	478
Chemistry	93	74	6	7	61	18	5	7	6	8	13	298
Computer Science	378	398	105	9	370	54	24	16	16	6	11	1387
Earth Science	200	96	39	5	163	51	6	10	14	3	8	595
Engineering	217	142	39	10	198	30	10	9	12	1	8	676
Materials Science	101	94	31	2	96	19	10	5	3	2	2	365
Mathematics	123	190	41	2	129	18	4	1	9	0	0	517
Medicine	167	123	47	8	153	32	11	12	12	5	2	572
Philosophy	1067	167	87	18	863	227	80	56	26	8	28	2627
Sociology	532	121	69	9	471	102	14	35	12	5	31	1401
Total	3544	1823	588	110	3046	668	196	197	164	49	138	10523

AWA: Frequency of Annotations In Each Document Category



Category	27	28	29	30	31	32	33	34	35	36	37	Total
Agriculture	5.518	3.197	1.144	0.202	4.273	1.38	0.538	0.471	0.303	0.034	0.437	17.497
Astronomy	8.059	5.141	1.505	0.347	6.971	1.204	0.185	0.625	0.579	0.208	0.347	25.171
Biology	7.312	4.796	1.187	0.902	5.413	1.14	0.38	0.237	0.95	0.047	0.332	22.696
Chemistry	5.536	4.405	0.357	0.417	3.631	1.071	0.298	0.417	0.357	0.476	0.774	17.739
Computer Science	6.495	6.838	1.804	0.155	6.357	0.928	0.412	0.275	0.275	0.103	0.189	23.831
Earth Science	7.722	3.707	1.506	0.193	6.293	1.969	0.232	0.386	0.541	0.116	0.309	22.974
Engineering	7.117	4.657	1.279	0.328	6.494	0.984	0.328	0.295	0.394	0.033	0.262	22.171
Materials Science	4.769	4.438	1.464	0.094	4.533	0.897	0.472	0.236	0.142	0.094	0.094	17.233
Mathematics	2.924	4.517	0.975	0.048	3.067	0.428	0.095	0.024	0.214	0	0	12.292
Medicine	8.064	5.939	2.269	0.386	7.388	1.545	0.531	0.579	0.579	0.241	0.097	27.618
Philosophy	15.57	2.437	1.27	0.263	12.593	3.312	1.167	0.817	0.379	0.117	0.409	38.334
Sociology	16.584	3.772	2.151	0.281	14.682	3.18	0.436	1.091	0.374	0.156	0.966	43.673
Total	95.67	53.844	16.911	3.616	81.695	18.038	5.074	5.453	5.087	1.625	4.216	291.229



Result: In Mathematics documents, the number of sentences with 'Surprise' (id=36) or 'Trend' (id=37) is 0. Besides that, the 'Background' (id=30) has only 2 sentences and 'Position' (id=34) has only 1 sentences out of 11 Mathematics documents. While other categories of documents are having higher number of sentences in the mentioned annotations. The following hypothesis can be derived:

$H_{0:}$ AWA sub categories annotations 'Surprise', 'Trend', 'Background' and 'Position' work well with Mathematics documents.

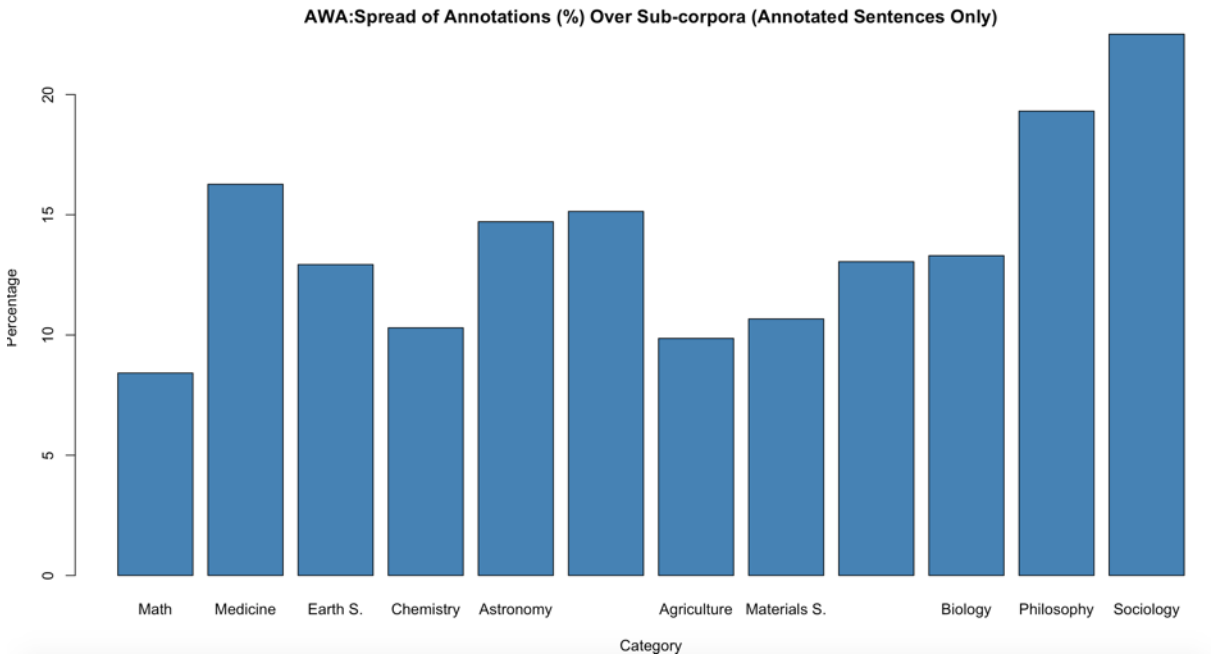
H_1: AWA sub categories annotations 'Surprise', 'Trend', 'Background' and 'Position' not work well with Mathematics documents.

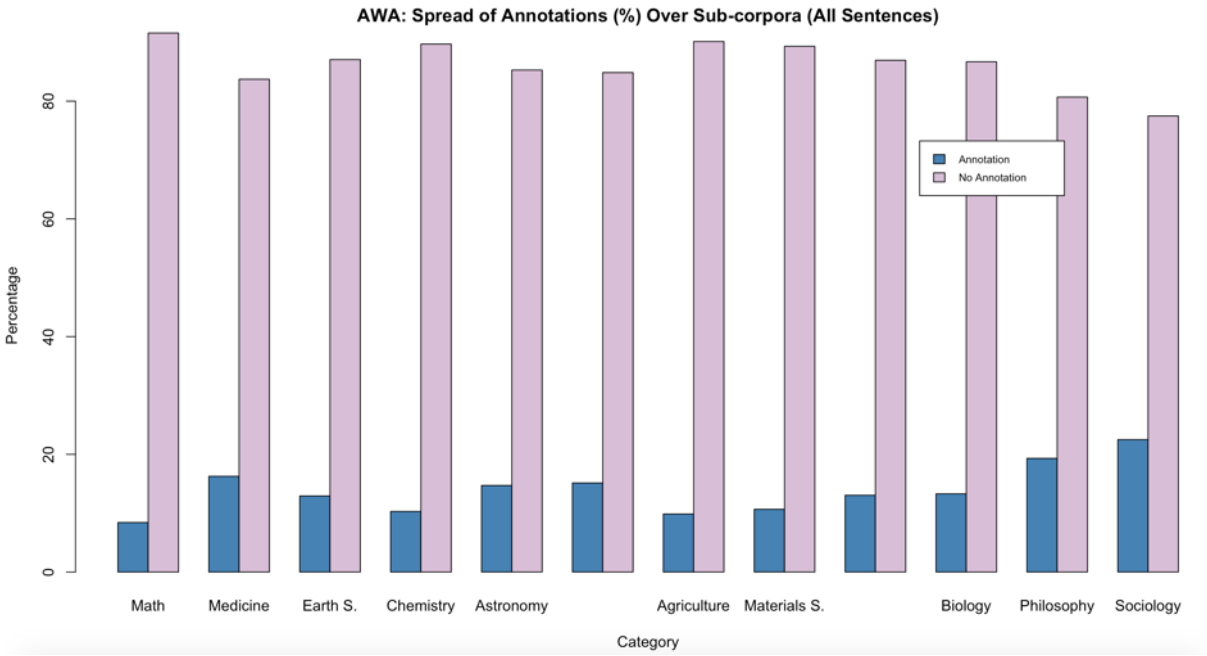
3.6 Test 3: Spread of annotated sentences over sub-corpora (document category)

This test shows the percentage of annotated sentences (opposite to Test 2 above) in each category of documents of the corpus data. Notice that one sentence with two annotations will

be counted as one as this test focus on number of annotated sentences and not number of annotations.

