Database Schema

(Corpora and annotation tools)

Prepared by: Yoong Kuan Goh (Andrew)

Department: Connected Intelligence Center in UTS

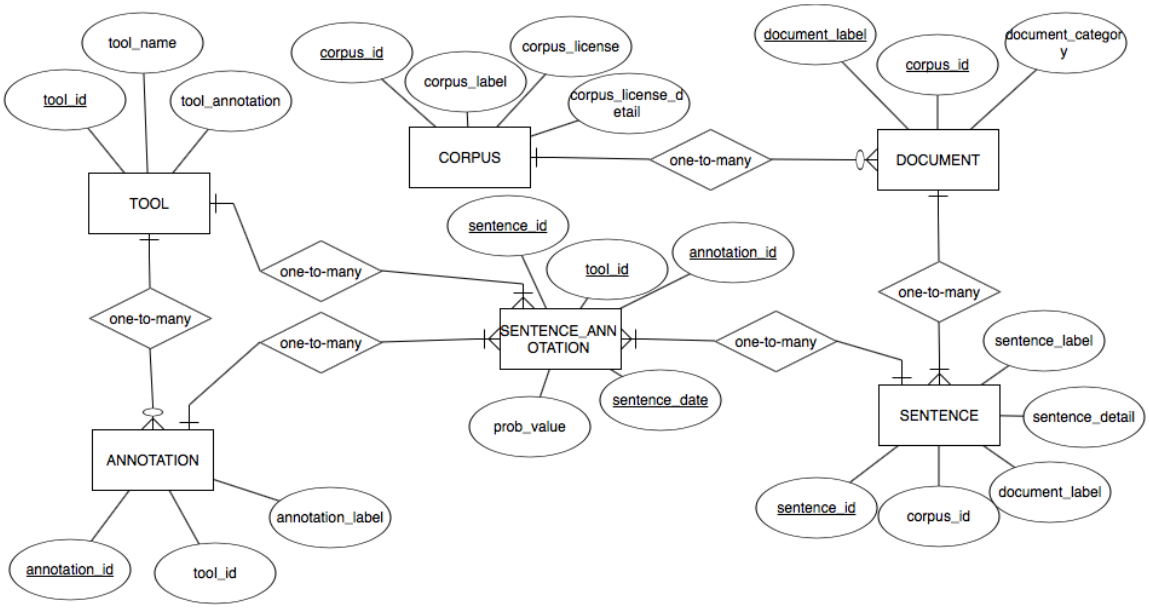
Contents

1. Introduction

This report describes the structure of a database scheme that been setup for the annotation analysis project involving four different annotation tools across three different corpora. Each corpus is broken down to document and sentence level. A MySQL database is used to store the annotation results.

Database name: AWA

ERD diagram:



2.0 Annotation tools

The annotation tools are AntMover, AWA, AWA3 and RWT. The information of these annotation tools is being stored in a table called 'TOOL'.

Field	Type	Null	Key	Default	Extra
tool_id	int(11)	NO	PRI	NULL	auto_increment
tool_name	varchar(50)	YES		NULL	
tool_annotation	varchar(50)	YES		NULL	

Fields:

a) tool_id → store the id of each tool

b) tool_name → store the name of each tool

c) tool_annotation → a short description of each tool

3.0 Annotation types

Each annotation tool has specific annotation scheme. The number of annotation type in each scheme may be different. These data are stored in a table called 'ANNOTATION' and this table is the child table of 'TOOL' table.

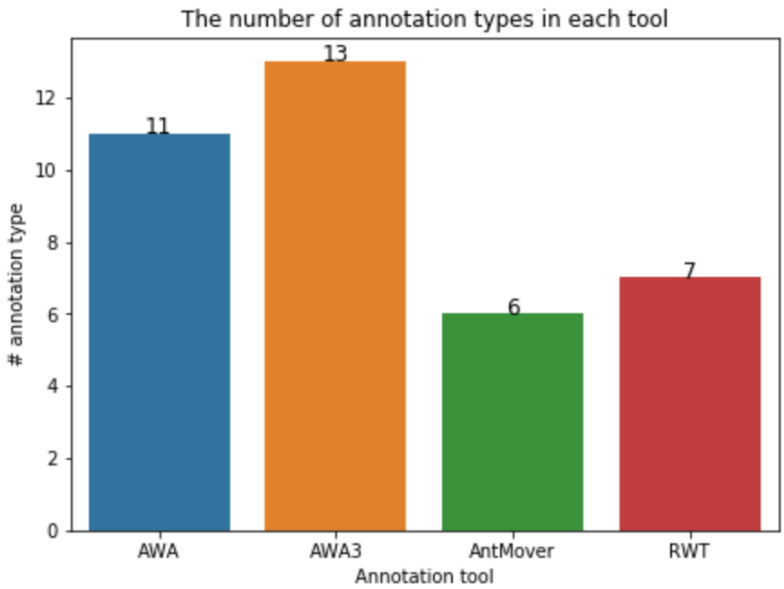
Field	Type	Null	Key	Default	Extra
annotation_id	int(11)	NO	PRI	NULL	auto_increment
annotation_label	varchar(50)	YES		NULL	
tool_id	int(11)	YES	MUL	NULL	

Fields:

a) annotation_id → store the id of each annotation type

b) annotation_label → store the name of each annotation type

c) tool_id → store the referencing tool_id from the 'TOOL' table



4.0 Corpus

The corpora used as the raw data of the annotation analysis are Elsevier (OA-STM), BAWE and PMC. The table that stored the type of corpus is 'CORPUS'.

Field	Type	Null	Key	Default	Extra
corpus_id	int(11)	NO	PRI	NULL	auto_increment
corpus_label	varchar(50)	YES		NULL	
corpus_license	tinyint(1)	YES		NULL	
corpus_license_detail	varchar(250)	YES		NULL	

Fields:

a) corpus_id → store the id of each corpus

b) corpus_label → store the name of each corpus

c) corpus_license → state whether licensing is required for the corpus. 0 = no and 1 = yes

d) corpus_license_detail → describe about the license.

5.0 Corpus (Document)

Each corpus contains a lot of files. Each file is called as a document. The document details are stored in a table called 'DOCUMENT'. This table is a child table of 'CORPUS' table.

Field	Type	Null	Key	Default	Extra
document_label	varchar(100)	NO	PRI	NULL	
document_category	varchar(50)	YES		NULL	
corpus_id	int(11)	NO	PRI	NULL	

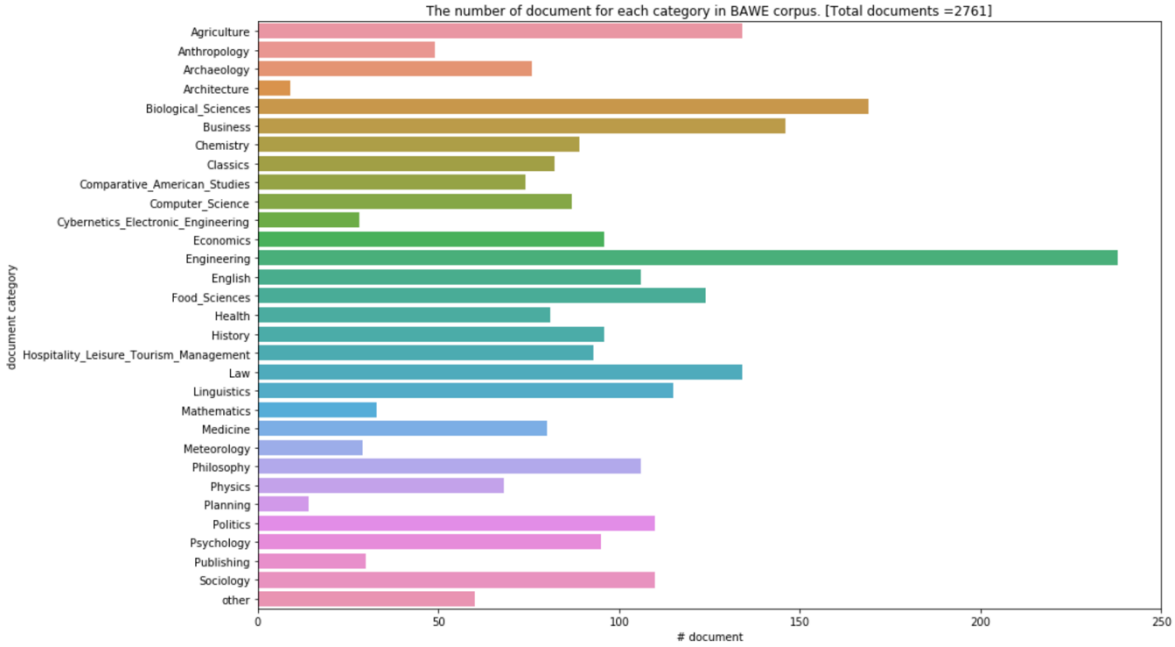
Fields:

a) document_label → store the name of the file/document

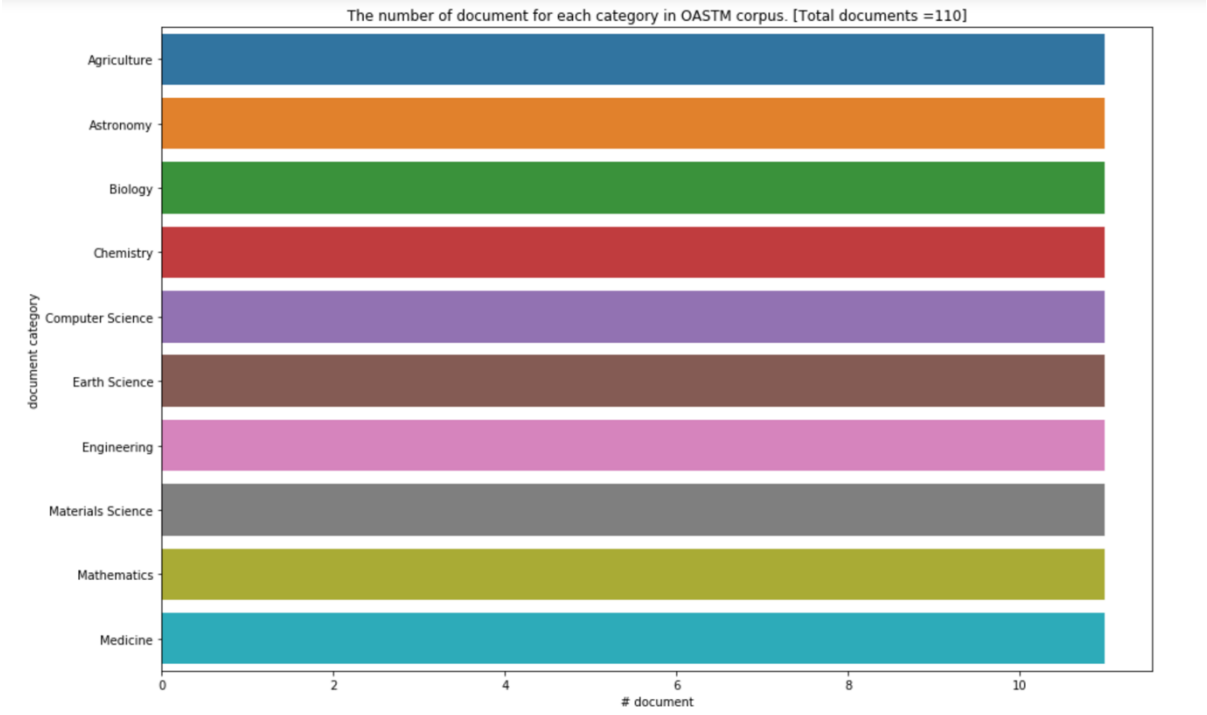
b) document_category → store the category of the document