Category	Annotated	Not Annotated	Total	Annotated (%)	Annotated
Mathematics	354	3852	4206	8.42	91.58
Medicine	337	1734	2071	16.27	83.73
Earth Science	335	2255	2590	12.93	87.07
Chemistry	173	1507	1680	10.30	89.70
Astronomy	635	3683	4318	14.71	85.29
Computer Science	881	4939	5820	15.14	84.86
Agriculture	293	2679	2972	9.86	90.14
Materials Science	226	1892	2118	10.67	89.33
Engineering	398	2651	3049	13.05	86.95
Biology	280	1826	2106	13.30	86.70
Philosophy	1323	5530	6853	19.31	80.69
Sociology	722	2486	3208	22.51	77.49
Total	5957	35034	40991		





Result: It appears that Mathematics sub-corpora has the least number of annotated sentences. It means AWA less likely to annotate sentences in Mathematics documents. The same hypothesis as test 2 above can be derived from here too.

4.0 AntMover and AWA2 reliability

This section will compare the reliability of AntMover and AWA. In this test, AntMover and AWA will act as 'raters' and the number of sentences will be the subjects. The agreement between AntMover and AWA to annotate same sentences will be evaluated.

4.1 Test 1: Percentage agreement in annotating sentence

Subjects = 40991 (sentences)

Raters = 2 (AntMover and AWA)

%-agree = 14.5

Result: There is 14.5% similarity in agreement of annotating the same sentences.

4.2 Test 2: Krippendorf's alpha and Cohen's Kappa for reliability

Based on Krippendorff (2004) and Fulcher (2010), the minimum acceptable inter-rater agreement value for Krippendorf's alpha is 0.80 while Cohen's kappa is 0.61.

AntMover and AWA	Krippendorf's alpha	Cohen's kappa
Value	-0.746	0

Result: The negative value of Krippendorf's alpha (-0.775) shows that there is no agreement between AntMover and AWA. On the other hand, both have strong disagreement. In addition, the Cohen's kappa value is 0 and this proof that the agreement in annotating same sentence is happened purely by chance (no agreement). In short, there is no inter-rater reliability between these tools.

4.3: Test 3: Krippendorf's alpha and Cohen's Kappa for reliability over sub-corpora

This test the reliability of AntMover and AWA over each document category

Category	Subject	Agreement (%)	Krippendorf's alpha	Cohen's kappa
Mathematics	4206	8.42	-0.845	0
Medicine	2071	16.3	-0.72	0
Earth Science	2590	12.9	-0.771	0
Chemistry	1680	10.3	-0.813	0
Astronomy	4318	14.7	-0.743	0
Computer Science	5820	15.1	-0.737	0
Agriculture	2972	9.86	-0.82	0
Materials Science	2118	10.7	-0.807	0
Engineering	3049	13.1	-0.769	0
Biology	2106	13.3	-0.765	0
Philosophy	6853	19.3	-0.676	0
Sociology	3208	22.5	-0.632	0
Total	30930			

Result: In general, all categories have negative values in Krippendorf's alpha and 0 values in Cohen's kappa. This means both tools are in disagreement in annotating sentences in each category. The category with highest disagreement is Mathematics which accounted with -0.845 in Krippendorf's alpha value. Based on the percentage agreement, Medicine category has the highest agreement (16.3%) follow by Computer Science (15.1%).

4.4: Test 4: Reliability test between certain AntMover's moves and AWA's annotations

This test the agreement between certain AntMover's moves and AWA's annotations. The test will cover percentage agreement test, Krippendorf's alpha and Cohen's kappa.

	AWA 'main category' annotation		
AntMover move	Agreement (%)	Krippendorf's alpha	Cohen's kappa
2	5.56	-0.895	-0.885
4	12.1	-0.784	-0.603
6	12	-0.786	-0.73

	AWA 'sub category' annotation		
AntMover move	Agreement (%)	Krippendorf's alpha	Cohen's kappa
2	5.91	-0.888	-0.854
4	5.11	-0.903	-0.878
6	10.4	-0.812	-0.583

	AWA 'sub category' annotation		
AntMover move	Agreement (%)	Krippendorf's alpha	Cohen's kappa
1	annotation id=36		
	0	-0.989	-0.984
3	annotation id=30		
	0	-0.996	-0.959
	annotation id=33		
	0	-0.997	-0.96
	annotation id=34		
	0	-0.997	-0.959
	annotation id=35		
	0.647	-0.984	-0.981
	annotation id=37		
	0.352	-0.989	-0.991

Result: The results had shown strong disagreement between the AntMover's moves and AWA's annotations.

5.0 AntMover and AWA2 association

5.1 Test 1: Association between AntMover's moves and AWA's annotations

The test will compare the association between AntMover's moves and AWA's annotations.

Null hypothesis, $\mathbf{H_0}$: AntMover's moves and AWA's annotations are independent

Alternate hypothesis, $\mathbf{H_1}$: AntMover's moves and AWA's annotations are dependent

	AWA's annotations											
	Main category			Sub category								
AntMover's moves	Important	Summary	Important&Summary	Background	Contrast	Emphasis	Novelty	Position	Question	Surprise	Trend	Total
Claiming_centrality	7	2	0	0	2	6	0	1	0	0	0	18
Announcing_principal_findings	1769	800	256	31	1520	321	90	81	76	25	69	5038
Evaluation_of_research	1035	330	152	32	848	243	47	95	49	14	35	2880
Making_topic_generalization	502	60	26	45	372	71	19	17	29	6	28	1175
Indicating_a_gap	20	3	4	0	21	2	0	0	2	0	1	53
Announcing_present_research	211	628	150	2	283	25	40	3	8	4	5	1359
Total	3544	1823	588	110	3046	668	196	197	164	49	138	10523

Chisquare test result:

X-squared = 1432.4, df = 50, p-value < 2.2e-16

Result: The p-value is smaller than 0.05, so reject null hypothesis. There is a significant association between AntMover's moves and AWA's annotations. The values of contributing cells to the total Chi-square score is shown at section 5.2.

5.2 Test 2: Positive and negative association between AntMover and AWA

Cells with the highest absolute standardized residuals contribute the most to the total Chi-square score. The negative sign means negative association.

	Background	Contrast	Emphasis	Important	Important&Summary	Novelty	Position	Question	Summary	Surprise	Trend
Claiming_centrality	-0.434	-1.406	4.544	0.381	-1.003	-0.579	1.142	-0.530	-0.633	-0.290	-0.486
Announcing_principal_findings	-2.985	1.616	0.066	1.755	-1.520	-0.396	-1.371	-0.284	-2.464	0.318	0.361
Evaluation_of_research	0.345	0.497	4.451	2.089	-0.704	-0.907	5.595	0.614	-7.563	0.161	-0.451
Making_topic_generalizations	9.335	1.729	-0.416	5.342	-4.894	-0.617	-1.065	2.498	-10.062	0.226	3.208
Indicating_a_gap	-0.744	1.445	-0.744	0.509	0.603	-0.994	-0.996	1.292	-2.040	-0.497	0.366
Announcing_present_research	-3.238	-5.565	-6.597	-11.531	8.499	2.919	-4.449	-2.864	25.585	-0.925	-3.037

Result: The blue colour represents positive association while orange colour represents negative association (no association in this context). The volume of the 'round' shape represents the strength of the association. From the picture, the following can be derived

a. Strong positive association:

- 'making topic generalization' → 'background'
- 'evaluation of research' → 'emphasis'
- 'evaluation of research' → 'position'
- 'announcing present research' → 'summary'
- 'announcing present research' → 'important&summary'
- 'claiming centrality' → 'emphasis'

b. Strong negative association

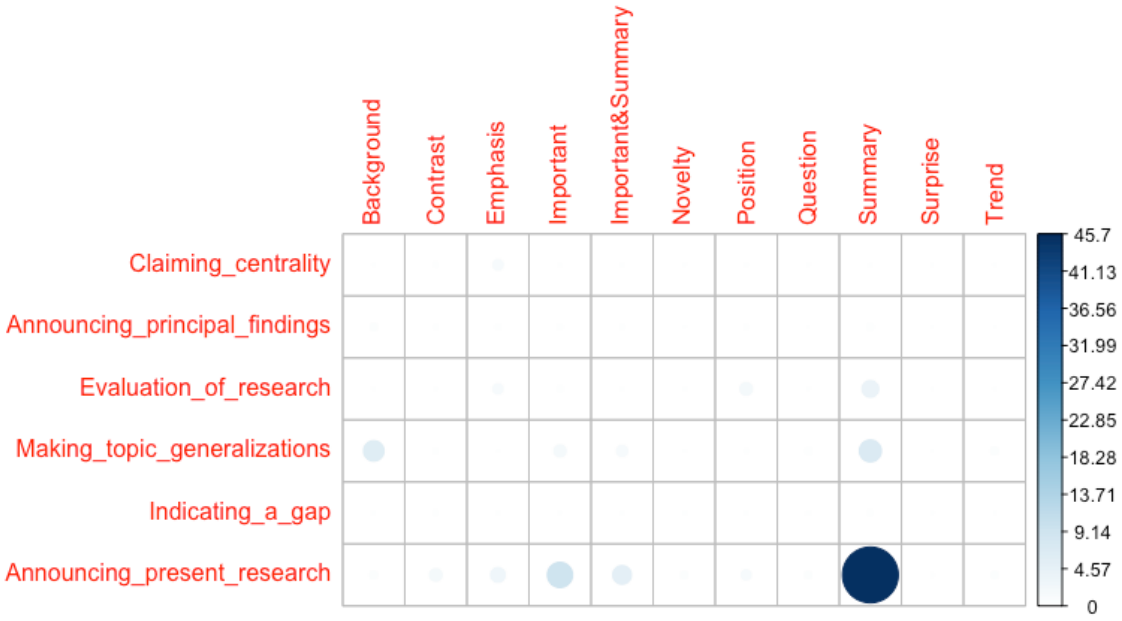
- 'announcing present research' → 'important'
- 'announcing present research' → 'emphasis'
- 'making topic generalization' → 'summary'