c) corpus_id → store the referencing corpus_id from the 'TOOL' table

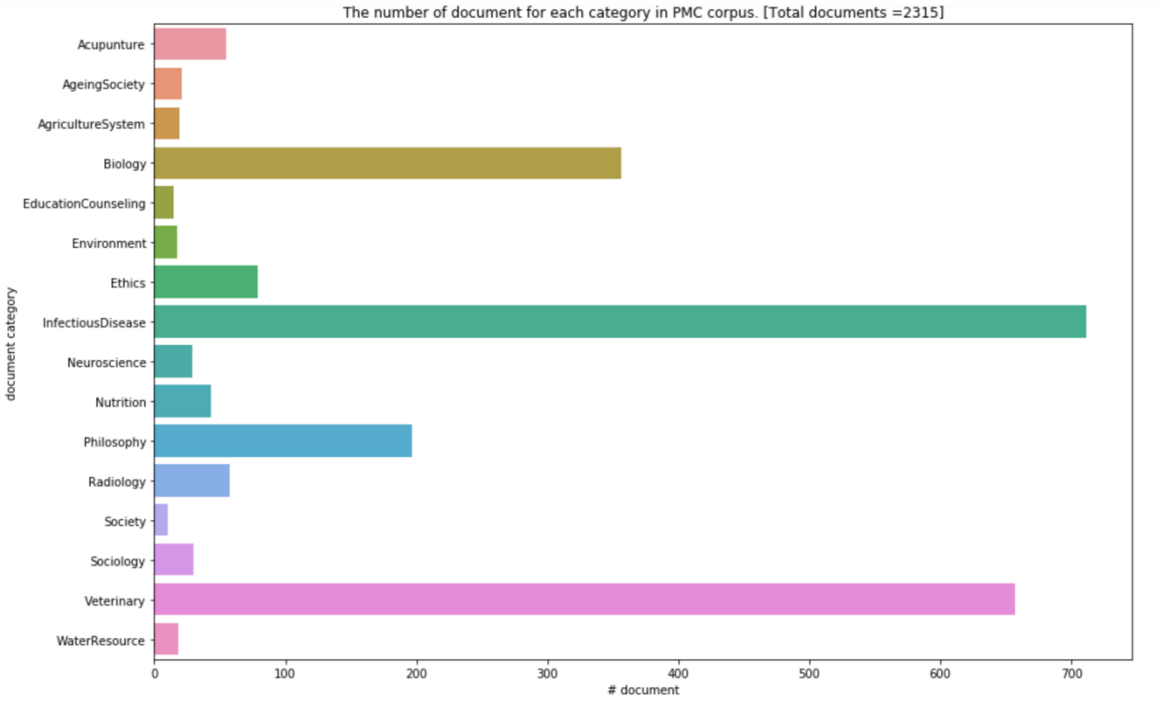
a) BAWE corpus



b) Elsevier/OA-STM corpus



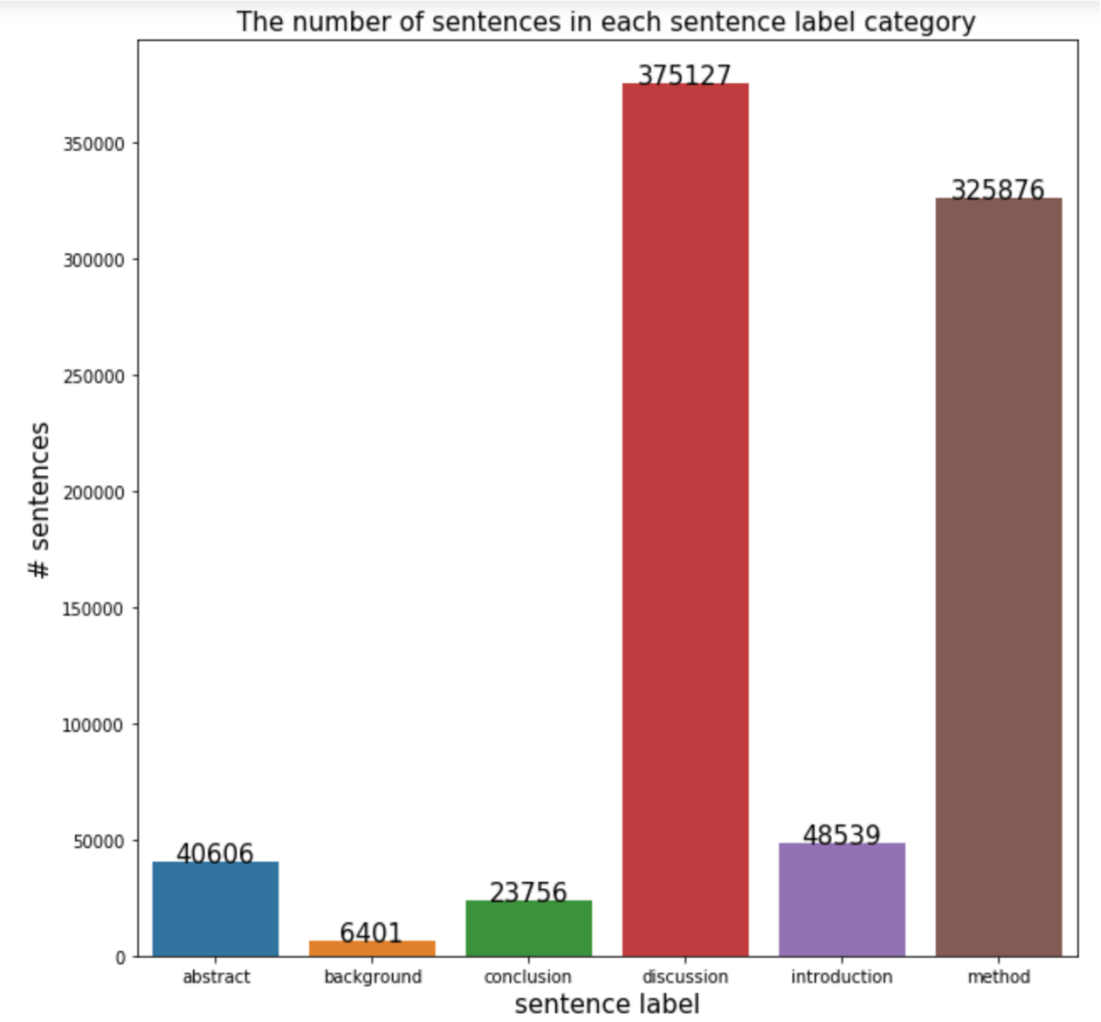
c) PMC corpus

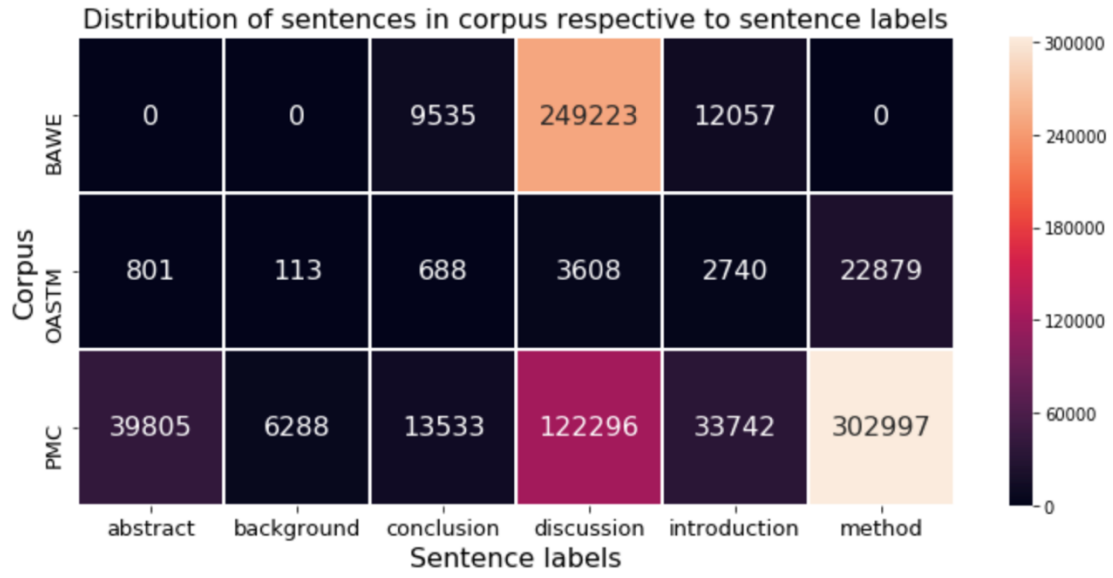


6.0 Sentences

Each document is broken down into sentences. Each sentence is stored in a table called 'SENTENCE'. The document labels of the sentences are stored together in the same table for easy referencing. Each sentence can be categorized into either 'abstract', 'introduction', 'background', 'discussion', 'method' or 'conclusion' as sentence label.

Field	Type	Null	Key	Default	Extra
sentence_id	bigint(20)	NO	PRI	NULL	auto_increment
sentence_detail	varchar(2500)	YES	MUL	NULL	
sentence_label	varchar(350)	YES		NULL	
document_label	varchar(100)	YES	MUL	NULL	
corpus_id	int(11)	YES		NULL	





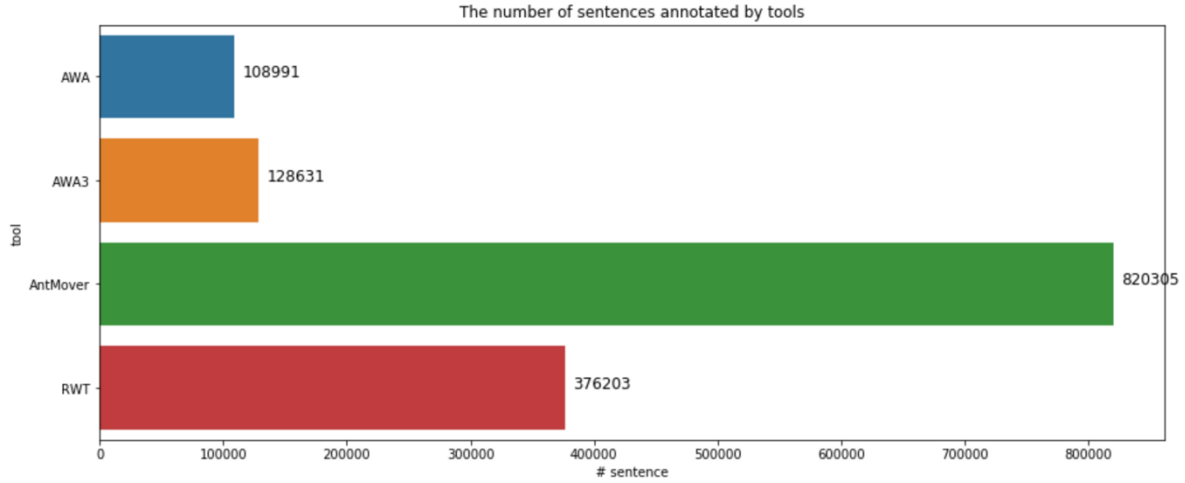
Total number of sentences: **820,305**

Total number of distinct sentences: **629,275**

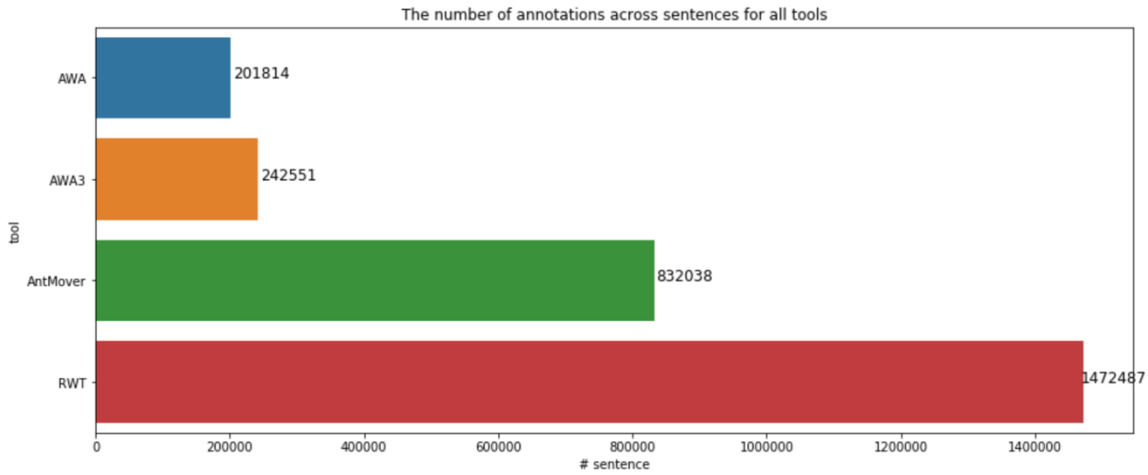
7.0 Sentence Annotations

Each sentence in 'SENTENCE' table will be annotated with all the tools from the 'TOOL' table. The annotated sentences will be stored in 'SENTENCE_ANNOTATION' table. Since RWT annotation involves the probabilistic values for each of its annotation, a column called 'prob_value' is created to store the probabilities. Sentences annotated by other tools apart from RWT will always have probability value as 1.

Field	Type	Null	Key	Default	Extra
sentence_id	bigint(20)	NO	PRI	NULL	
tool_id	int(11)	NO	PRI	NULL	
sentence_date	date	NO	PRI	NULL	
annotation_id	int(11)	NO	PRI	NULL	
prob_value	float	YES		0	



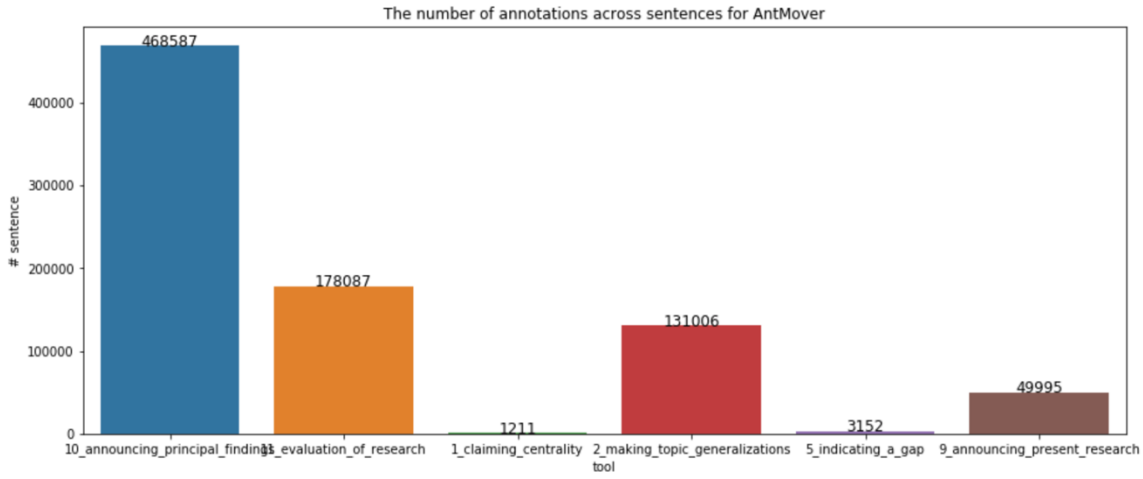
From the above graph, not all tools will annotate every sentence. Only AntMover tool will annotated every sentence. The tool that annotate the least sentences is AWA followed by AWA3 and RWT. The different between AWA and AWA3 is 18.02% (19640 sentences). In total **820,305** sentences, **AWA3 only annotate 15.68% of the sentences**. This shows that there are still plenty of improvement or enhancement can be made to AWA3 so that it can recognize and annotate more sentences.



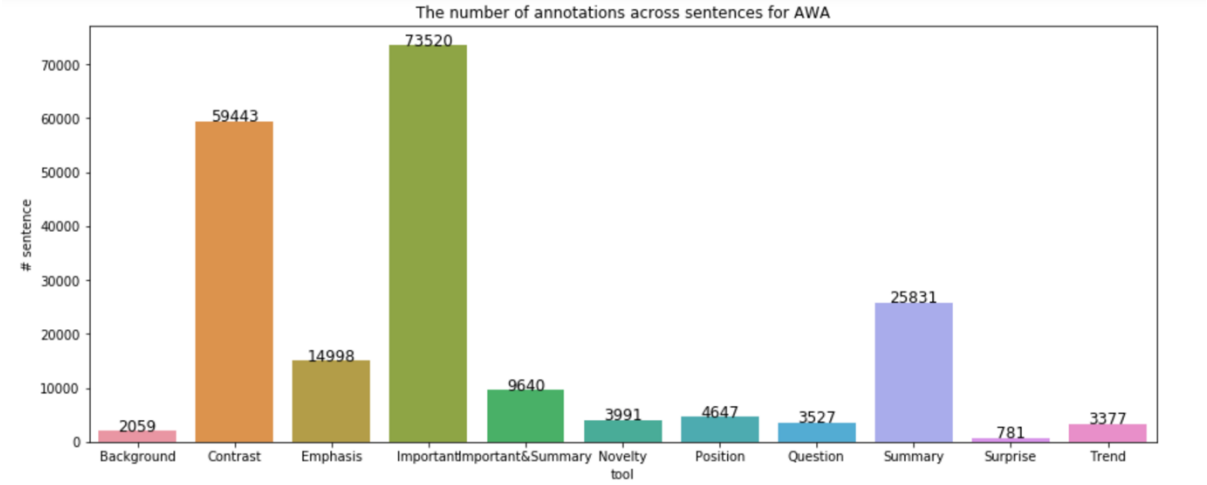
The above graph shows the total number of annotations in all the sentences. The RWT has the highest number of annotations because each sentence is annotated and given different probabilities values for each annotation. As an example, sentence A would be annotated and be given 7 probabilities value that show how likely this sentence being classify in each RWT annotations. So, a sentence will be annotated with 7 different annotation with different probability value.

Meanwhile, even though AntMover is supposed to annotated every annotation with only one annotation but surprisingly the number of annotations is higher than the total number of sentences. This is because, AntMover annotated a same sentence but in different files with different annotations. As an example, a sentence “Click here for additional data file” from file A had been annotated with “making_topic_generalizations” annotation but being annotated with “announcing_present_research” annotation at file B. Therefore, a same sentence had been considered annotated twice in this case.

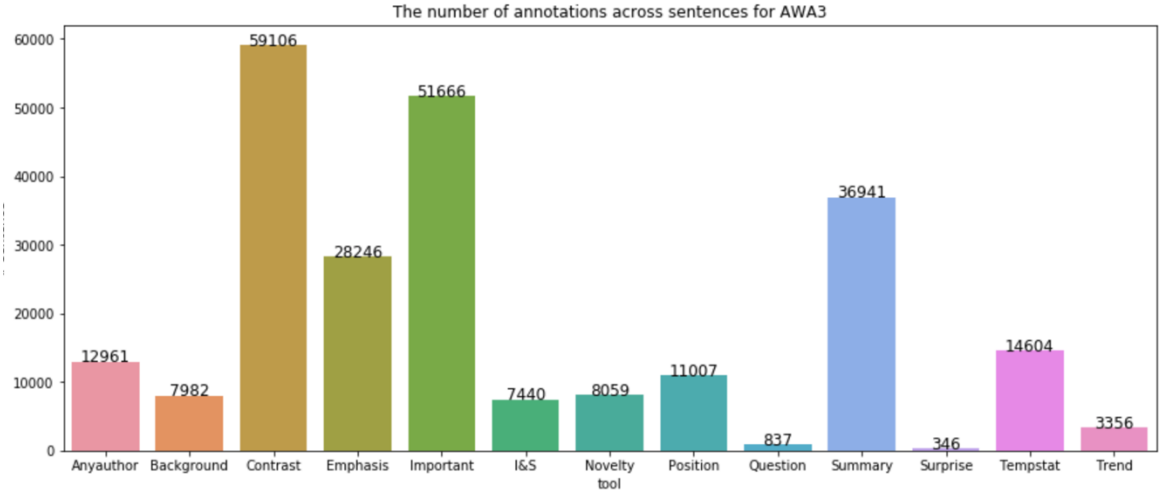
7.1 AntMover annotation distribution



7.2 AWA annotation distribution

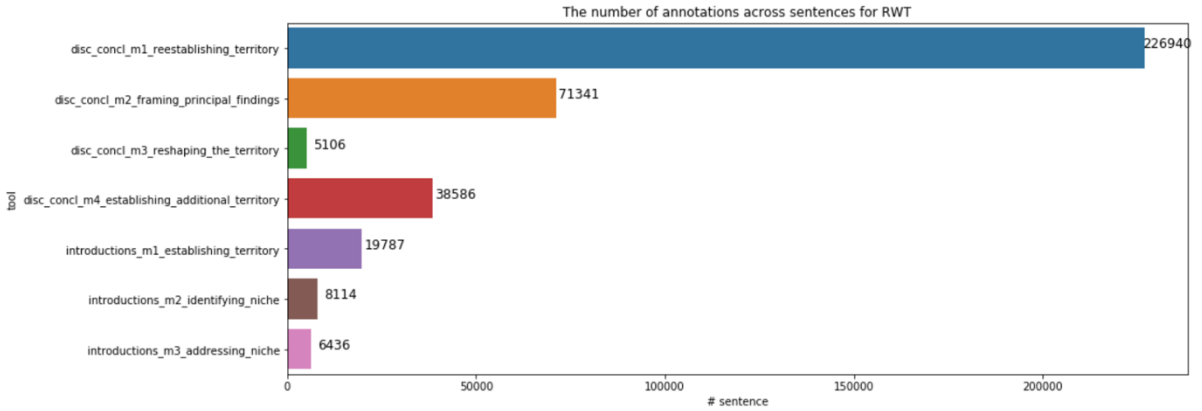


7.3 AWA3 annotation distribution



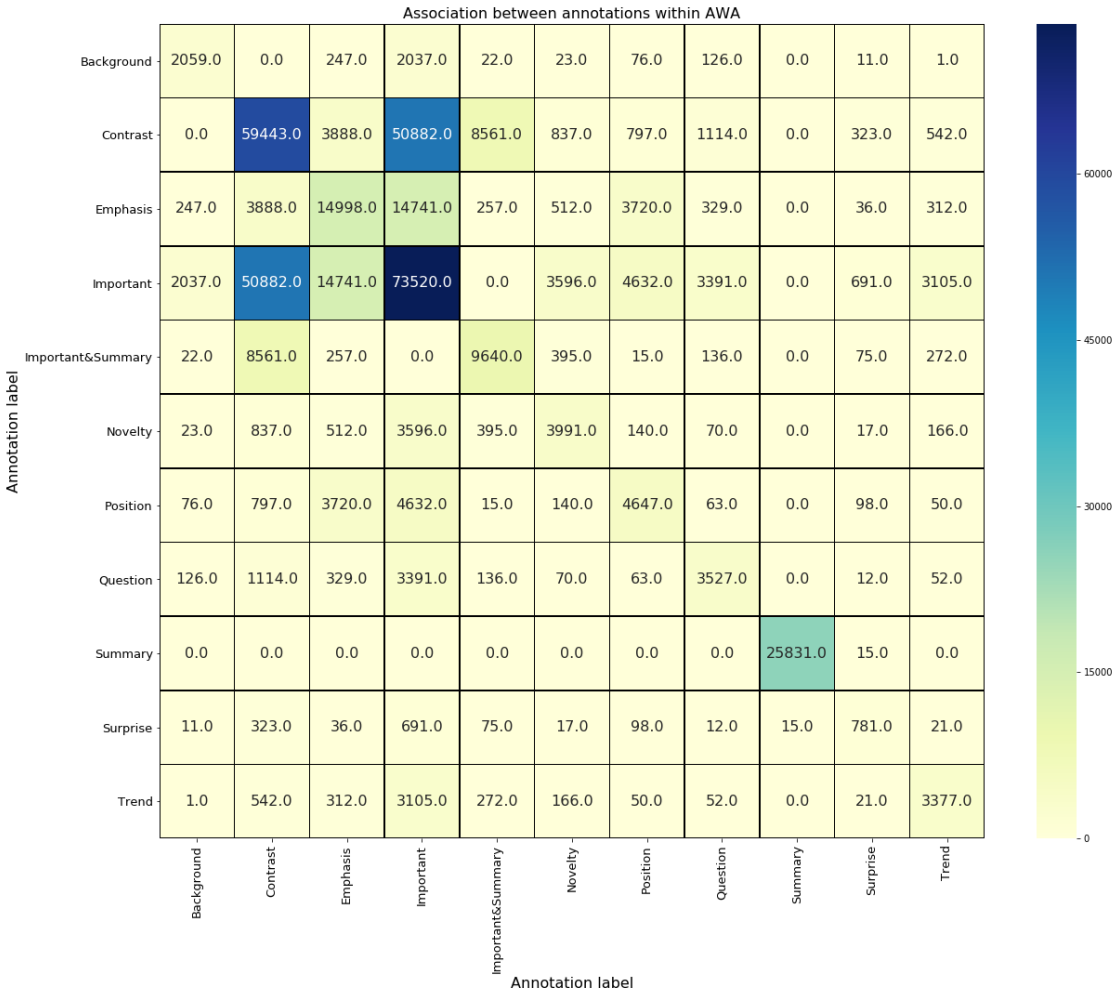
7.4 RWT annotation distribution

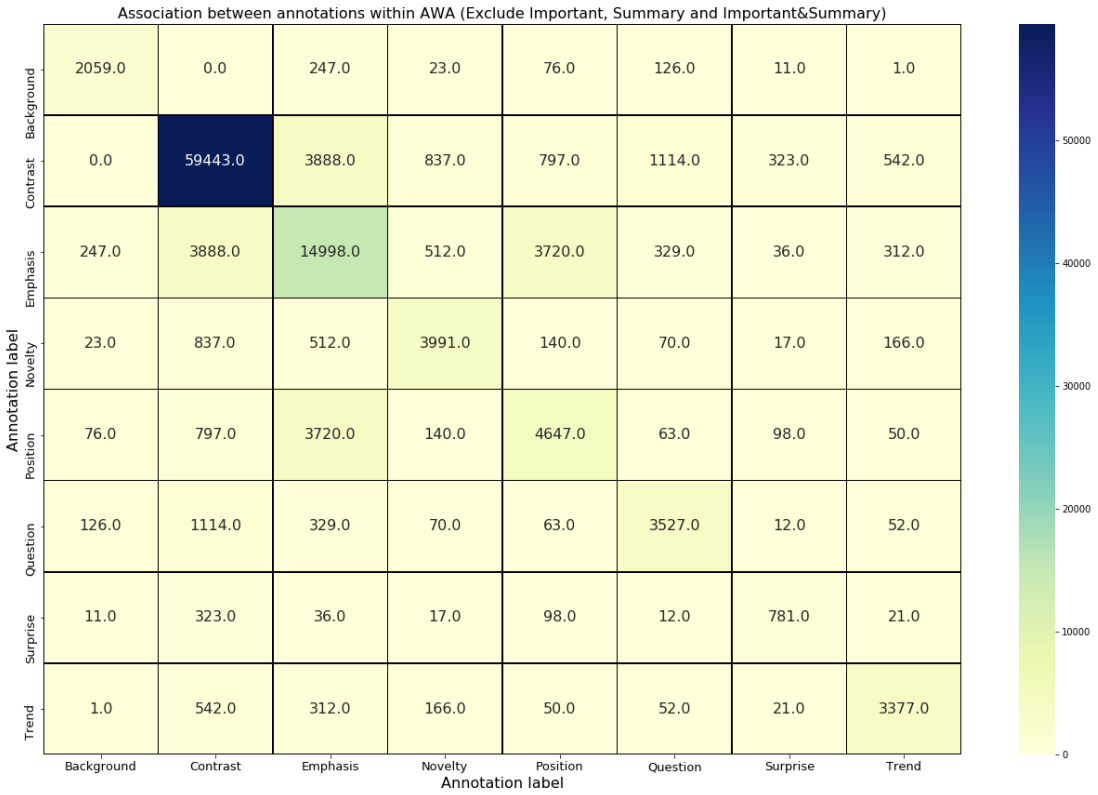
As each sentence would be given probability value for each annotation, only the record with the highest probability for each sentence will be taken in this distribution.



8.0 Association within AWA

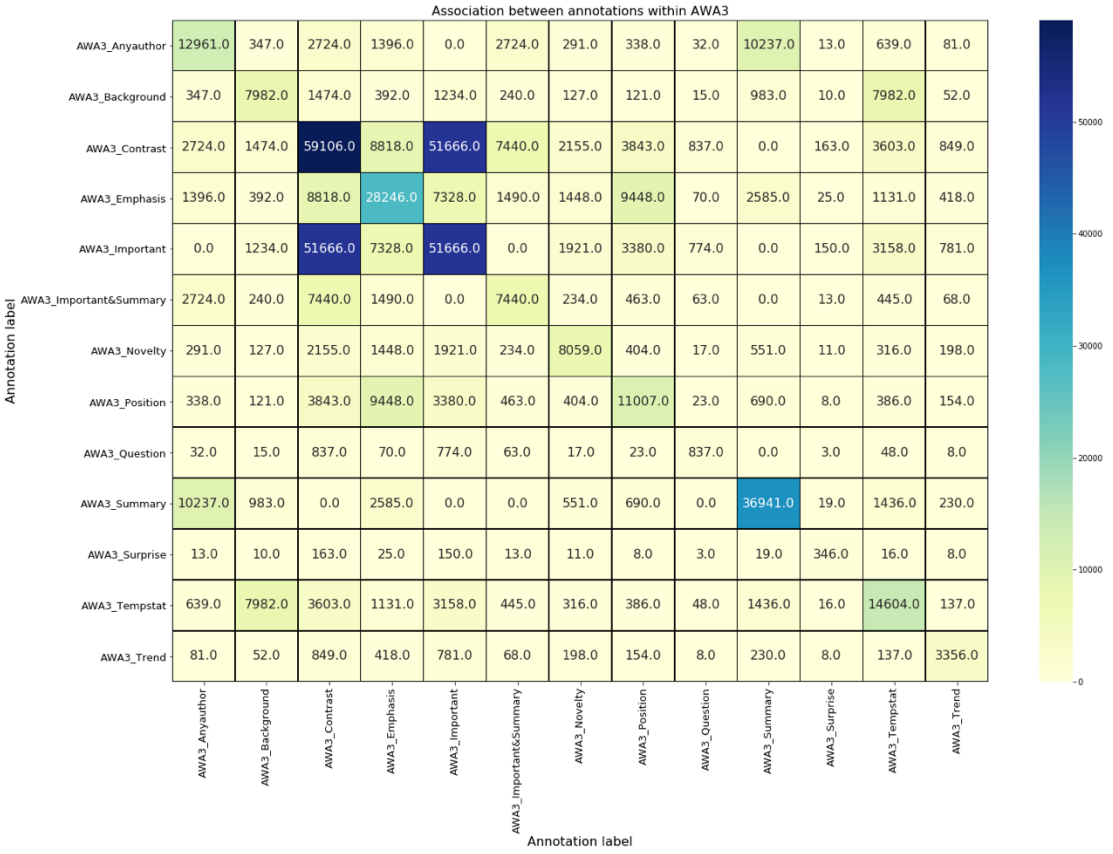
Since AWA tool annotated a sentence with different possible annotations, it is worth to see the association between these annotations. As shown in figure below, it seems that any sentence annotated with ‘Contrast’ will also be annotated with ‘Important’. Apart from that, most of the sentences annotated with ‘Emphasis’ will also be annotated with ‘Important’. Meanwhile, the we also can see that ‘Position’ annotation does have strong association with not only ‘Important’ annotation but also ‘Emphasis’. Suppose two annotations are having very strong association, then either one is not necessary to be included as another separate annotation as it just cause redundancy.

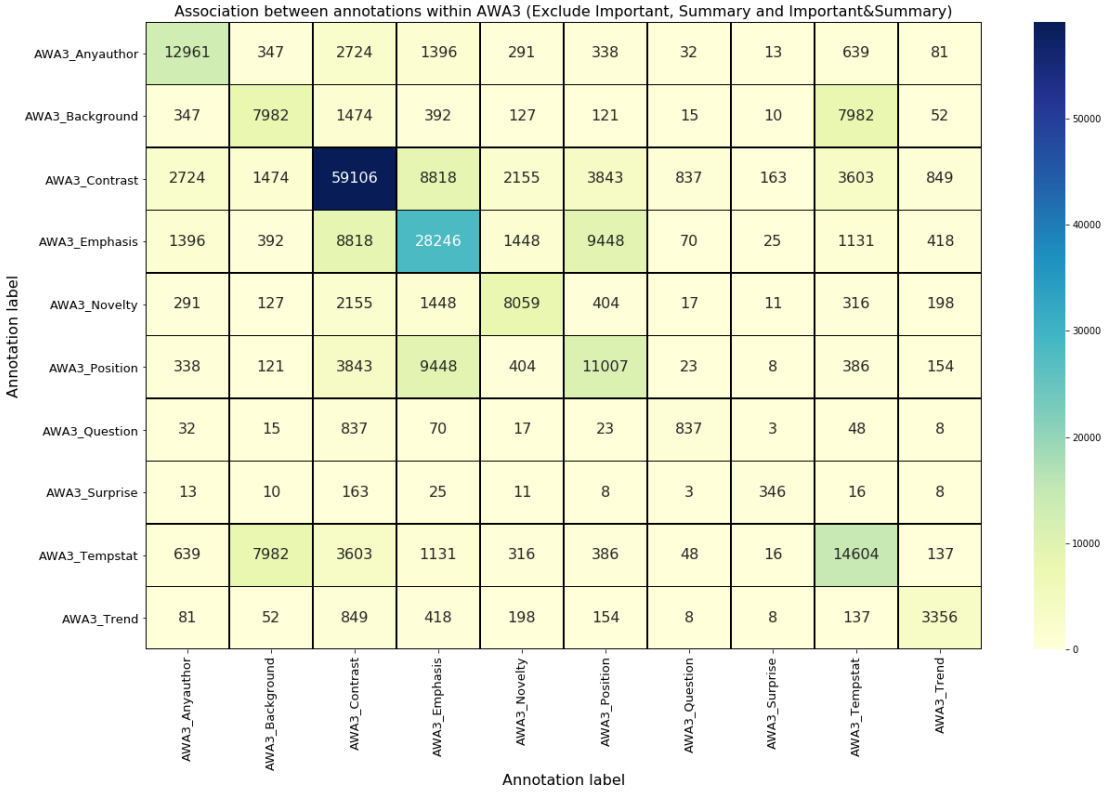

Association between annotations within AWA
Annotation label
Background 2059.0 0.0 247.0 2037.0 22.0 23.0 76.0 126.0 0.0 11.0 1.0
Contrast 0.0 59443.0 3888.0 50882.0 8561.0 837.0 797.0 1114.0 0.0 323.0 542.0
Emphasis 247.0 3888.0 14998.0 14741.0 257.0 512.0 3720.0 329.0 0.0 36.0 312.0
Important 2037.0 50882.0 14741.0 73520.0 0.0 3596.0 4632.0 3391.0 0.0 691.0 3105.0
Important&Summary 22.0 8561.0 257.0 0.0 9640.0 395.0 15.0 136.0 0.0 75.0 272.0
Novelty 23.0 837.0 512.0 3596.0 395.0 3991.0 140.0 70.0 0.0 17.0 166.0
Position 76.0 797.0 3720.0 4632.0 15.0 140.0 4647.0 63.0 0.0 98.0 50.0
Question 126.0 1114.0 329.0 3391.0 136.0 70.0 63.0 3527.0 0.0 12.0 52.0
Summary 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 25831.0 15.0 0.0
Surprise 11.0 323.0 36.0 691.0 75.0 17.0 98.0 12.0 15.0 781.0 21.0
Trend 1.0 542.0 312.0 3105.0 272.0 166.0 50.0 52.0 0.0 21.0 3377.0
Background Contrast Emphasis Important Important&Summary Novelty Position Question Summary Surprise Trend
Annotation label
60000
45000
30000
15000
0



9.0 Association within AWA3

In AWA3, there are two additional annotations called 'AWA3_Anyauthor' and 'AWA3_Tempstat' being considered. AWA3 also annotates each sentence with possible more than one annotations. The figures below show the association between the annotations within AWA3. One significant different from AWA is that the 'Position annotation in AWA3 does have strong association with 'Contrast' annotation too.