- ‘evaluation of research → ‘summary’

5.3 Test 3: Degree of association (%) between AntMover and AWA

The contribution in percentage just shows the significant contribution.

	Background	Contrast	Emphasis	Important	Important&Summary	Novelty	Position	Question	Summary	Surprise	Trend
Claiming_centrality	0.013	0.138	1.442	0.010	0.070	0.023	0.091	0.020	0.028	0.006	0.016
Announcing_principal_findings	0.622	0.182	0.000	0.215	0.161	0.011	0.131	0.006	0.424	0.007	0.009
Evaluation_of_research	0.008	0.017	1.383	0.305	0.035	0.057	2.186	0.026	3.993	0.002	0.014
Making_topic_generalizations	6.084	0.209	0.012	1.993	1.672	0.027	0.079	0.435	7.068	0.004	0.718
Indicating_a_gap	0.039	0.146	0.039	0.018	0.025	0.069	0.069	0.116	0.291	0.017	0.009
Announcing_present_research	0.732	2.162	3.038	9.283	5.043	0.595	1.382	0.573	45.698	0.060	0.644



Result: This picture can be interpreted the same way as above. ‘Making topic generalization’ is significantly associated with ‘background’. In general sense, this is true as most of the content in a ‘background’ category of an article generalizes or the topic of the article. On the other hand, the ‘question’ and ‘surprise’ categories have zero association with ‘announcing principal findings’. ‘Announcing principal findings’ should have strong association with ‘emphasis’ or ‘background’ in AWA annotation. As an example, a sentence with ‘principal findings’ should be recognized by AWA as ‘emphasis’ or ‘background’ of certain research/facts. Furthermore, another category to be concerned is the ‘indicating a gap’. This

category has very weak association in all AWA annotations. It supposes to have high association in ‘question’, ‘emphasis’, ‘position’ or ‘trend’.

5.4 Test 4: Association between AntMover and AWA over sub-corpora

This test is similar to test 5.1 but it restricted to category of document (sub-corpora). The result of the chisquare for each category of documents is as follow.

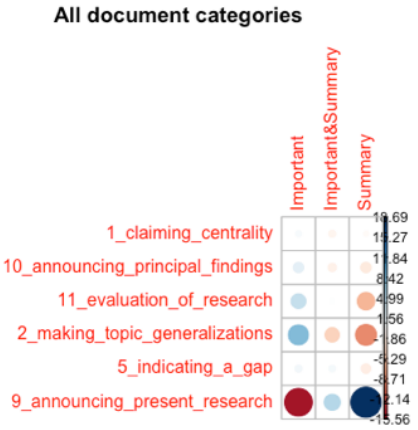
Category	Chisquare p-value	Null Hypothesis	Independent (Yes/No)
Mathematics	3.404e-15	reject	No
Medicine	1.569e-09	reject	No
Earth Science	1.777e-06	reject	No
Chemistry	0.0006613	reject	No
Astronomy	2.2e-16	reject	No
Computer Science	2.2e-16	reject	No
Agriculture	0.007117	reject	No
Materials Science	0.0004878	reject	No
Engineering	1.512e-13	reject	No
Biology	2.634e-05	reject	No
Philosophy	2.2e-16	reject	No
Sociology	2.17e-14	reject	No

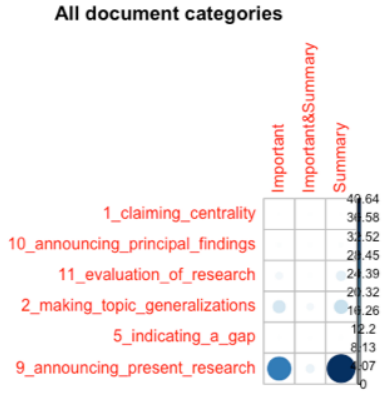
Result: All categories show that AWA and AntMover are associated.

5.5 Association between AntMover and AWA (main category and sub category)

Association Test (Main Category)

X-squared = 859.64, df = 10, **p-value < 2.2e-16**

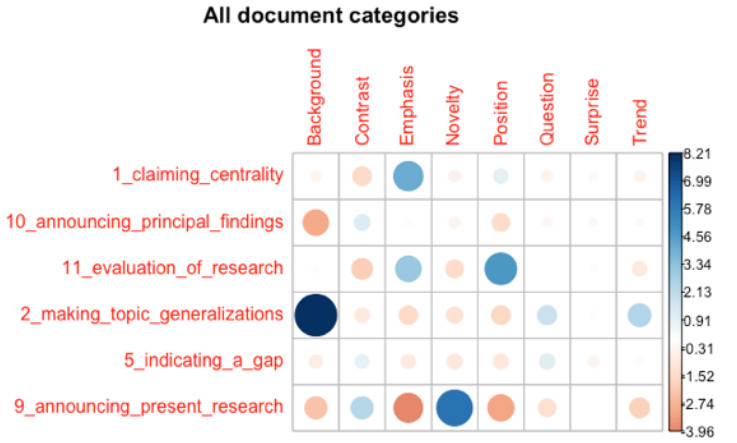


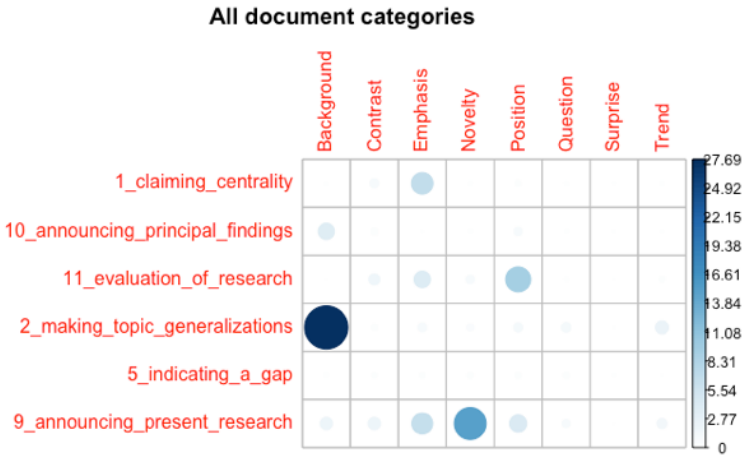


Result: The p-value is lower than 0.05. So, reject the null hypothesis. There is association between AntMover and AWA.

Association Test (Sub Category)

X-squared = 243.4, df = 35, **p-value < 2.2e-16**





Result: The p-value is lower than 0.05. So, reject the null hypothesis. There is association between AntMover and AWA.

6.0 AWA2: Sub corpora annotations distribution comparison

6.1 Non-normal distribution assumption

H_0: There is no significant different between the sub-corpora annotations distribution

H_1: There is significant different between the sub-corpora annotations distribution

Wilcoxon signed rank test (all annotations)

	Mathematics	Medicine	Earth	Chemistry	Astronomy	Computer	Agriculture	Materials	Engineering	Biology	Philosophy	Sociology
Mathematics		0.05371	0.1475	0.7002	0.2402	0.2061	0.2619	0.1422	0.2298	0.3981	0.3063	0.3503
Medicine			0.9658	0.4767	0.9188	0.1533	1	0.3063	0.7646	0.9658	0.8385	1
Earth Science				1	0.8938	0.4131	0.9658	0.6835	0.8311	0.8588	1	0.7598
Chemistry					0.5771	0.9593	0.8311	0.7002	0.4765	1	1	0.9658
Astronomy						0.5195	0.8588	0.824	0.8311	0.7555	1	1
Computer Science							0.4648	0.6888	0.4648	0.7596	0.3737	0.5536
Agriculture								0.5771	0.7002	0.9188	0.5195	0.3203
Materials Science									1	0.9291	0.3056	0.6565
Engineering										0.4491	0.5771	0.5047
Biology											0.8984	0.9658
Philosophy												0.9291

Friedman chi-squared = 12.736, df = 11, **p-value = 0.311**

Result: All p-values are above 0.05. So, the null hypothesis (H_0) is not rejected. There is no significant different on how AWA annotates different sub-corpora.