Association between annotations within AWA3
Annotation label
AWA3_Anyauthor 12961.0 347.0 2724.0 1396.0 0.0 2724.0 291.0 338.0 32.0 10237.0 13.0 639.0 81.0
AWA3_Background 347.0 7982.0 1474.0 392.0 1234.0 240.0 127.0 121.0 15.0 983.0 10.0 7982.0 52.0
AWA3_Contrast 2724.0 1474.0 59106.0 8818.0 51666.0 7440.0 2155.0 3843.0 837.0 0.0 163.0 3603.0 849.0
AWA3_Emphasis 1396.0 392.0 8818.0 28246.0 7328.0 1490.0 1448.0 9448.0 70.0 2585.0 25.0 1131.0 418.0
AWA3_Important 0.0 1234.0 51666.0 7328.0 51666.0 0.0 1921.0 3380.0 774.0 0.0 150.0 3158.0 781.0
AWA3_Important&Summary 2724.0 240.0 7440.0 1490.0 0.0 7440.0 234.0 463.0 63.0 0.0 13.0 445.0 68.0
AWA3_Novelty 291.0 127.0 2155.0 1448.0 1921.0 234.0 8059.0 404.0 17.0 551.0 11.0 316.0 198.0
AWA3_Position 338.0 121.0 3843.0 9448.0 3380.0 463.0 404.0 11007.0 23.0 690.0 8.0 386.0 154.0
AWA3_Question 32.0 15.0 837.0 70.0 774.0 63.0 17.0 23.0 837.0 0.0 3.0 48.0 8.0
AWA3_Summary 10237.0 983.0 0.0 2585.0 0.0 0.0 551.0 690.0 0.0 36941.0 19.0 1436.0 230.0
AWA3_Surprise 13.0 10.0 163.0 25.0 150.0 13.0 11.0 8.0 3.0 19.0 346.0 16.0 8.0
AWA3_Tempstat 639.0 7982.0 3603.0 1131.0 3158.0 445.0 316.0 386.0 48.0 1436.0 16.0 14604.0 137.0
AWA3_Trend 81.0 52.0 849.0 418.0 781.0 68.0 198.0 154.0 8.0 230.0 8.0 137.0 3356.0
AWA3_Anyauthor AWA3_Background AWA3_Contrast AWA3_Emphasis AWA3_Important AWA3_Important&Summary AWA3_Novelty AWA3_Position AWA3_Question AWA3_Summary AWA3_Surprise AWA3_Tempstat AWA3_Trend
Annotation label
50000
40000
30000
20000
10000
0

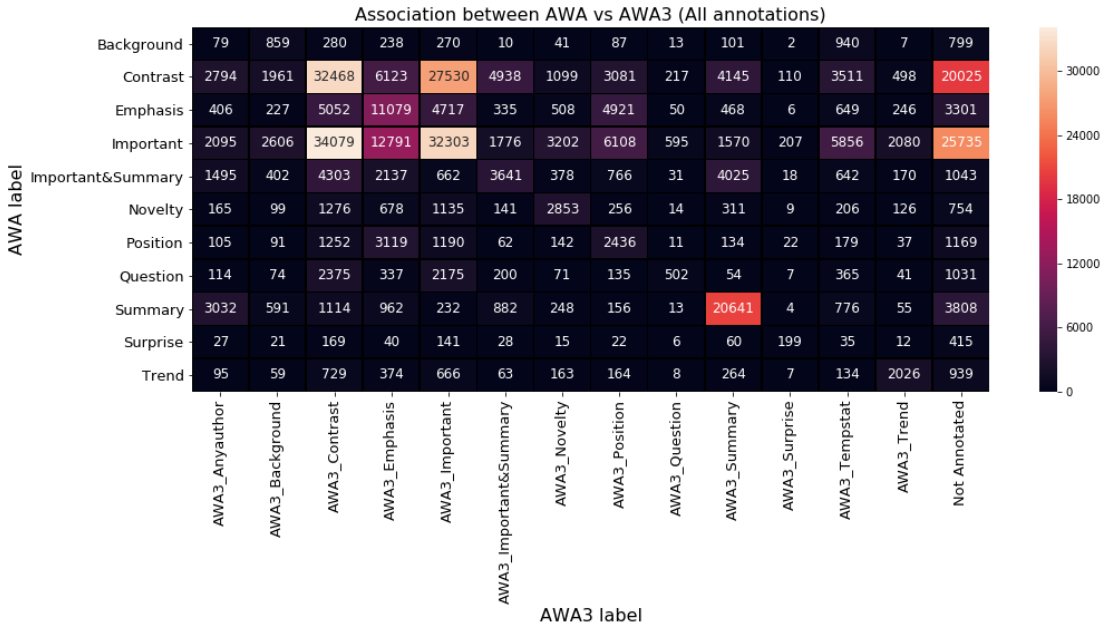

Association between annotations within AWA3 (Exclude Important, Summary and Important&Summary)
Annotation label
AWA3_Anyauthor 12961 347 2724 1396 291 338 32 13 639 81
AWA3_Background 347 7982 1474 392 127 121 15 10 7982 52
AWA3_Contrast 2724 1474 59106 8818 2155 3843 837 163 3603 849
AWA3_Emphasis 1396 392 8818 28246 1448 9448 70 25 1131 418
AWA3_Novelty 291 127 2155 1448 8059 404 17 11 316 198
AWA3_Position 338 121 3843 9448 404 11007 23 8 386 154
AWA3_Question 32 15 837 70 17 23 837 3 48 8
AWA3_Surprise 13 10 163 25 11 8 3 346 16 8
AWA3_Tempstat 639 7982 3603 1131 316 386 48 16 14604 137
AWA3_Trend 81 52 849 418 198 154 8 8 137 3356
AWA3_Anyauthor AWA3_Background AWA3_Contrast AWA3_Emphasis AWA3_Novelty AWA3_Position AWA3_Question AWA3_Surprise AWA3_Tempstat AWA3_Trend
Annotation label
50000
40000
30000
20000
10000
0

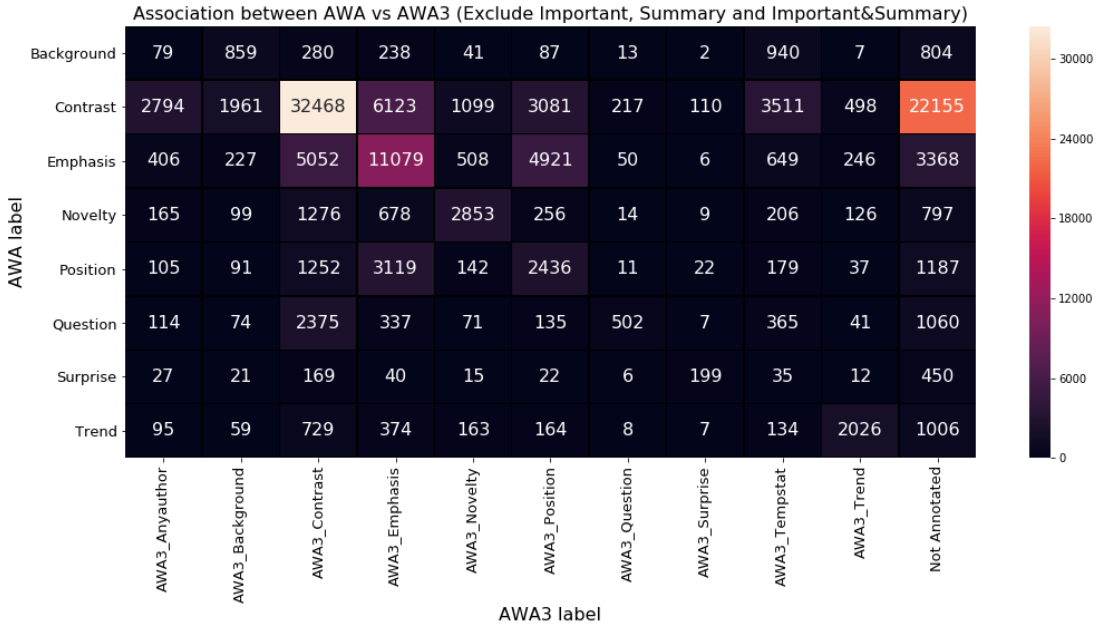
10.0 AWA and AWA3 Associations

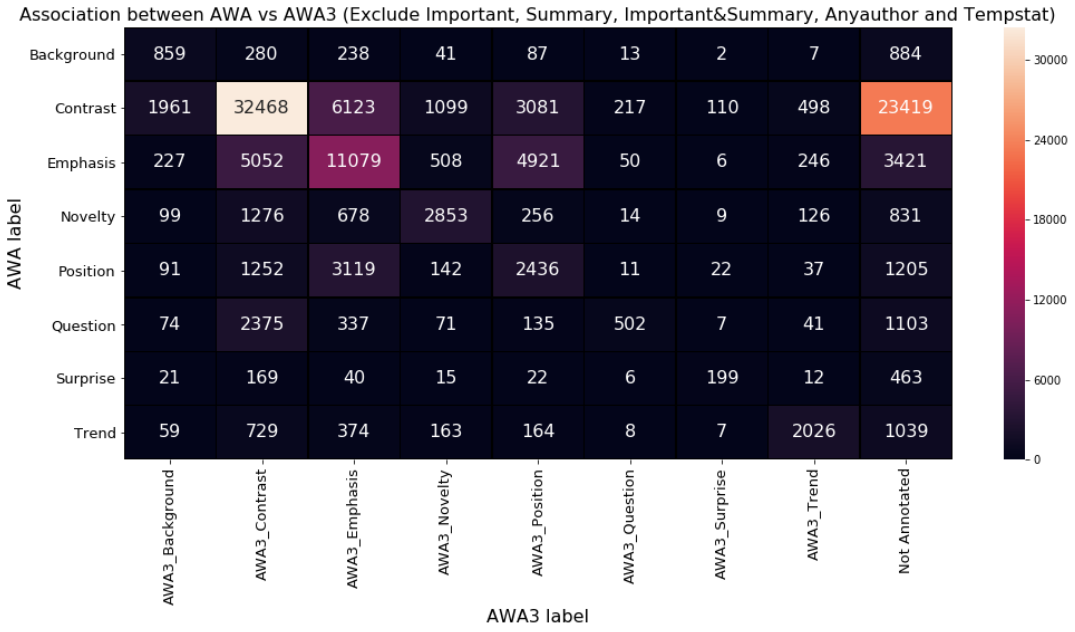
AWA3 is somewhat supposed to be an improved version of AWA. So, it will be important to compare the association of these two tools.

10.1 AWA vs AWA3

Here, we will look at the association between AWA vs AWA3 in three different ways. One is association across all annotations, second is excluding the association 'Important', 'Summary' and 'Important&Summary' and lastly is the association excluding the 'Important', 'Summary', 'Important&Summary', 'AWA3_Anyuthor' and 'AWA3_Tempstat'.



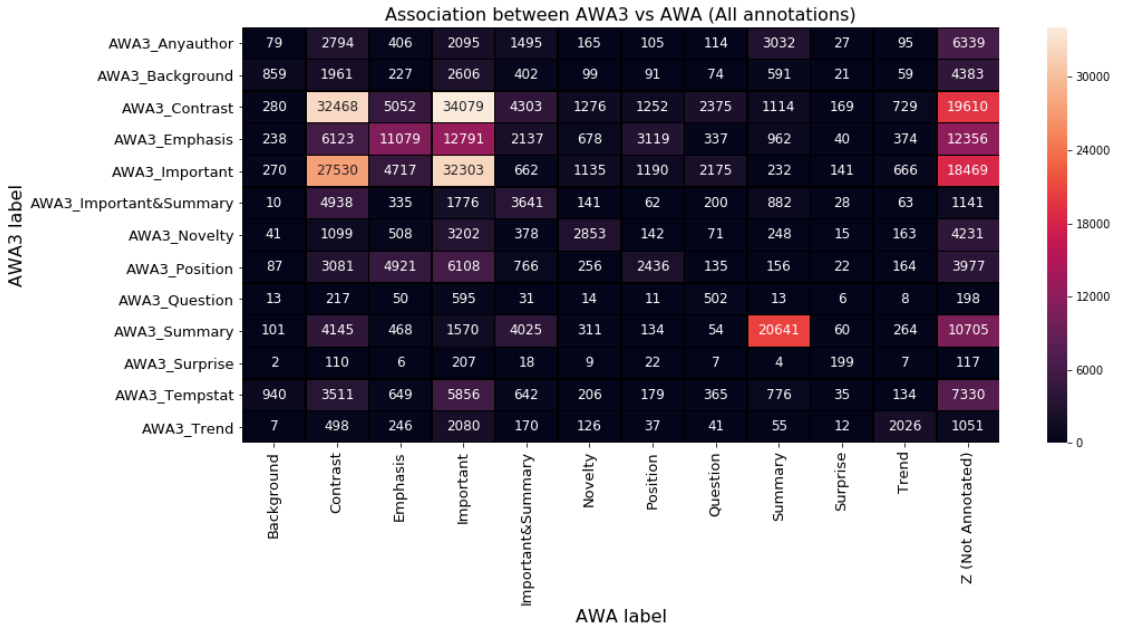


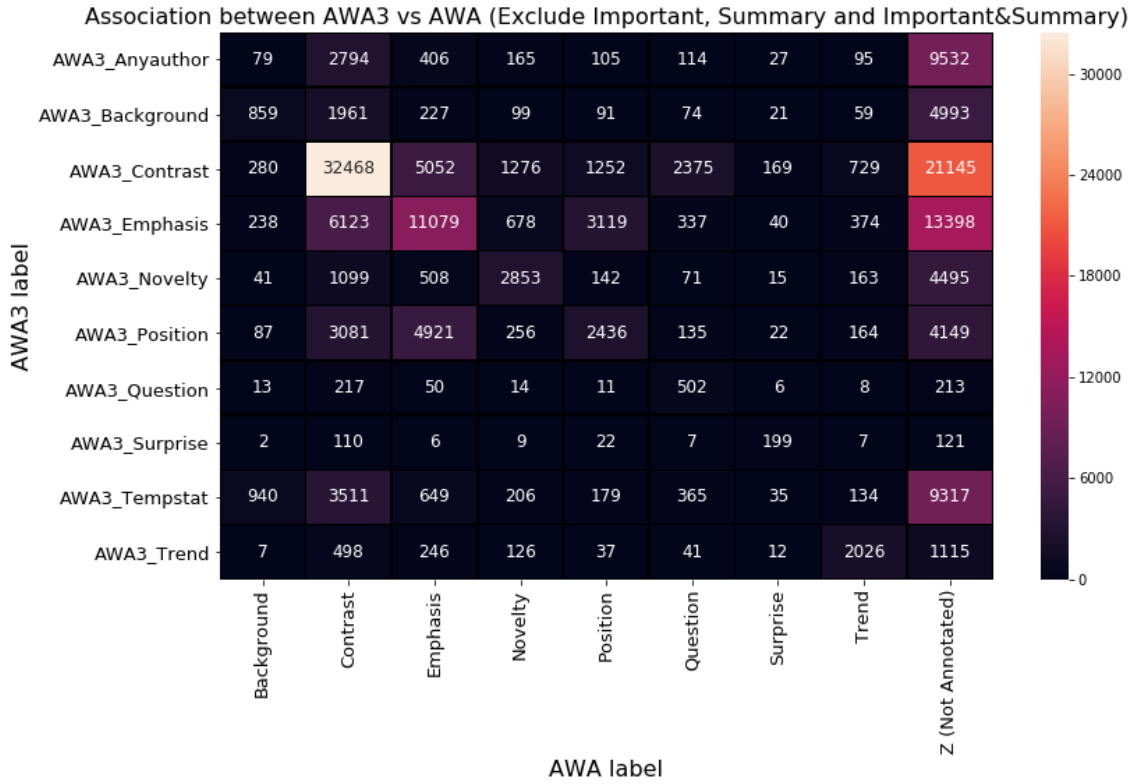


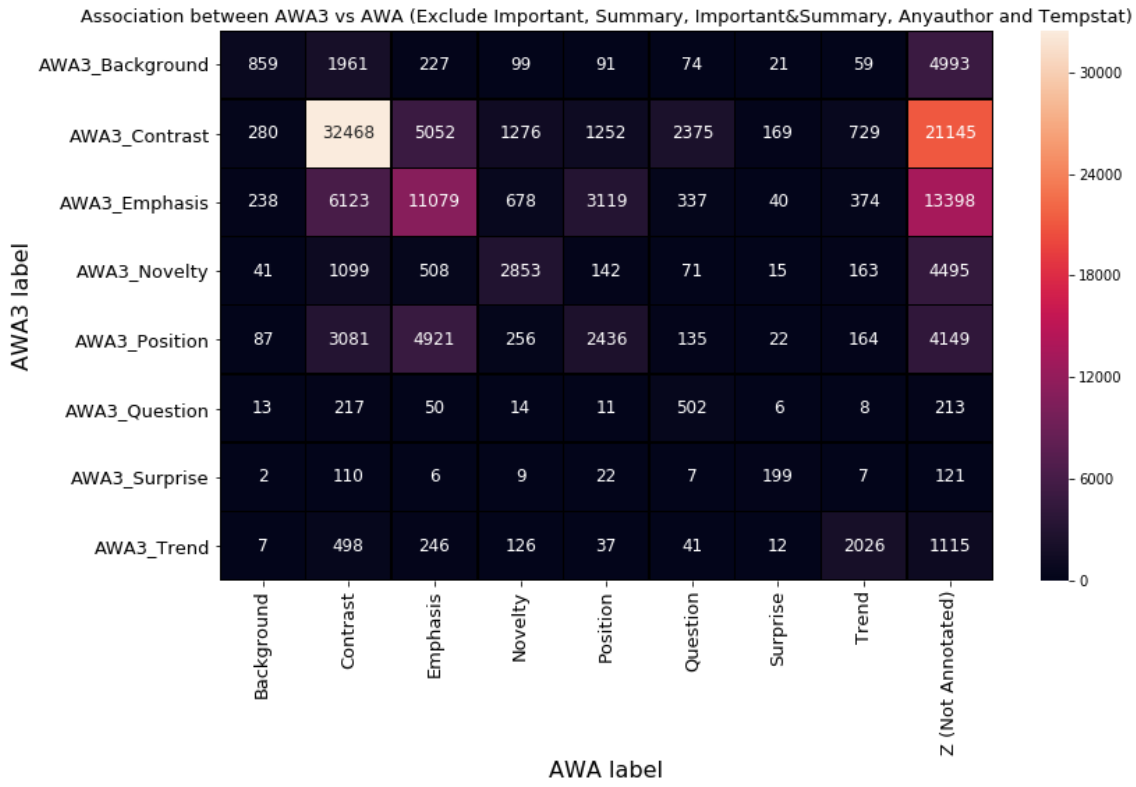
From the above diagrams, there are indeed some changes been made to AWA to become AWA3. As an example (diagram 3), there are many sentences (23419 sentences) that are annotated as 'Contrast' in AWA is no longer being annotated in AWA3. Apart from that, sentences annotated as 'Question' in AWA has no strong association with 'AWA3_Question' but strong association with 'AWA3_Contrast' as well as 1103 sentences in this annotation not being annotated. Apart from that, some sentences annotated with 'Contrast' in AWA have been shifted to be annotated as 'Emphasis' in AWA3. Meanwhile, some sentences annotated with 'Emphasis in AWA have been shifted to be annotated as 'Contrast' in AWA3.

10.2 AWA3 vs AWA

Similar as 10.1 but this time is AWA3 vs AWA. In this way, we can see how many sentences that are annotated with AWA3 but not annotated with AWA.


Association between AWA3 vs AWA (All annotations)
AWA3_Anyauthor 79 2794 406 2095 1495 165 105 114 3032 27 95 6339
AWA3_Background 859 1961 227 2606 402 99 91 74 591 21 59 4383
AWA3_Contrast 280 32468 5052 34079 4303 1276 1252 2375 1114 169 729 19610
AWA3_Emphasis 238 6123 11079 12791 2137 678 3119 337 962 40 374 12356
AWA3_Important 270 27530 4717 32303 662 1135 1190 2175 232 141 666 18469
AWA3_Important&Summary 10 4938 335 1776 3641 141 62 200 882 28 63 1141
AWA3_Novelty 41 1099 508 3202 378 2853 142 71 248 15 163 4231
AWA3_Position 87 3081 4921 6108 766 256 2436 135 156 22 164 3977
AWA3_Question 13 217 50 595 31 14 11 502 13 6 8 198
AWA3_Summary 101 4145 468 1570 4025 311 134 54 20641 60 264 10705
AWA3_Surprise 2 110 6 207 18 9 22 7 4 199 7 117
AWA3_Tempstat 940 3511 649 5856 642 206 179 365 776 35 134 7330
AWA3_Trend 7 498 246 2080 170 126 37 41 55 12 2026 1051
AWA3 label
Background Contrast Emphasis Important Important&Summary Novelty Position Question Summary Surprise Trend Z (Not Annotated)
AWA label
30000 24000 18000 12000 6000 0

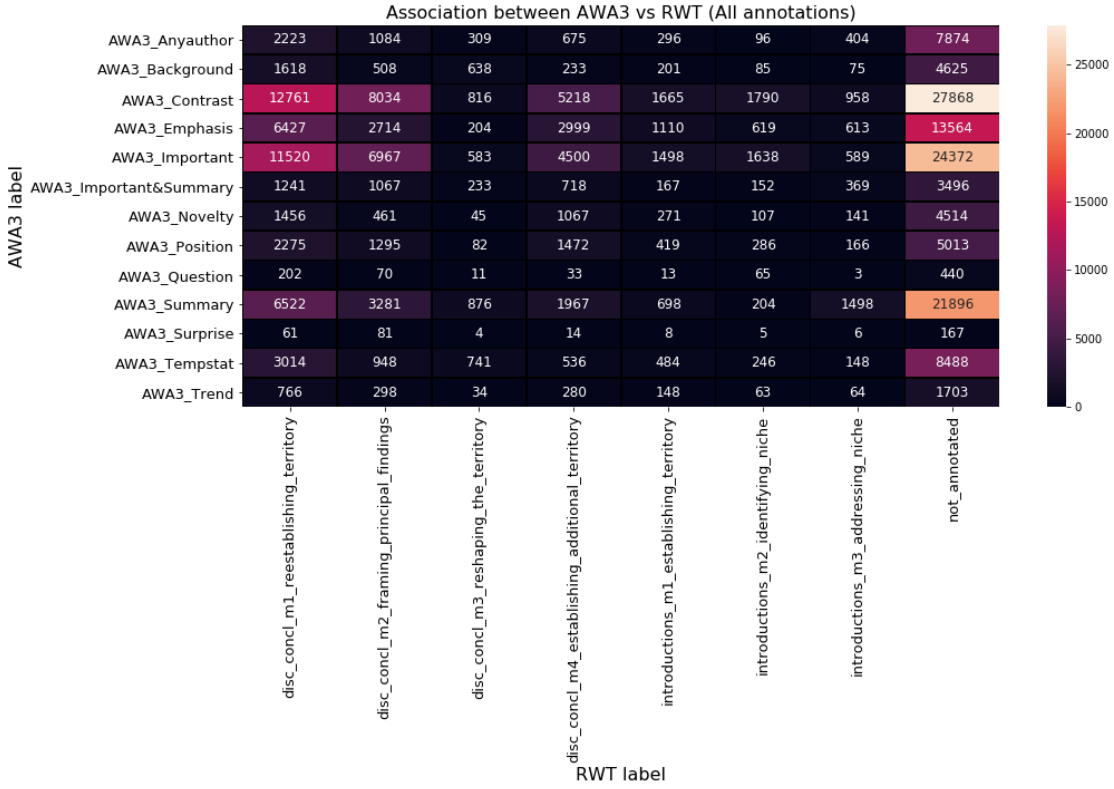

Association between AWA3 vs AWA (Exclude Important, Summary and Important&Summary)
AWA3_Anyauthor 79 2794 406 165 105 114 27 95 9532
AWA3_Background 859 1961 227 99 91 74 21 59 4993
AWA3_Contrast 280 32468 5052 1276 1252 2375 169 729 21145
AWA3_Emphasis 238 6123 11079 678 3119 337 40 374 13398
AWA3_Novelty 41 1099 508 2853 142 71 15 163 4495
AWA3_Position 87 3081 4921 256 2436 135 22 164 4149
AWA3_Question 13 217 50 14 11 502 6 8 213
AWA3_Surprise 2 110 6 9 22 7 199 7 121
AWA3_Tempstat 940 3511 649 206 179 365 35 134 9317
AWA3_Trend 7 498 246 126 37 41 12 2026 1115
AWA3 label
Background Contrast Emphasis Novelty Position Question Surprise Trend Z (Not Annotated)
AWA label
30000 24000 18000 12000 6000 0

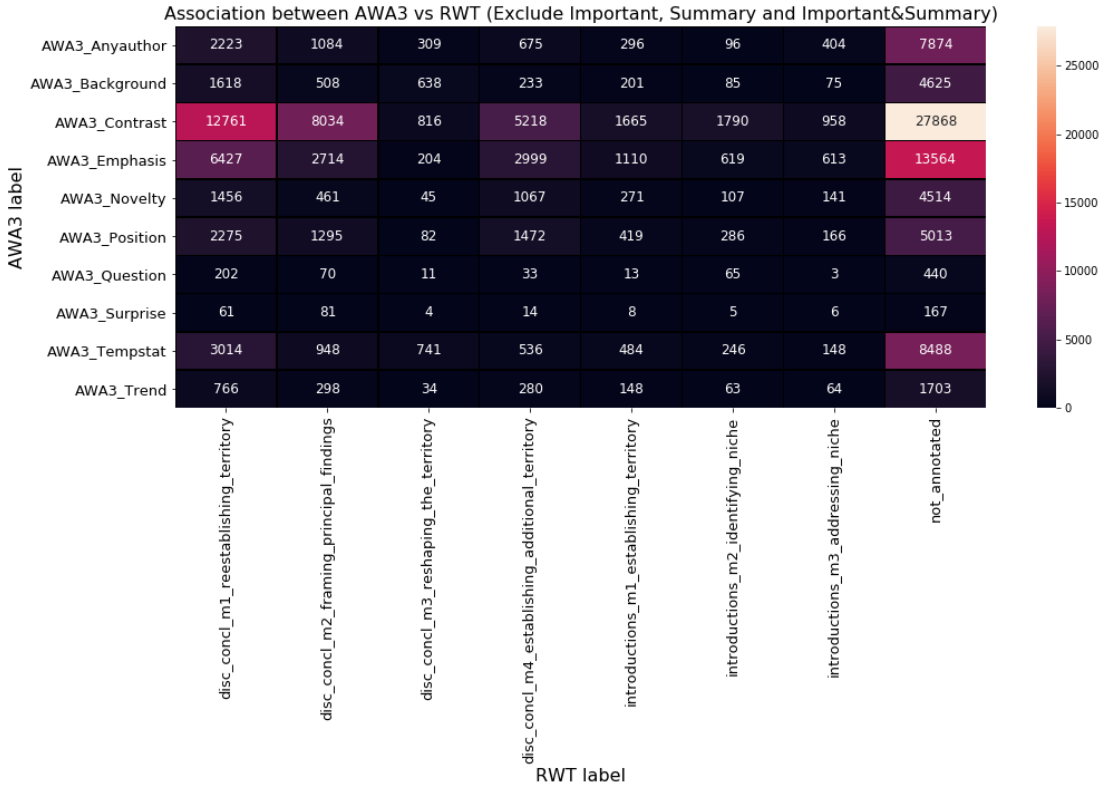


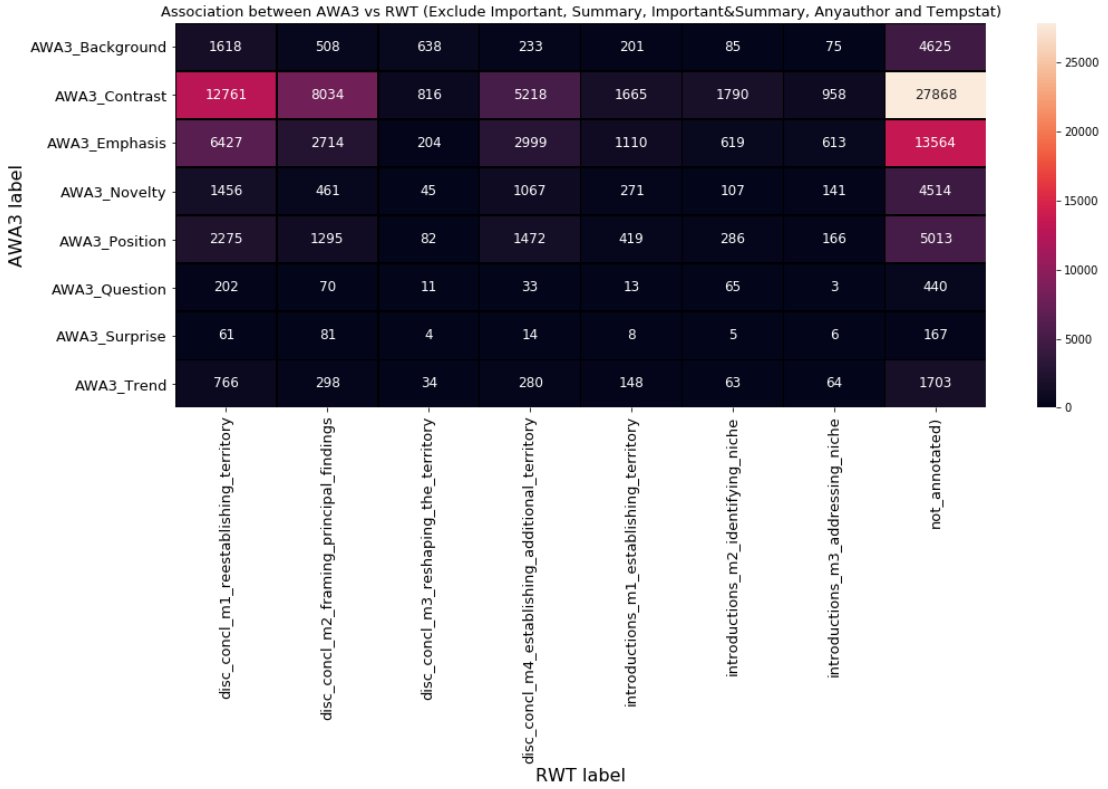
From the above diagrams, the obvious different between AWA3 vs AWA is that many sentences annotated by AWA3 are not annotated with any AWA annotation. That can be noticed in the last column of each of the diagrams.