Wilcoxon signed rank test (main category annotations only)

	Main category annotations											
	Mathematics	Medicine	Earth Science	Chemistry	Astronomy	Computer Science	Agriculture	Materials Science	Engineering	Biology	Philosophy	Sociology
Mathematics		1	0.75	0.75	0.75	0.75	0.75	1	0.75	0.75	0.75	0.75
Medicine			0.75	1	1	1	0.75	0.75	1	1	0.75	0.75
Earth Science				1	1	0.75	1	0.75	1	1	0.75	0.75
Chemistry					1	0.5862	1	0.75	1	1	1	1
Astronomy						0.75	0.5862	0.75	0.75	1	0.75	0.75
Computer Science							0.75	1	0.75	0.75	0.75	0.75
Agriculture								0.75	0.75	0.75	0.75	0.75
Materials Science									0.75	0.75	0.75	0.75
Engineering										1	0.75	0.75
Biology											0.75	0.75
Philosophy												1

Friedman chi-squared = 3.2051, df = 11, **p-value = 0.9877**

Result: All p-values are above 0.05. So, the null hypothesis (H_0) is not rejected. There is no significant different on how AWA annotates different sub-corpora.

Wilcoxon signed rank test (sub category annotations only)

	Sub category annotations											
	Mathematics	Medicine	Earth Science	Chemistry	Astronomy	Computer Science	Agriculture	Materials Science	Engineering	Biology	Philosophy	Sociology
Mathematics		0.007813	0.007813	0.1953	0.01563	0.05469	0.05149	0.0584	0.01415	0.0584	0.04206	0.0584
Medicine			0.7422	0.4002	0.7998	0.0754	0.7422	0.141	0.7422	1	0.3525	0.6406
Earth Science				0.7422	0.5276	0.07813	0.8438	0.2049	0.9453	0.7792	0.6721	1
Chemistry					0.3828	0.2719	0.8438	0.3828	0.1829	1	0.9453	0.5469
Astronomy						0.07813	0.9453	0.293	0.7422	0.8438	0.7422	0.6726
Computer Science							0.1953	0.8885	0.07813	0.3517	0.06836	0.1422
Agriculture								0.1484	0.4609	0.6726	0.6406	0.3125
Materials Science									0.6236	0.8885	0.02917	0.207
Engineering										0.4002	0.1953	0.2334
Biology											0.5469	0.8438
Philosophy												0.8885

Friedman chi-squared = 22.992, df = 11, **p-value = 0.01772**

Result: The Wilcoxon signed rank test shows that 'Mathematics' documents are significantly annotated differently by AWA compare to 'Medicine', 'Earth Science', 'Astronomy', 'Engineering' and 'Philosophy' documents. Even the Friedman test shows that in general AWA annotates differently across different category of documents (the main contribution is 'Mathematics' category). So, the null hypothesis is rejected.

7.0 AntMover and AWA2 association (BAWE corpus)

7.1 Test 1: Association between AntMover's moves and AWA's annotations

The test will compare the association between AntMover's moves and AWA's annotations in BAWE corpus instead of Elsevier corpus.

Null hypothesis, $\mathbf{H_0}$: AntMover's moves and AWA's annotations are independent

Alternate hypothesis, $\mathbf{H_1}$: AntMover's moves and AWA's annotations are dependent

	AWA's annotations											
	Main category			Sub category								
AntMover's moves	Important	Summary	Important&Summary	Background	Contrast	Emphasis	Novelty	Position	Question	Surprise	Trend	Total
Claiming_centrality	35	3	0	0	10	23	3	1	0	1	3	79
Announcing_principal_findings	16372	2528	988	159	12352	3822	734	1354	610	127	658	39704
Evaluation_of_research	1480	159	99	48	1065	429	84	174	54	19	50	3661
Making_topic_generalization	5344	409	204	367	3813	1128	168	413	210	36	214	12306
Indicating_a_gap	46	4	0	0	42	3	0	0	3	0	0	98
Announcing_present_research	758	421	149	4	705	114	70	24	14	1	24	2284
Total	24035	3524	1440	578	17987	5519	1059	1966	891	184	949	58132

Chisquare test result:

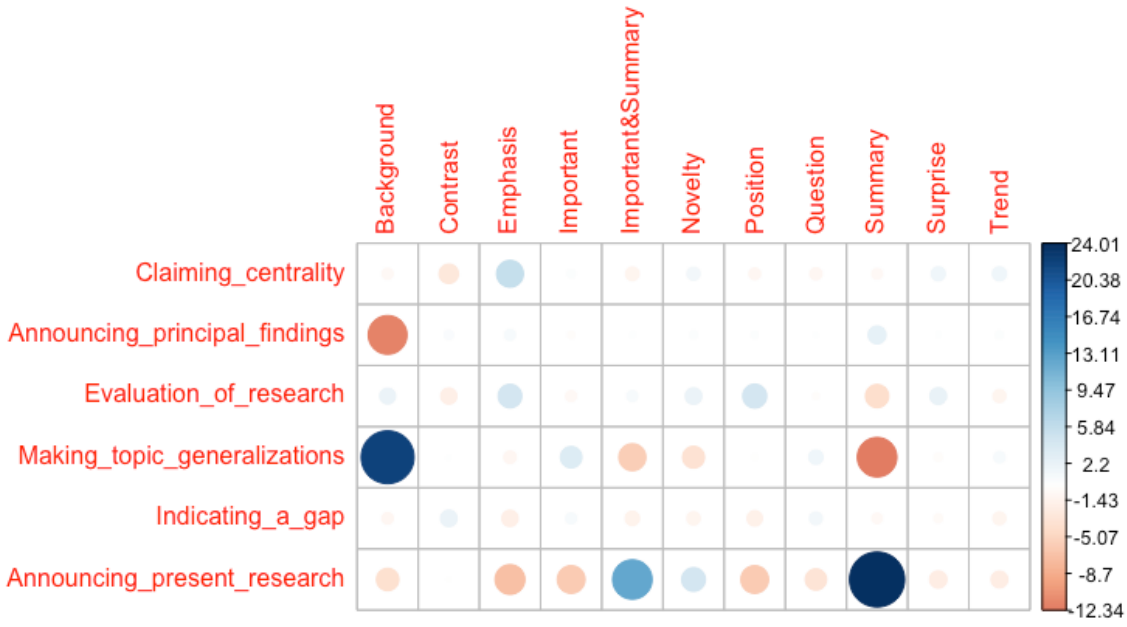
X-squared = 1914.6, df = 50, **p-value < 2.2e-16**

Result: The p-value is smaller than 0.05, so reject null hypothesis. There is a significant association between AntMover's moves and AWA's annotations. The values of contributing cells to the total Chi-square score is shown at section 5.2.

7.2 Test 2: Positive and negative association between AntMover and AWA

Cells with the highest absolute standardized residuals contribute the most to the total Chi-square score. The negative sign means negative association.

	Background	Contrast	Emphasis	Important	Important&Summary	Novelty	Position	Question	Summary	Surprise	Trend
Claiming_centrality	-0.886	-2.921	5.660	0.409	-1.399	1.301	-1.023	-1.100	-0.818	1.500	1.506
Announcing_principal_findings	-11.866	0.604	0.856	-0.342	0.143	0.398	0.306	0.059	2.469	0.119	0.386
Evaluation_of_research	1.923	-2.014	4.368	-0.865	0.873	2.119	4.510	-0.282	-4.224	2.177	-1.263
Making_topic_generalizations	22.117	0.086	-1.180	3.589	-5.775	-3.752	-0.156	1.557	-12.338	-0.473	0.925
Indicating_a_gap	-0.987	2.121	-2.067	0.861	-1.558	-1.336	-1.821	1.222	-0.796	-0.557	-1.265
Announcing_present_research	-3.926	-0.064	-6.984	-6.064	12.287	4.402	-6.058	-3.551	24.012	-2.317	-2.176



Result: The blue colour represents positive association while orange colour represents negative association (no association in this context). The volume of the 'round' shape represents the strength of the association. From the picture, the following can be derived

a. Strong positive association:

- 'making topic generalization' → 'background'
- 'evaluation of research' → 'emphasis'

- ‘evaluation of research’ → ‘position’
- ‘announcing present research’ → ‘summary’
- ‘announcing present research’ → ‘important&summary’
- ‘claiming centrality’ → ‘emphasis’

b. Strong negative association

- ‘announcing present research’ → ‘important’
- ‘announcing present research’ → ‘emphasis’
- ‘making topic generalization’ → ’summary’
- ‘evaluation of research → ‘summary’

7.3 Test 3: Degree of association (%) between AntMover and AWA

The contribution in percentage just shows the significant contribution.