11. AWA3 and RWT associations

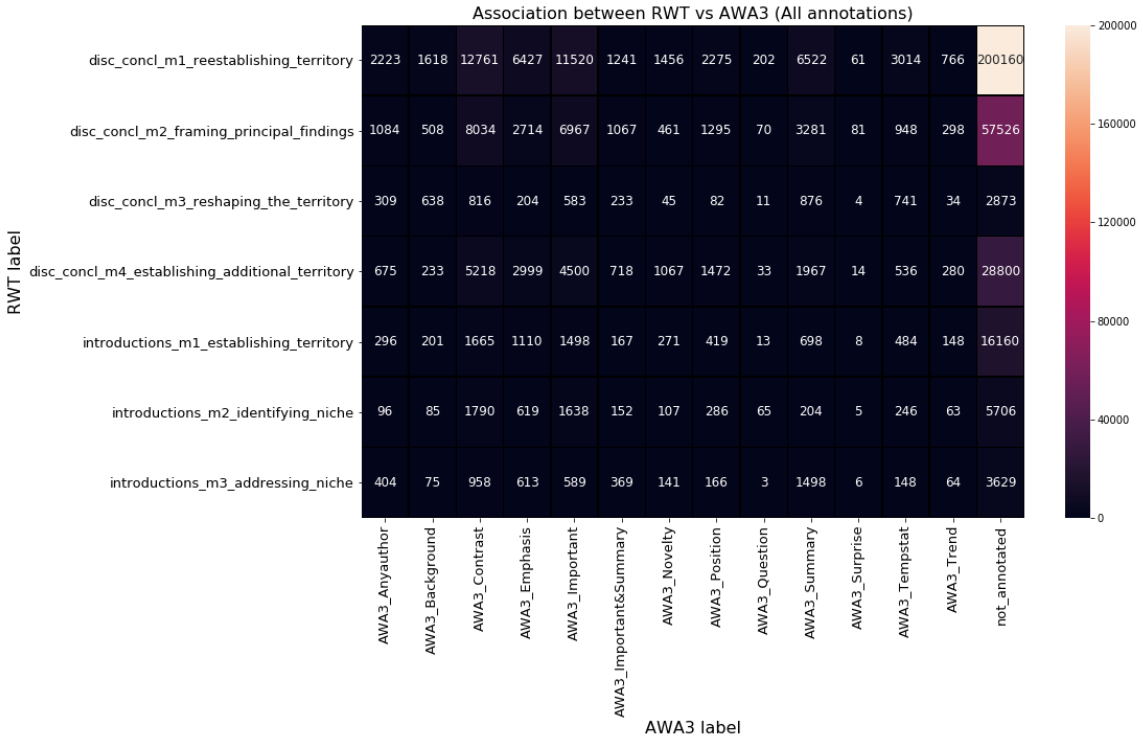
11.1 AWA3 VS RWT

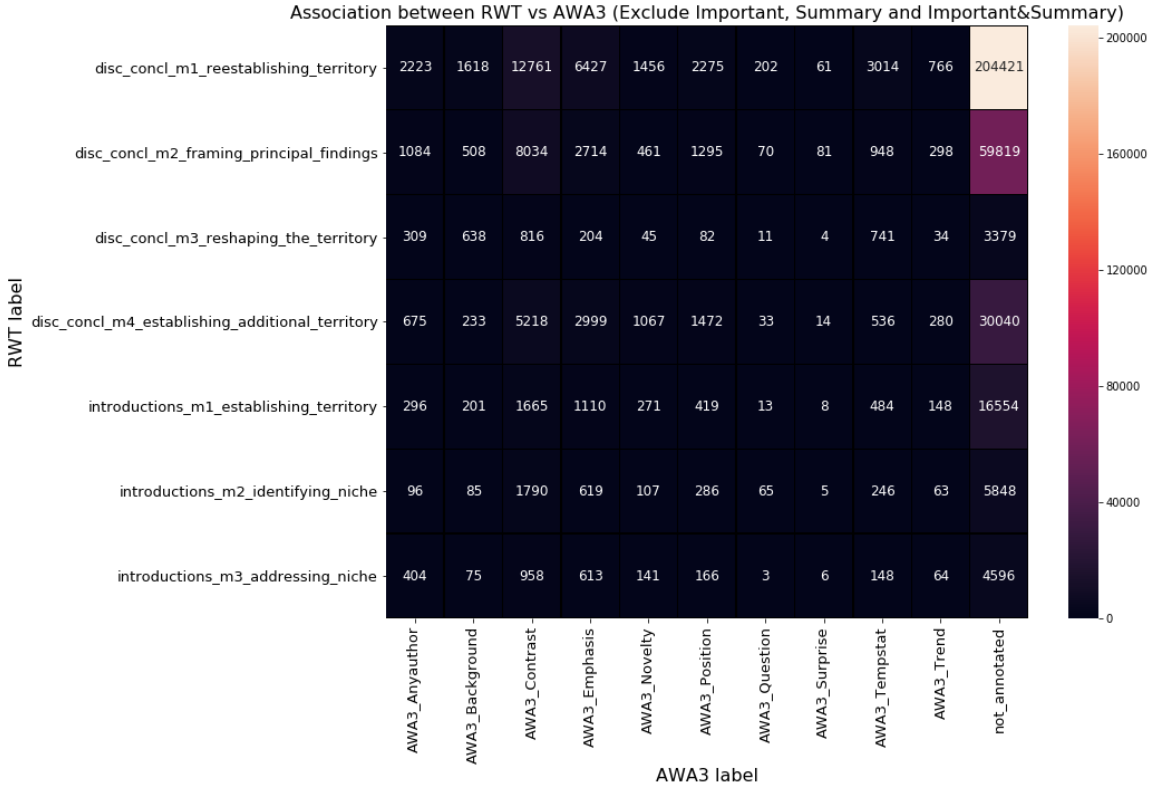


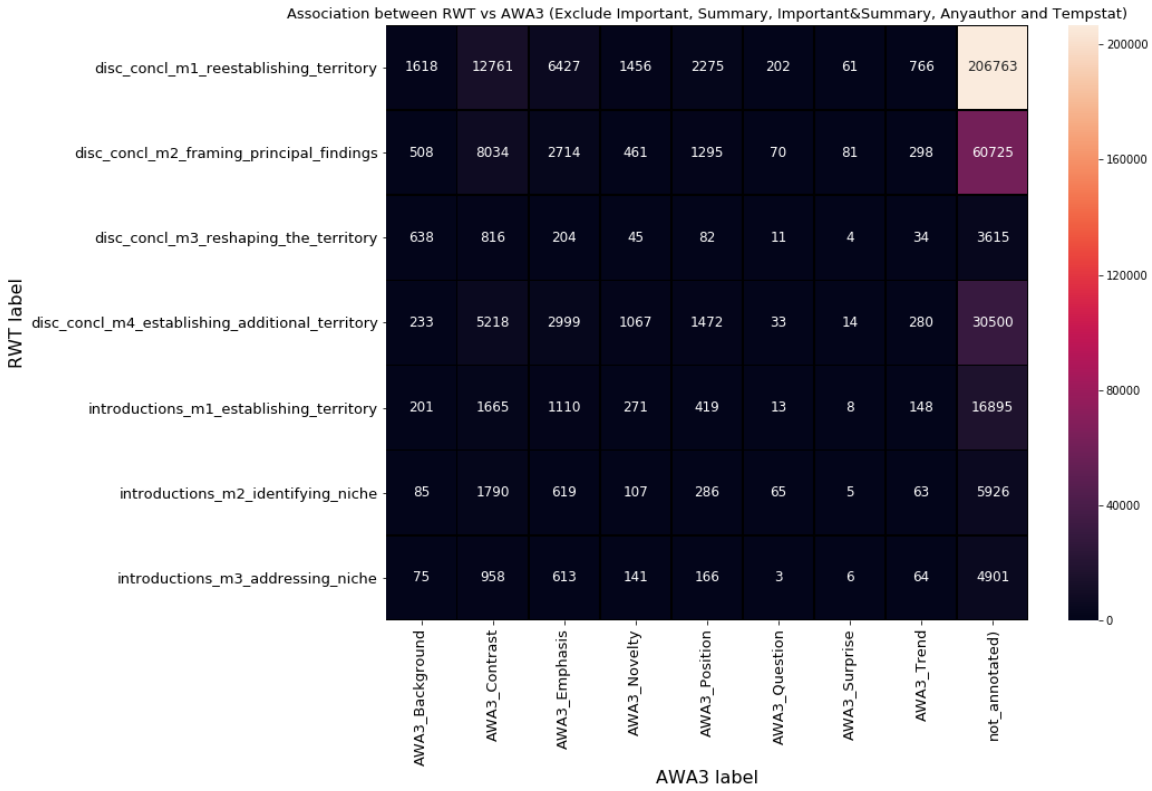




11.2 RWT VS AWA3

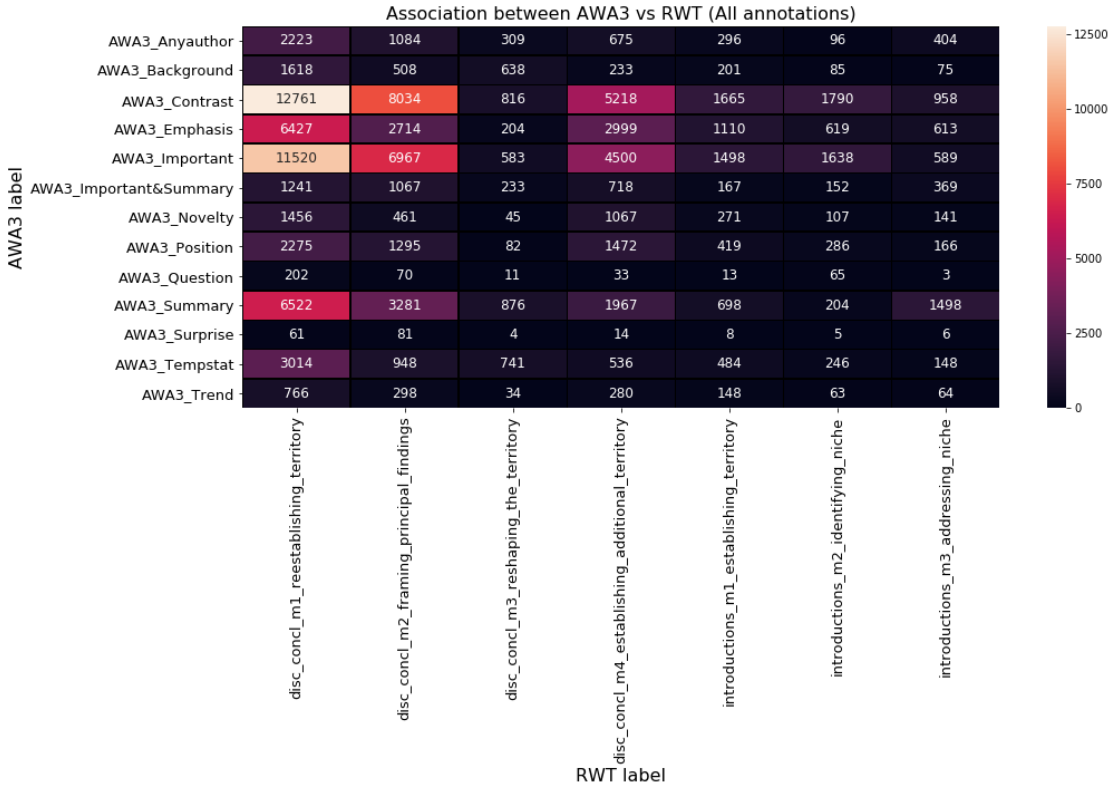


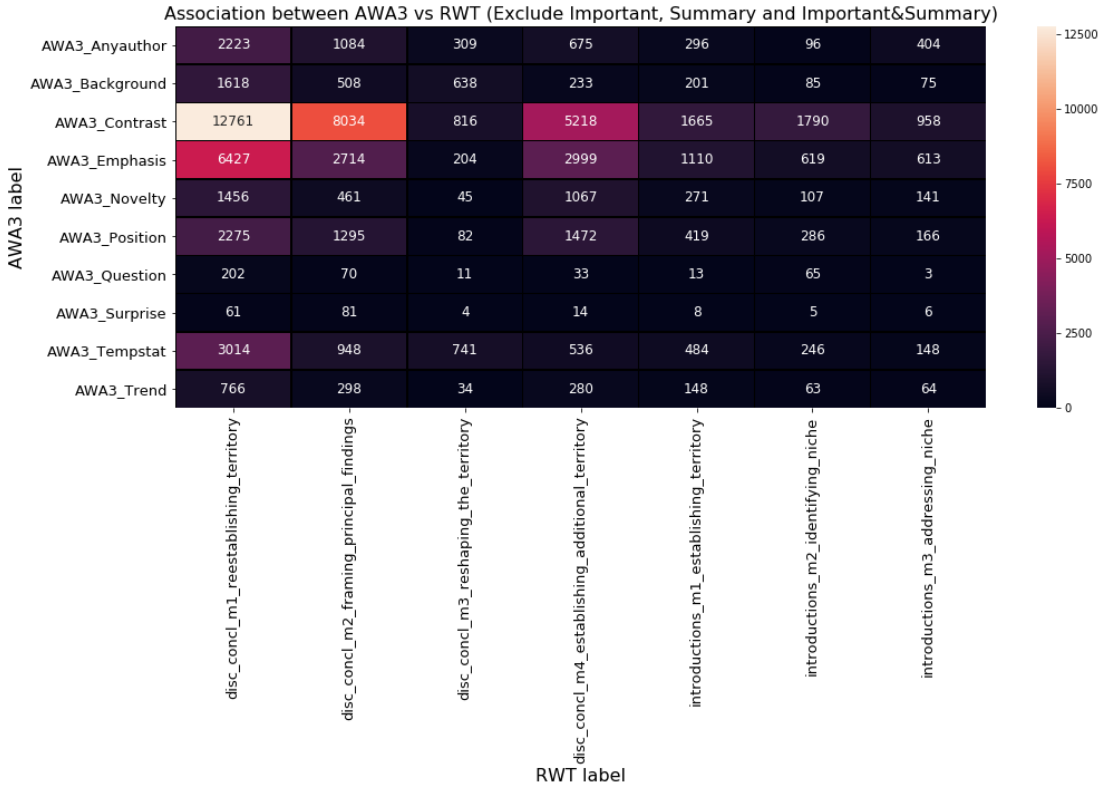

Association between RWT vs AWA3 (Exclude Important, Summary and Important&Summary)
RWT label
disc_concl_m1_reestablishing_territory
disc_concl_m2_framing_principal_findings
disc_concl_m3_reshaping_the_territory
disc_concl_m4_establishing_additional_territory
introductions_m1_establishing_territory
introductions_m2_identifying_niche
introductions_m3_addressing_niche
2223 1618 12761 6427 1456 2275 202 61 3014 766 204421
1084 508 8034 2714 461 1295 70 81 948 298 59819
309 638 816 204 45 82 11 4 741 34 3379
675 233 5218 2999 1067 1472 33 14 536 280 30040
296 201 1665 1110 271 419 13 8 484 148 16554
96 85 1790 619 107 286 65 5 246 63 5848
404 75 958 613 141 166 3 6 148 64 4596
AWA3_Anyauthor
AWA3_Background
AWA3_Contrast
AWA3_Emphasis
AWA3_Novelty
AWA3_Position
AWA3_Question
AWA3_Surprise
AWA3_Tempstat
AWA3_Trend
not_annotated
AWA3 label
200000
160000
120000
80000
40000
0

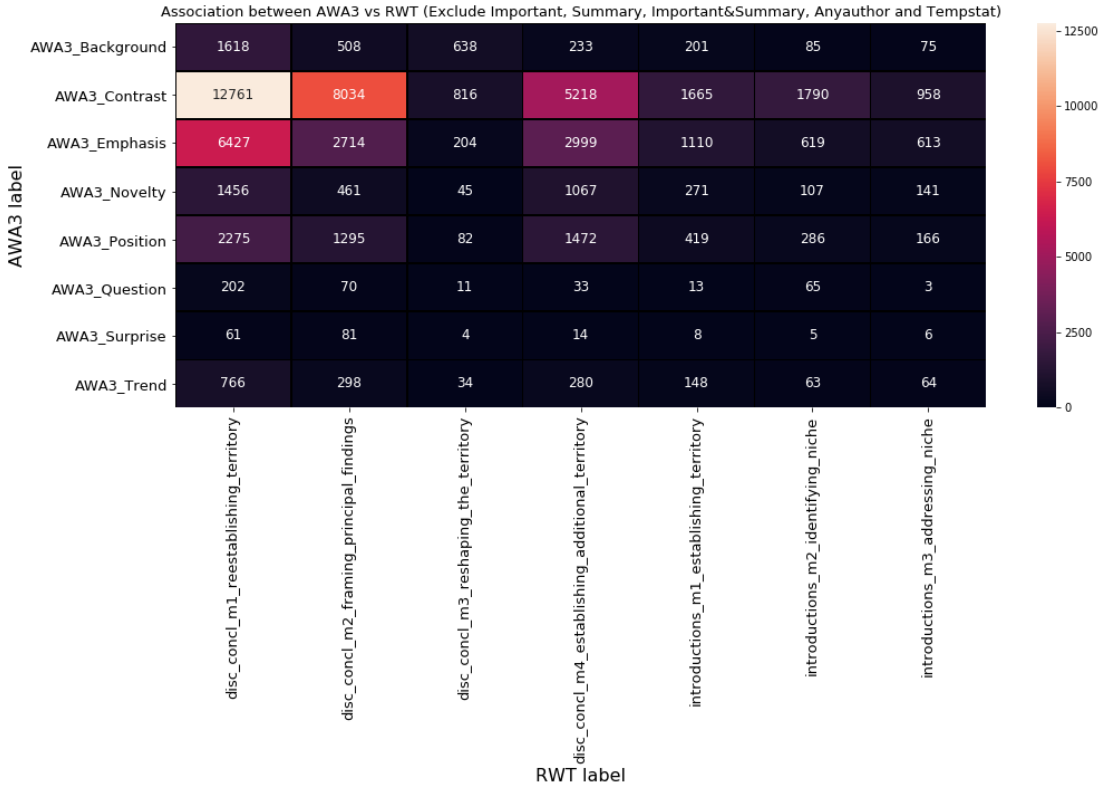

Association between RWT vs AWA3 (Exclude Important, Summary, Important&Summary, Anyauthor and Tempstat)
RWT label
disc_concl_m1_reestablishing_territory
disc_concl_m2_framing_principal_findings
disc_concl_m3_reshaping_the_territory
disc_concl_m4_establishing_additional_territory
introductions_m1_establishing_territory
introductions_m2_identifying_niche
introductions_m3_addressing_niche
1618 12761 6427 1456 2275 202 61 766 206763
508 8034 2714 461 1295 70 81 298 60725
638 816 204 45 82 11 4 34 3615
233 5218 2999 1067 1472 33 14 280 30500
201 1665 1110 271 419 13 8 148 16895
85 1790 619 107 286 65 5 63 5926
75 958 613 141 166 3 6 64 4901
AWA3_Background
AWA3_Contrast
AWA3_Emphasis
AWA3_Novelty
AWA3_Position
AWA3_Question
AWA3_Surprise
AWA3_Trend
not_annotated)
AWA3 label
200000
160000
120000
80000
40000
0

12. AWA3 and RWT associations (Ignoring sentences that not annotated)

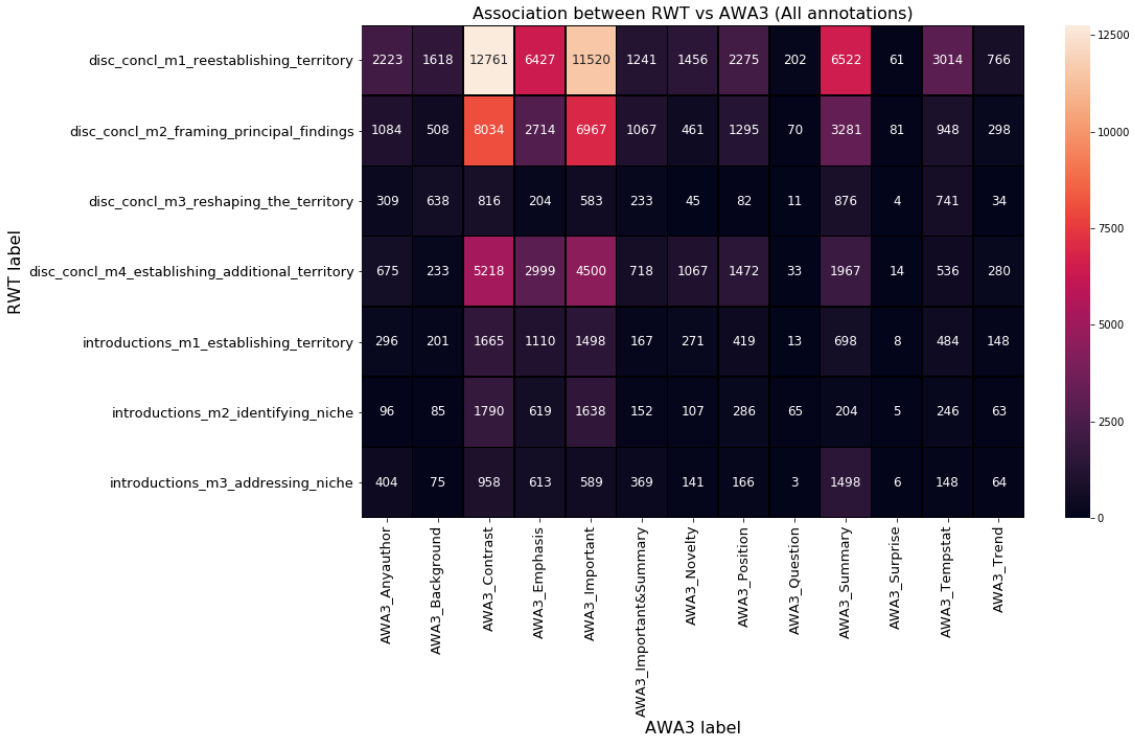
12.1 AWA3 VS RWT

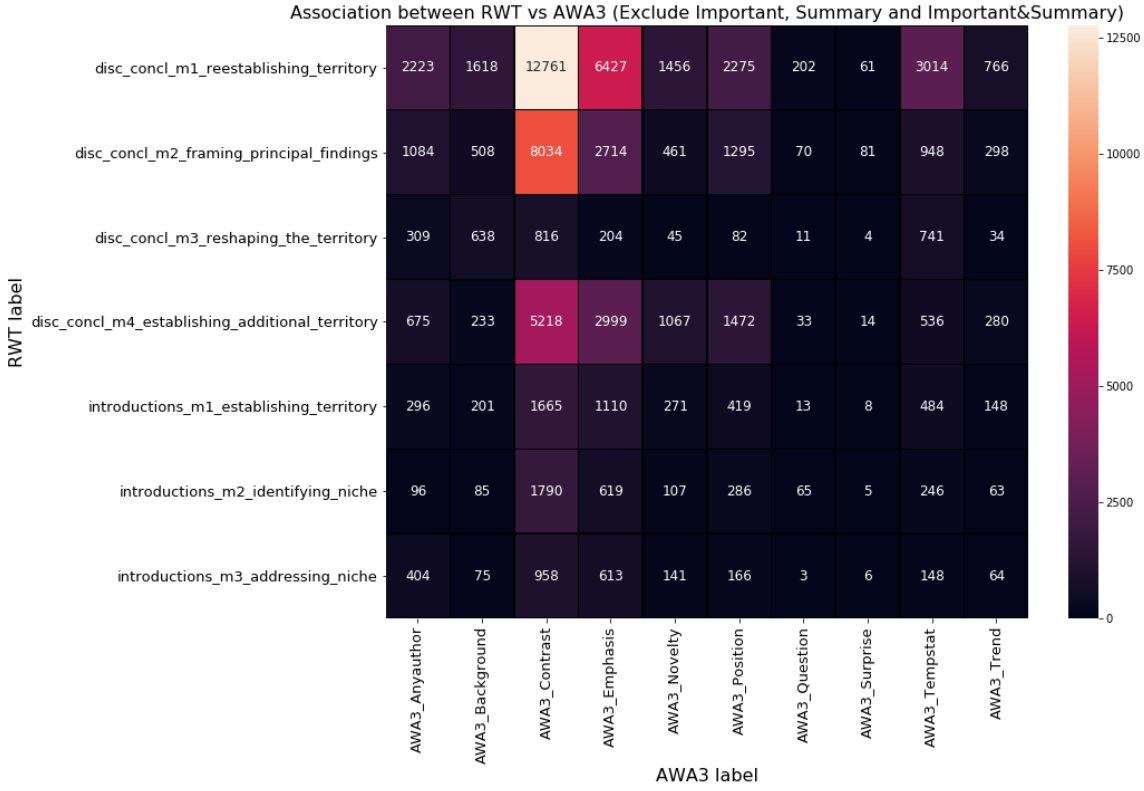


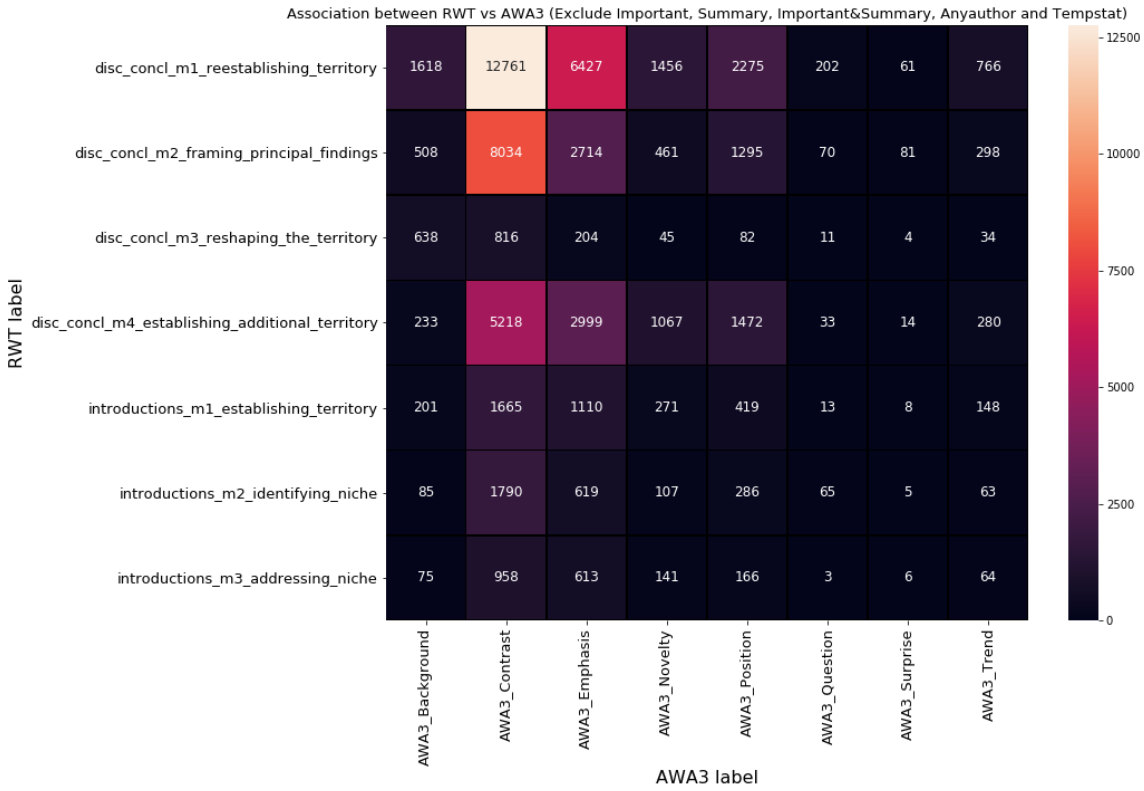




12.2 RWT VS AWA3





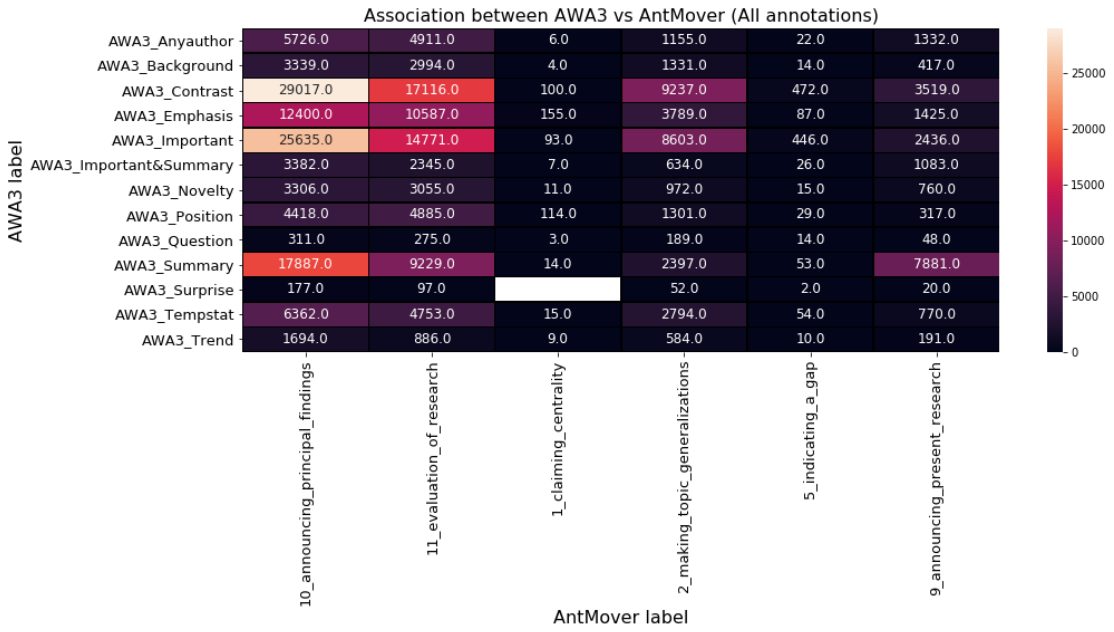


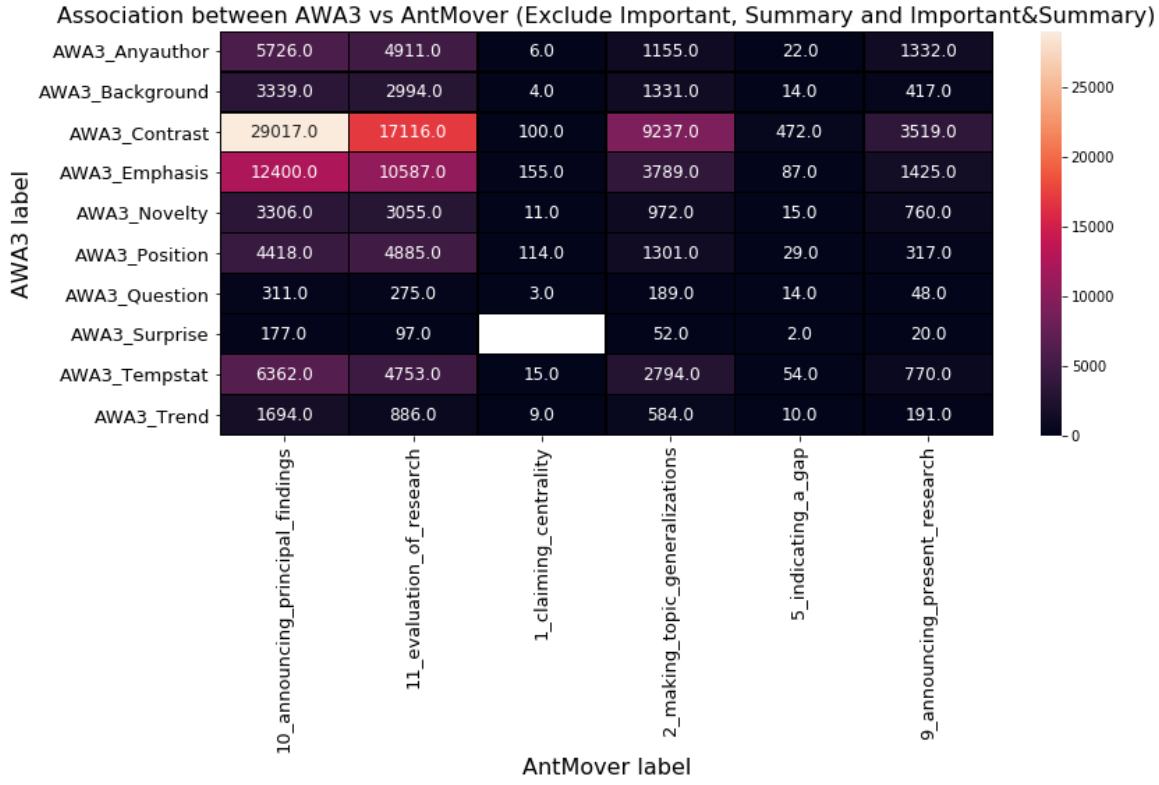
From the section 11 to 12, it is clearly to see that there is no direct association between AWA3 and RWT annotations. The only things can be derived from the above diagrams are some annotations such as 'disc_concl_m1_reestablishing_territory' and