	Background	Contrast	Emphasis	Important	Important&Summary	Novelty	Position	Question	Summary	Surprise	Trend
Claiming_centrality	0.041	0.446	1.673	0.009	0.102	0.088	0.055	0.063	0.035	0.117	0.118
Announcing_principal_findings	7.355	0.019	0.038	0.006	0.001	0.008	0.005	0.000	0.318	0.001	0.008
Evaluation_of_research	0.193	0.212	0.996	0.039	0.040	0.235	1.063	0.004	0.932	0.248	0.083
Making_topic_generalizations	25.549	0.000	0.073	0.673	1.742	0.735	0.001	0.127	7.951	0.012	0.045
Indicating_a_gap	0.051	0.235	0.223	0.039	0.127	0.093	0.173	0.078	0.033	0.016	0.084
Announcing_present_research	0.805	0.000	2.548	1.920	7.886	1.012	1.917	0.658	30.115	0.280	0.247

Result: This picture can be interpreted the same way as above. 'Making topic generalization' is significantly associated with 'background'. In general sense, this is true as most of the content in a 'background' category of an article generalizes or the topic of the article. On the other hand, the 'question' and 'surprise' categories have zero association with 'announcing principal findings'. 'Announcing principal findings' should have strong association with 'emphasis' or 'background' in AWA annotation. As an example, a sentence with 'principal findings' should be recognized by AWA as 'emphasis' or 'background' of certain research/facts. Furthermore, another category to be concerned is the 'indicating a gap'. This category has very weak association in all AWA annotations. It supposes to have high association in 'question', 'emphasis', 'position' or 'trend'.

7.4 Test 4: Association between AntMover and AWA over sub-corpora

This test is similar to test 5.1 but it restricted to category of document (sub-corpora). The result of the chisquare for each category of documents is as follow.

Category	Number of Sentences	Chisquare p-value	Null Hypothesis	AntMover and AWA Independent (Yes/No)
Agriculture	2064	0.001819	reject	No
Anthropology	1037	0.07275	accept	Yes
Archaeology	2056	2.2E-16	reject	No
Architecture	225	0.8425	accept	Yes
BiologicalSciences	1955	5.286E-08	reject	No
Business	3273	2.973E-08	reject	No
Chemistry	528	0.1797	accept	Yes
Classics	1688	2.461e-05	reject	No
ComparativeAmericanStudies	1649	0.009326	reject	No
ComputerScience	1081	0.03042	reject	No
CyberneticsElectronicEngineering	488	0.2763	accept	Yes
Economics	1940	5.771e-08	reject	No
Engineering	2757	4.367e-16	reject	No
English	2052	0.004935	reject	No
FoodSciences	537	0.02008	reject	No
Health	2341	2.065e-06	reject	No

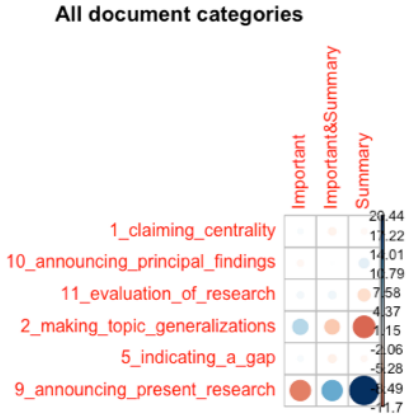
History	2763	1.565e-12	reject	No
HospitalityLeisureTourismManagement	2636	3.374e-05	reject	No
Law	4455	6.16e-12	reject	No
Linguistics	3145	2.2e-16	reject	No
Mathematics	600	0.01512	reject	No
Medicine	934	2.913e-05	reject	No
Meteorology	209	0.652	accept	Yes
other	1409	0.9929	accept	Yes
Philosophy	3873	0.001844	reject	No
Physics	1300	2.2e-16	reject	No
Planning	83	0.9972	accept	Yes
Politics	4350	2.426e-16	reject	No
Psychology	2547	5.301e-14	reject	No
Publishing	355	0.9006	accept	Yes
Sociology	3802	1.595e-05	reject	No

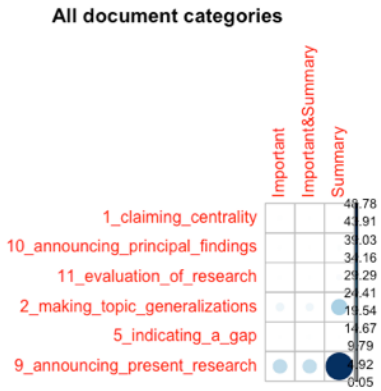
Result: Most categories show that AWA and AntMover are associated. There are some showing the association is independent but these categories all having less test subjects (number of sentences). Thus, the chisquare p-value might not be correct.

7.5 Association between AntMover and AWA (main category and sub category)

Association Test (Main Category)

X-squared = 856.2, df = 10, **p-value < 2.2e-16**

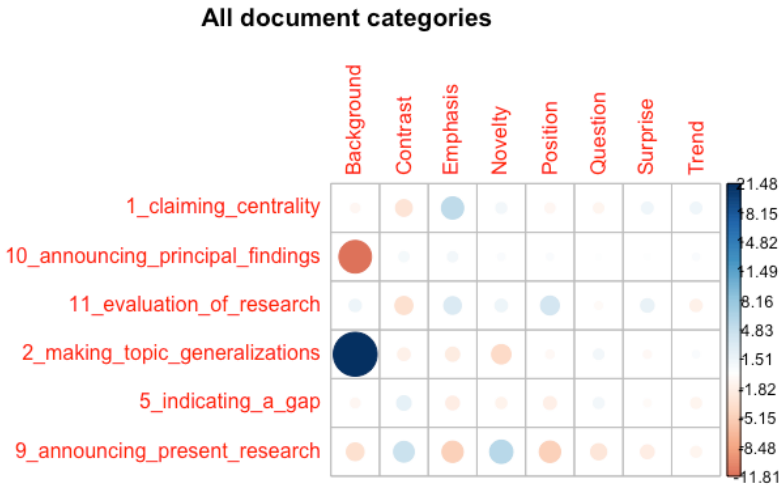


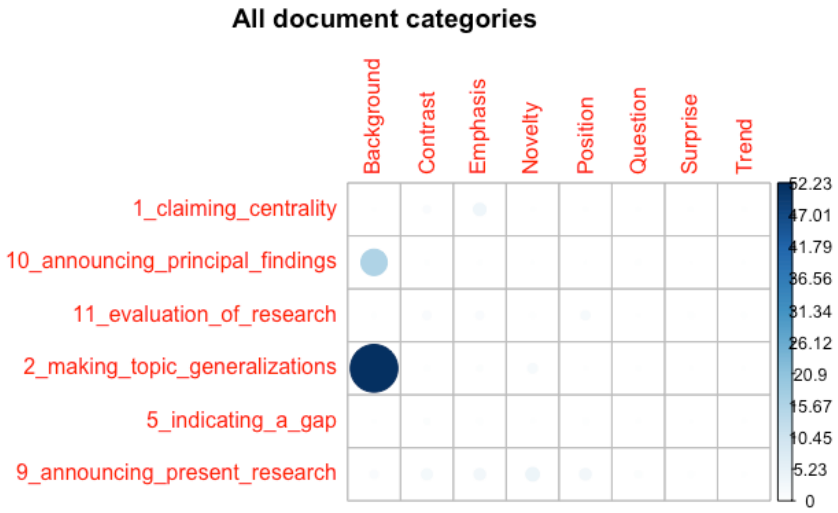


Result: The p-value is lower than 0.05. So, reject the null hypothesis. There is association between AntMover and AWA.

Association Test (Sub Category)

X-squared = 882.99, df = 35, **p-value < 2.2e-16**





Result: The p-value is lower than 0.05. So, reject the null hypothesis. There is association between AntMover and AWA.