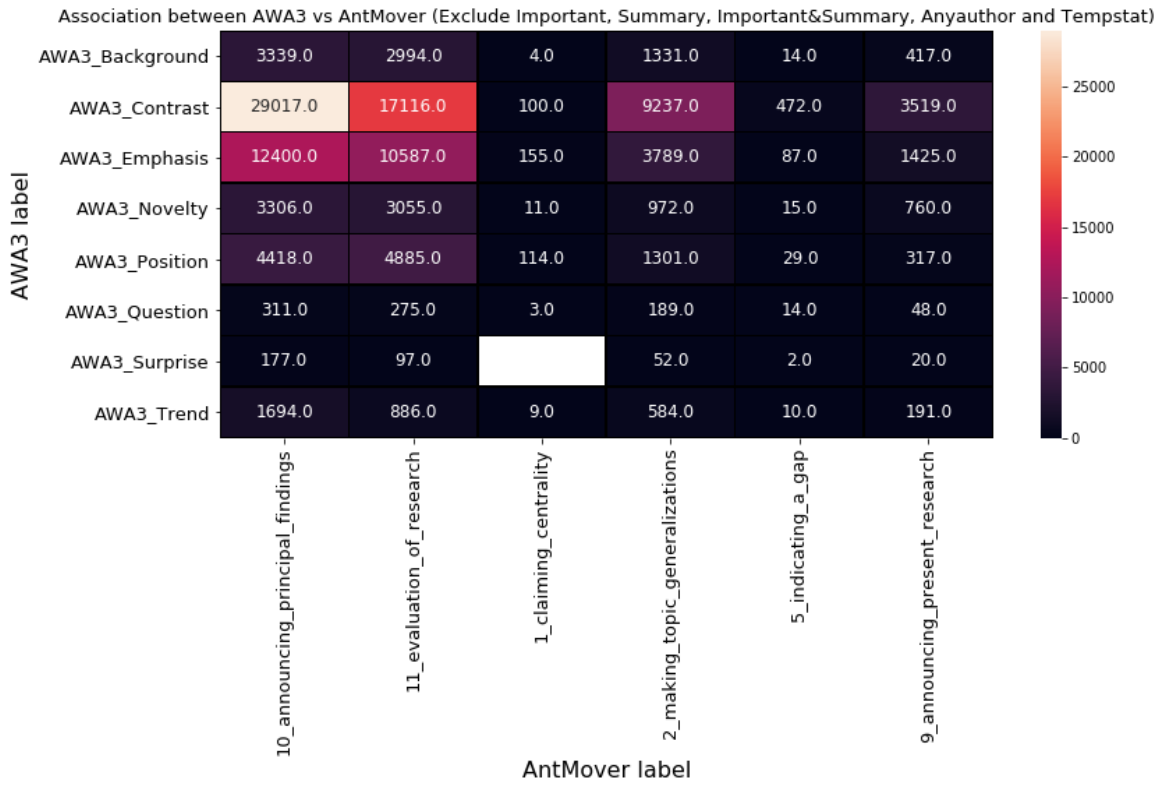
'disc_concl_m2_framing_principle_finding' are having high probability to be selected whenever that sentences are annotated as 'AWA3_Contrast' and 'AWA3_Emphasis' by AWA3.

13. AWA3 and AntMover associations

13.1 AWA3 VS AntMover



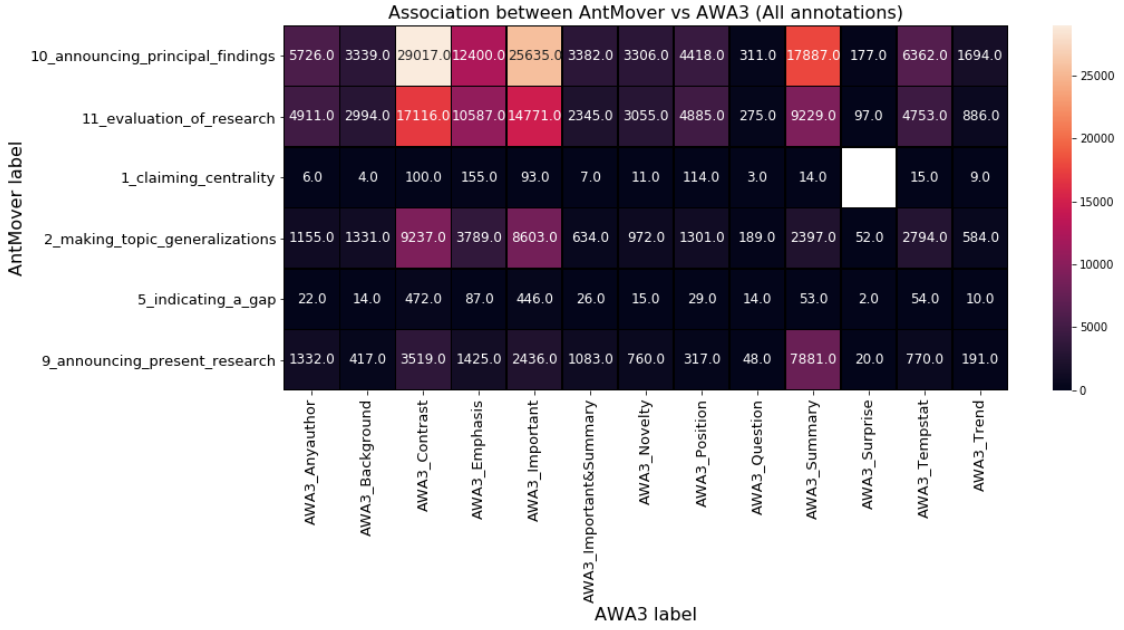


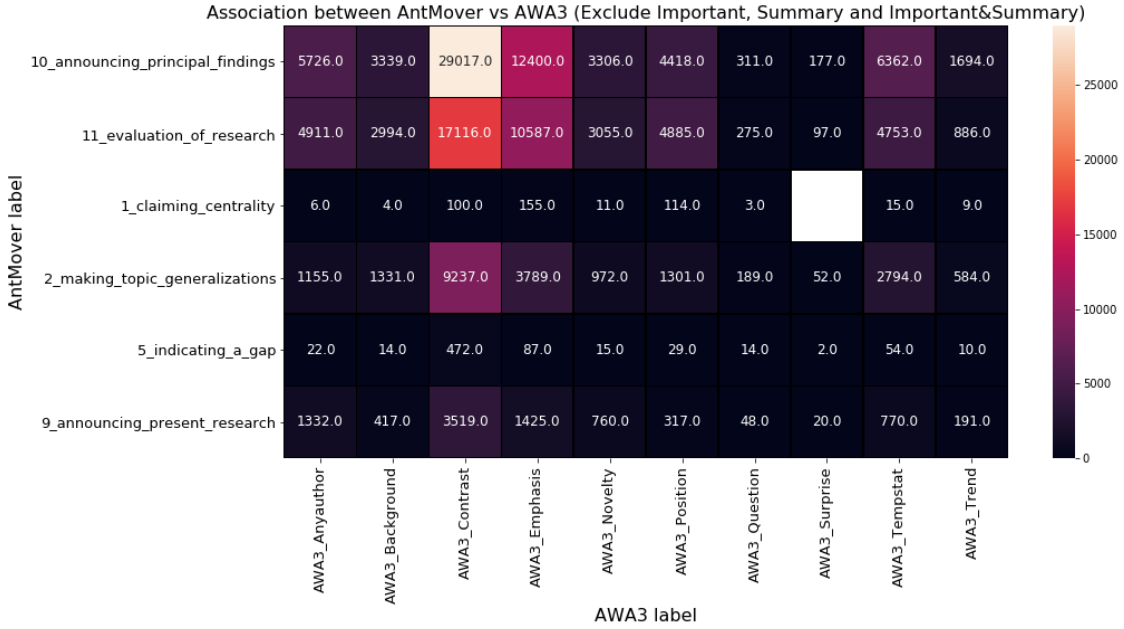


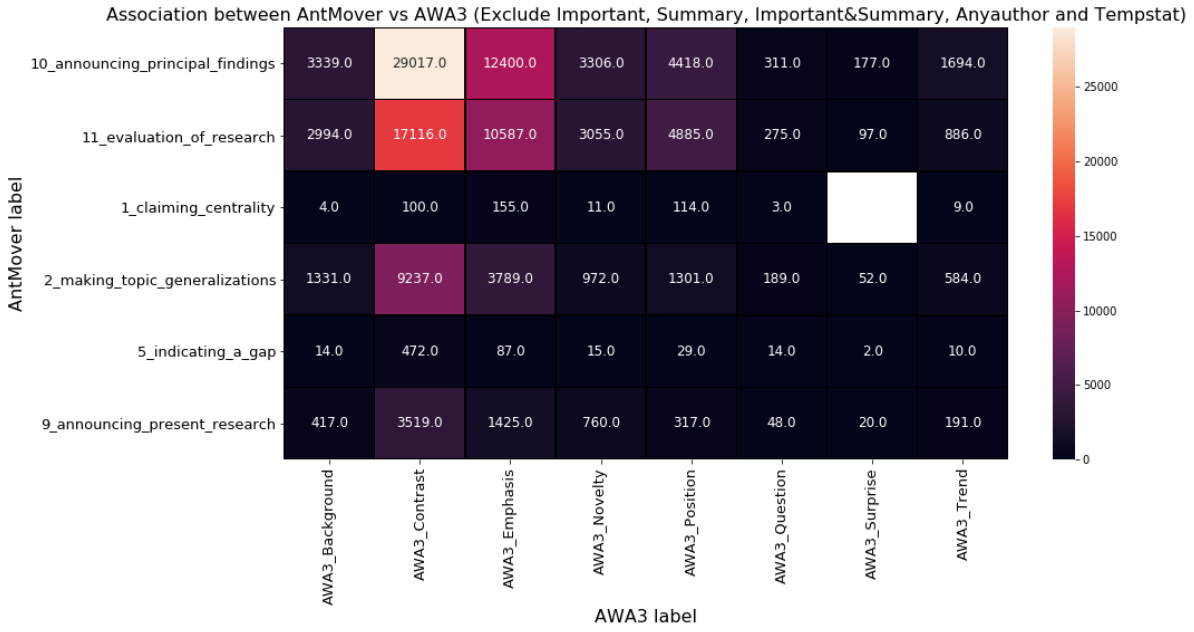
From the above three diagrams, we can see that there is no obvious association between AWA3 and AntMover annotations. However, there is one interesting thing to be noticed. There is no sentence being annotated as 'AWA3_Surprise' is also being annotated as '1_claiming_centrality'.

13.2 AntMover VS AWA3

Since AntMover tool is annotating every sentence, we would not include sentences not annotated by AWA3 in this analysis or else it might mask some of the interesting facts from the analysis between the annotations of these tools.


Association between AntMover vs AWA3 (All annotations)
AntMover label
10_announcing_principal_findings 5726.0 3339.0 29017.0 12400.0 25635.0 3382.0 3306.0 4418.0 311.0 17887.0 177.0 6362.0 1694.0
11_evaluation_of_research 4911.0 2994.0 17116.0 10587.0 14771.0 2345.0 3055.0 4885.0 275.0 9229.0 97.0 4753.0 886.0
1_claiming_centrality 6.0 4.0 100.0 155.0 93.0 7.0 11.0 114.0 3.0 14.0 15.0 9.0
2_making_topic_generalizations 1155.0 1331.0 9237.0 3789.0 8603.0 634.0 972.0 1301.0 189.0 2397.0 52.0 2794.0 584.0
5_indicating_a_gap 22.0 14.0 472.0 87.0 446.0 26.0 15.0 29.0 14.0 53.0 2.0 54.0 10.0
9_announcing_present_research 1332.0 417.0 3519.0 1425.0 2436.0 1083.0 760.0 317.0 48.0 7881.0 20.0 770.0 191.0
AWA3_Anyauthor
AWA3_Background
AWA3_Contrast
AWA3_Emphasis
AWA3_Important
AWA3_Important&Summary
AWA3_Novelty
AWA3_Position
AWA3_Question
AWA3_Summary
AWA3_Surprise
AWA3_Tempstat
AWA3_Trend
AWA3 label
25000
20000
15000
10000
5000
0


Association between AntMover vs AWA3 (Exclude Important, Summary and Important&Summary)
AntMover label
10_announcing_principal_findings 5726.0 3339.0 29017.0 12400.0 3306.0 4418.0 311.0 177.0 6362.0 1694.0
11_evaluation_of_research 4911.0 2994.0 17116.0 10587.0 3055.0 4885.0 275.0 97.0 4753.0 886.0
1_claiming_centrality 6.0 4.0 100.0 155.0 11.0 114.0 3.0 15.0 9.0
2_making_topic_generalizations 1155.0 1331.0 9237.0 3789.0 972.0 1301.0 189.0 52.0 2794.0 584.0
5_indicating_a_gap 22.0 14.0 472.0 87.0 15.0 29.0 14.0 2.0 54.0 10.0
9_announcing_present_research 1332.0 417.0 3519.0 1425.0 760.0 317.0 48.0 20.0 770.0 191.0
AWA3_Anyauthor
AWA3_Background
AWA3_Contrast
AWA3_Emphasis
AWA3_Novelty
AWA3_Position
AWA3_Question
AWA3_Surprise
AWA3_Tempstat
AWA3_Trend
AWA3 label
25000
20000
15000
10000
5000
0



As stated before, one interesting fact is that there is no sentence exist which is annotated as '1_claiming_centrality' and 'AWA3_Surprise'.

14. Conclusion

Overall, there are some changes have been done on AWA3 from the previous version of AWA. However, these changes seem just a small changes as can see from the number of sentences being annotated as well as the comparison between AWA3 and AWA. Meanwhile, other tools such as RWT and AntMover seem do not have direct association to AWA3. There are some annotations from RWT or AntMover can be seen to have relations to certain annotations in AWA3 but these relations are not 1:1 relations. This report is not serve as a very detail statistical report but it is a report that provides the overall architecture of the database content as well as some basic comparison of the contents based on the tools. Therefore, this report can be significant for people who are intended to use the corpora in this database for research purpose.