8.0 Association between main categories and sub categories

8.1 AWA2

This test is to determine the association between the main categories and sub categories of annotation in AWA 2

	Background	Contrast	Emphasis	Novelty	Position	Question	Surprise	Trend	sub_other
Important	87	1333	325	85	100	114	28	78	69
Important&Summary	2	378	12	19	2	10	2	5	11
Summary	0	0	0	0	0	0	4	0	1504

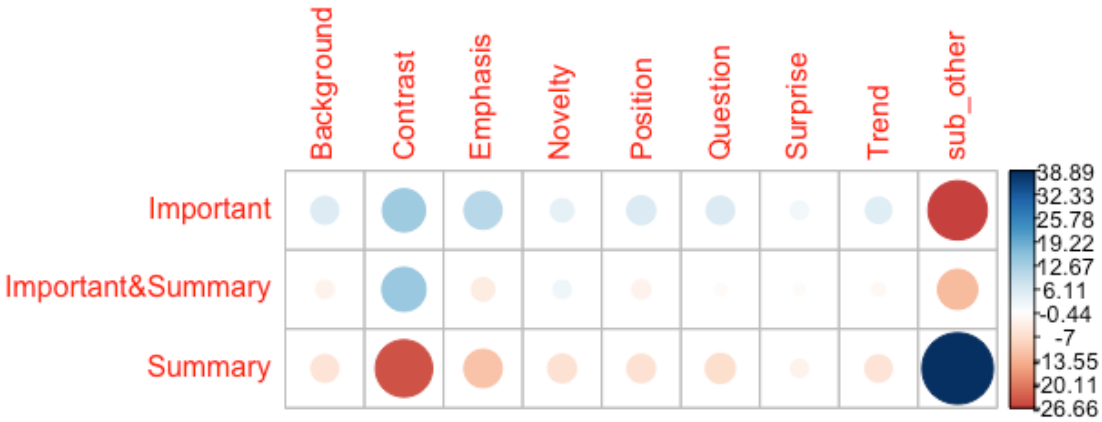
The number of sentences associated between main categories and sub categories. 'sub_other' column refers to sentences that only annotated with main categories and no any sub category.

Pearson's Chi-squared test

X-squared = 4016.8, df = 16, p-value < 2.2e-16

	Background	Contrast	Emphasis	Novelty	Position	Question	Surprise	Trend	sub_other
Important	5.755	13.985	10.869	3.982	6.201	5.906	2.327	5.086	-26.664
Important&Summary	-2.417	14.639	-3.962	2.411	-2.676	-0.861	-0.842	-1.276	-12.096
Summary	-5.675	-24.881	-11.042	-6.134	-6.075	-6.698	-2.367	-5.480	38.886

The Chi-square statistic for each cell.



8.2 AWA 3

This test is to determine the association between the main categories and sub categories of annotation in AWA 3

	sub_categories								
main_categories	AWA3_Background	AWA3_Contrast	AWA3_Emphasis	AWA3_Novelty	AWA3_Position	AWA3_Question	AWA3_Surprise	AWA3_Trend	sub_other
AWA3_Important	32	1296	128	36	53	30	3	18	0
AWA3_Important&Summary	15	273	36	10	14	2	2	1	0
AWA3_Summary	44	0	106	21	25	0	2	3	1746
main_other	227	0	427	131	170	0	11	66	0

The number of sentences associated between main categories and sub categories. ‘sub_other’ column refers to sentences that only annotated with main categories and no any sub category. Meanwhile, the ‘main_other’ row refers to sentences that only annotated with sub category only with no any main category.

Pearson's Chi-squared test

X-squared = 7054.4, df = 24, p-value < 2.2e-16

	sub_categories								
main_categories	AWA3_Background	AWA3_Contrast	AWA3_Emphasis	AWA3_Novelty	AWA3_Position	AWA3_Question	AWA3_Surprise	AWA3_Trend	sub_other
AWA3_Important	-6.995	34.951	-6.505	-3.512	-3.458	6.100	-1.172	-1.967	-23.780
AWA3_Important&Summary	-1.630	15.150	-1.971	-1.111	-1.100	-0.193	0.626	-2.112	-11.183
AWA3_Summary	-7.283	-24.898	-10.207	-6.470	-7.717	-3.556	-1.917	-5.388	40.213
main_other	19.656	-18.127	23.262	13.905	15.543	-2.589	3.724	11.082	-19.122

The Chi-square statistic for each cell.

In order to compare against AWA2 which does not have ‘main_other’ row, combine ‘main_other’ row with ‘AWA_Important’ row.

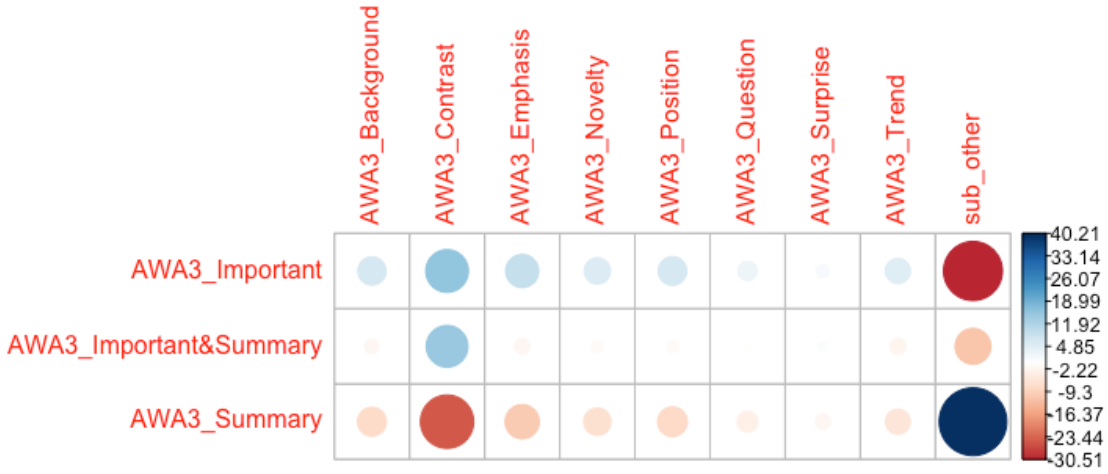
	AWA3_Background	AWA3_Contrast	AWA3_Emphasis	AWA3_Novelty	AWA3_Position	AWA3_Question	AWA3_Surprise	AWA3_Trend	sub_other
AWA3_Important	259	1296	555	167	223	30	14	84	0
AWA3_Important&Summary	15	273	36	10	14	2	2	1	0
AWA3_Summary	44	0	106	21	25	0	2	3	1746

Pearson's Chi-squared test

X-squared = 4356.7, df = 16, p-value < 2.2e-16

	AWA3_Background	AWA3_Contrast	AWA3_Emphasis	AWA3_Novelty	AWA3_Position	AWA3_Question	AWA3_Surprise	AWA3_Trend	sub_other
AWA3_Important	6.866	15.878	9.508	5.976	7.046	3.131	1.420	5.412	-30.514
AWA3_Important&Summary	-1.630	15.150	-1.971	-1.111	-1.100	-0.193	0.626	-2.112	-11.183
AWA3_Summary	-7.283	-24.898	-10.207	-6.470	-7.717	-3.556	-1.917	-5.388	40.213

The Chi-square statistic for each cell.



9.0 Reliability between AWA2 and AWA3

Similarities:

- Each sentence have only one main category annotation

  Query: select count(annotation_id) as 'total_main_annotation', sentence_id from sentence_annotation where annotation_id in (<main category ids>) group by sentence_id order by total_main_annotation desc;

- Each sentence have more than one sub category annotation

  Query: select count(annotation_id) as 'total_sub_annotation', sentence_id from sentence_annotation where annotation_id in (<sub category ids>) group by sentence_id order by total_main_annotation desc;

Differences:

- In AWA2, each annotated sentence must belong to one main category annotation
- In AWA3, not each annotated sentence belongs to main category annotation

9.1 Comparison between types of annotations (sub categories only)

In this test, the below setup is considered.

- the test subjects are sentences in Elsevier corpus
- annotation scope only covers the sub categories
- One tool may annotate one sentence with more sub categories than the other tool.

	Elsevier Corpus	
Annotation label	AWA3	AWA2
Background	318	89
Contrast	1569	1711
Emphasis	697	337
Novelty	198	104
Position	262	102
Question	32	124
Summary	1917	1508
Surprise	18	34
Trend	88	83
Total	5099	4092

Additional info:

Total distinct different sentences annotated (sub categories) by AWA3 but not AWA2

Query:

select count(distinct sentence_id) from sentence_annotation where sentence_id in (select sentence_id from sentence where corpus_id=1)

and annotation_id in (18,19,20,21,22,23,24,25)

and sentence_id not in (select sentence_id from sentence_annotation where sentence_id in (select sentence_id from sentence where corpus_id=1)

and annotation_id in (10,11,12,13,14,15,16,17));

Result: 1103

Total distinct different sentences annotated (sub categories) by AWA2 but not AWA3

Query:

select count(distinct sentence_id) from sentence_annotation where sentence_id in (select sentence_id from sentence where corpus_id=1)

and annotation_id in (10,11,12,13,14,15,16,17)

and sentence_id not in (select sentence_id from sentence_annotation where sentence_id in (select sentence_id from sentence where corpus_id=1)

and annotation_id in (18,19,20,21,22,23,24,25));

Result: 834

9.2 Reliability test on annotation level

	scope: sentences annotated with AWA2 and AWA3				scope: sentences annotated with AWA2 or AWA3			
Annotation labels	total sentences	agreement (%)	Cohen's kappa	krippendorff's alpha	total sentences	agreement (%)	Cohen's kappa	krippendorff's alpha
Important	1277	65.4	-0.034	-0.209	2417	34.5	-0.37	-0.486
Summary	1630	80.1	-0.057	-0.11	2120	61.6	-0.168	-0.238
Important&Summary	497	33.8	-0.35	-0.493	533	31.5	-0.39	-0.519
Background	160	28.8	-0.116	-0.549	361	12.7	-0.259	-0.771
Contrast	1396	65.1	-0.205	-0.211	2371	38.3	-0.44	-0.445
Emphasis	434	62.4	-0.058	-0.23	763	37.3	-0.202	-0.454
Position	114	71.9	-0.106	-0.158	293	24.2	-0.223	-0.607
Novelty	187	38	-0.162	-0.446	220	37.3	-0.202	-0.454
Question	98	18.4	-0.128	-0.681	138	13	-0.218	-0.763
Trend	60	75	-0.125	-0.133	126	35.7	-0.471	-0.468
Surprise	18	61.1	-0.105	-0.207	41	26.8	-0.355	-0.558

The table above shows the reliability results from two different subsets.

- Sentences annotated with AWA2 and AWA3
    - Include sentences annotated by both tools
    - Exclude sentences that annotated by AWA2 only or AWA3 only

- Sentences annotated with AWA2 or AWA3
    - Include sentences annotated by both tools
    - Include sentences that annotated by AWA2 only or AWA3 only

Result: As shown from above, the kappa and alpha values are all having negatives values even though some percentage agreements are high. This is because both methods considered the possibility of agreement happen by chance. Therefore, I found out that kappa and alpha testing for reliability are not suitable for AWA2 and AWA3 because both tools will not annotate by chance. As a result, the kappa and alpha values will not be accurate. The **correct method to test the reliability** will be just the **percentage agreement method (agreement column)** which count the percentage of sentences having same annotation in both AWA2 and AWA3. **Based on the agreement column, the reliability of AWA2 and AWA3 in every annotation is low**.

9.3 Reliability test on all annotation level (AWA2 and AWA3 annotated sentences)

	AWA3											
AWA2	Important	Summary	Important&Summary	Background	Contrast	Emphasis	Position	Novelty	Question	Trend	Surprise	No_annotation
Important	835											420
Summary		1305										52
Important&Summary			168									241
Background				46								9
Contrast					909							283
Emphasis						271						13
Position							71					15
Novelty								82				7
Question									18			74
Trend										45		10
Surprise											11	6
No_annotation	22	273	88	105	204	150	101	25	6	5	1	

All categories
Percentage agreement
Subjects = 5871
Raters = 2
%-agree = 64.1
Kappa = 0.566
alpha = 0.521

Main categories
Percentage agreement
Subjects = 3404
Raters = 2
%-agree = 67.8
Kappa = 0.533
alpha = 0.528

Sub categories
Percentage agreement
Subjects = 2467
Raters = 2
%-agree = 58.9
Kappa = 0.425
alpha = 0.219

9.4 Reliability test on all annotation level (AWA2 or AWA3 annotated sentences)

	AWA3											
AWA2	Important	Summary	Important&Summary	Background	Contrast	Emphasis	Position	Novelty	Question	Trend	Surprise	No_annotation
Important	835											1121
Summary		1305										203
Important&Summary			168									260
Background				46								43
Contrast					909							802
Emphasis						271						66
Position							71					31
Novelty								82				22
Question									18			106
Trend										45		38
Surprise											11	23
No_annotation	461	612	105	272	660	426	191	116	14	43	7	

All categories	**Main categories**	**Sub categories**
Percentage agreement	Percentage agreement	Percentage agreement
Subjects = 9383	Subjects = 5070	Subjects = 4313
Raters = 2	Raters = 2	Raters = 2
%-agree = 40.1	%-agree = 45.5	%-agree = 33.7
Kappa = 0.263	Kappa = 0.235	Kappa = 0.0956
alpha = -0.08	alpha = -0.115	alpha = -0.105

10.0 Association between AntMover and AWA3

10.1 Test 1: Association between AntMover's moves and AWA's annotations

The test will compare the association between AntMover's moves and AWA's annotations.

Null hypothesis, $\mathbf{H_0}$: AntMover's moves and AWA's annotations are independent

Alternate hypothesis, $\mathbf{H_1}$: AntMover's moves and AWA's annotations are dependent

	AWA3's annotations											
	Main category			Sub category								
AntMover's moves	AWA3_Important	AWA3_Summary	AWA3_Important&Summary	AWA3_Background	AWA3_Contrast	AWA3_Emphasis	AWA3_Novelty	AWA3_Position	AWA3_Question	AWA3_Surprise	AWA3_Trend	Total
Claiming_centrality	5	0	0	0	5	9	0	8	0	0	0	27
Announcing_principal_findings	605	788	117	135	722	291	81	93	14	8	39	2893
Evaluation_of_research	421	408	84	125	505	270	62	126	6	6	25	2038
Making_topic_generalization	173	52	11	33	184	72	18	23	9	1	18	594
Indicating_a_gap	13	2	1	1	14	1	0	0	1	0	1	34
Announcing_present_research	79	668	60	24	139	54	37	12	2	3	5	1083
Total	1296	1918	273	318	1569	697	198	262	32	18	88	6669

Chisquare test result:

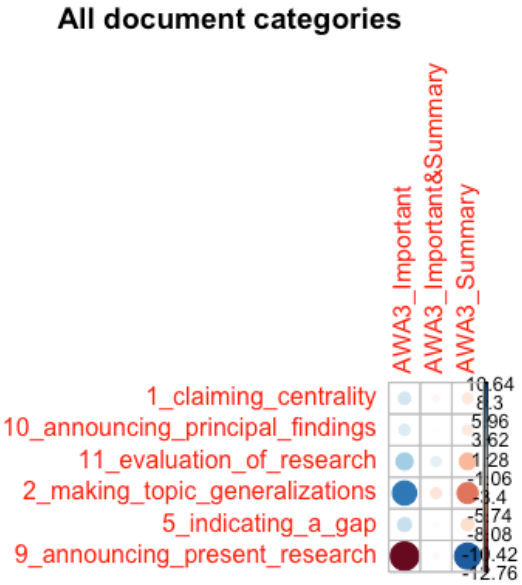
X-squared = 1004.7, df = 50, p-value < 2.2e-16

Result: The p-value is smaller than 0.05, so reject null hypothesis. There is a significant association between AntMover's moves and AWA3's annotations. The values of contributing cells to the total Chi-square score is shown at section 10.2.

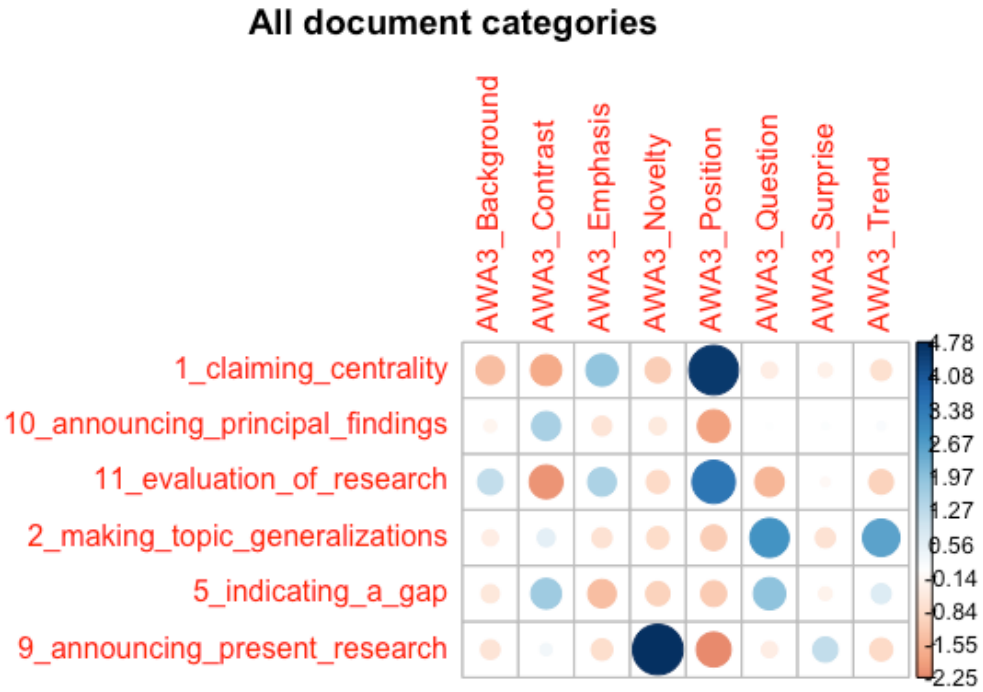
10.2 Test 2: Positive and negative association between AntMover and AWA3

10.3 Association between AntMover and AWA (main category and sub category)

Main category
Pearson's Chi-squared test
X-squared = 475.9, df = 10, p-value < 2.2e-16



Sub category
Pearson's Chi-squared test
X-squared = 126.84, df = 35, p-value = 2.445e-12



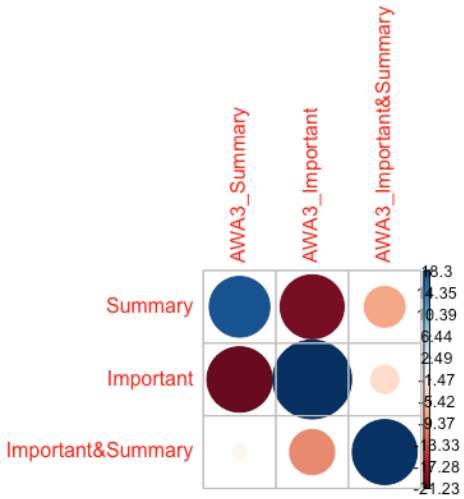
11.0 Association between AWA2 and AWA3

11.1 Main categories

	AWA3_Summary	AWA3_Important	AWA3_Important&Summary
Summary	1305	8	35
Important	53	835	53
Important&Summary	220	14	168

Pearson's Chi-squared test
X-squared = 2781.1, df = 4, p-value < 2.2e-16

	AWA3_Summary	AWA3_Important	AWA3_Important&Summary
Summary	18.301	-20.333	-8.233
Important	-21.234	30.923	-3.860
Important&Summary	-1.025	-10.077	20.982



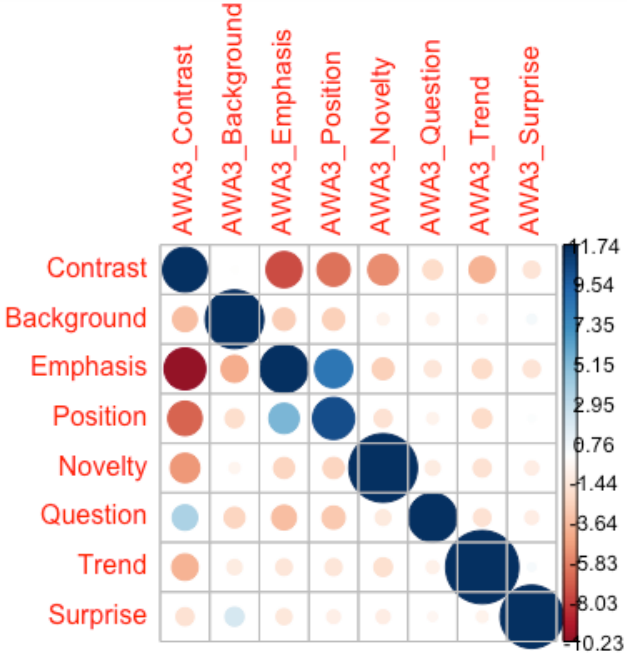
11.2 Sub categories

	AWA3_Contrast	AWA3_Background	AWA3_Emphasis	AWA3_Position	AWA3_Novelty	AWA3_Question	AWA3_Trend	AWA3_Surprise
Contrast	909	81	164	72	22	6	9	6
Background	12	46	4	0	2	0	1	1
Emphasis	90	9	271	128	13	2	5	1
Position	25	5	79	71	4	1	0	2
Novelty	26	7	18	6	82	0	0	0
Question	90	1	10	3	3	18	0	0
Trend	12	2	11	4	0	0	45	1
Surprise	5	4	2	1	0	0	0	11

Pearson's Chi-squared test

X-squared = 3969.5, df = 49, p-value < 2.2e-16

	AWA3_Contrast	AWA3_Background	AWA3_Emphasis	AWA3_Position	AWA3_Novelty	AWA3_Question	AWA3_Trend	AWA3_Surprise
Contrast	11.739	-0.094	-7.636	-6.399	-5.460	-2.187	-4.030	-1.648
Background	-3.549	20.231	-2.897	-2.798	-0.785	-0.861	-0.505	0.509
Emphasis	-10.226	-4.230	13.676	8.469	-2.725	-1.587	-2.211	-1.721
Position	-6.917	-2.033	5.382	10.367	-1.854	-0.760	-2.161	0.220
Novelty	-5.061	-0.657	-2.521	-2.583	27.674	-1.250	-1.863	-1.128
Question	3.743	-2.487	-3.538	-3.071	-1.388	14.003	-1.767	-1.070
Trend	-4.054	-1.290	-1.543	-1.641	-1.983	-0.918	31.515	0.378
Surprise	-1.850	2.066	-1.448	-1.046	-1.098	-0.508	-0.758	23.513



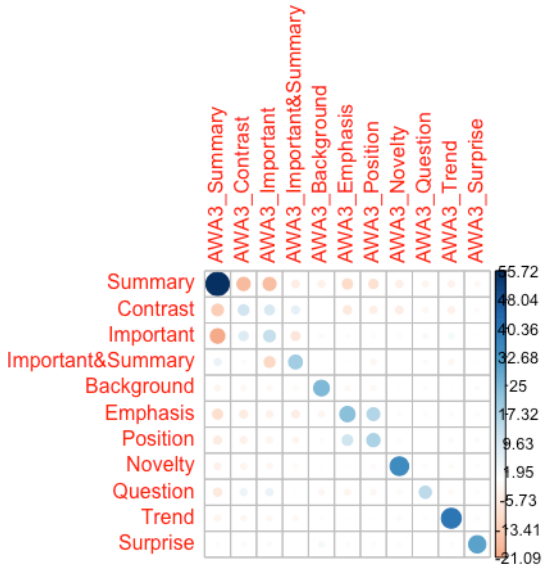
11.3 All categories

	AWA3										
	Summary	Contrast	Important	Important&Summary	Background	Emphasis	Position	Novelty	Question	Trend	Surprise
Summary	1305	43	8	35	20	35	4	10	0	2	1
Contrast	211	909	704	205	81	164	72	22	6	9	6
Important	53	888	835	53	104	299	145	77	22	46	9
Important&Summary	220	182	14	168	27	87	23	20	2	2	2
Background	6	12	10	2	46	4	0	2	0	1	1
Emphasis	25	90	86	4	9	271	128	13	2	5	1
Position	5	25	23	2	5	79	71	4	1	0	2
Novelty	14	26	21	5	7	18	6	82	0	0	0
Question	1	90	76	14	1	10	3	3	18	0	0
Trend	6	12	11	1	2	11	4	0	0	45	1
Surprise	5	5	3	2	4	2	1	0	0	0	11

Pearson's Chi-squared test

X-squared = 12102, df = 100, p-value < 2.2e-16

	AWA3										
	Summary	Contrast	Important	Important&Summary	Background	Emphasis	Position	Novelty	Question	Trend	Surprise
Summary	55.722	-17.538	-17.011	-5.32	-4.451	-10.214	-8.371	-4.714	-2.948	-3.867	-1.991
Contrast	-13.397	10.876	9.213	5.85	-0.449	-6.581	-4.891	-5.319	-2.174	-3.906	-1.125
Important	-21.09	8.301	13.363	-7.625	1.453	0.595	0.886	1.003	1.797	2.384	-0.323
Important&Summary	4.646	-1.174	-11.361	19.168	0.073	0.188	-2.658	-0.06	-1.157	-2.447	-0.557
Background	-2.846	-2.185	-1.797	-1.279	24.856	-1.805	-2.114	-0.185	-0.706	-0.073	1.157
Emphasis	-9.553	-6.048	-4.022	-5.357	-2.86	23.35	16.225	-1.014	-0.91	-1.096	-0.953
Position	-6.109	-4.302	-3.309	-2.955	-0.983	10.897	17.493	-0.778	-0.255	-1.667	1.231
Novelty	-3.958	-3.128	-2.674	-1.637	0.246	-0.538	-1.143	35.001	-1.031	-1.514	-0.842
Question	-6.677	4.301	4.61	0.469	-2.414	-2.951	-2.506	-1.182	14.758	-1.664	-0.925
Trend	-3.138	-2.558	-1.907	-1.873	-0.722	0.118	-0.427	-1.589	-0.743	40.134	1.041
Surprise	-0.793	-1.273	-1.48	0.082	2.604	-0.91	-0.571	-0.946	-0.443	-0.65	30.068



References

Fulcher, G. (2010). Practical Language Testing. London: Hodder Education.

Krippendorff, K. (2004). Content analysis: An introduction to its methodology (2nd ed.). Thousand Oaks, CA: Sage.

http://www.statisticshowto.com/wilcoxon-signed-rank-test/