## Data Processing Workflow

## Removing Unnecessary Columns from the Raw Data

All unnecessary columns were removed from the raw data. In addition, X1 and X2 columns were dropped since they were not mention if X4 table[[1]](#endnote-1)[[2]](#footnote-1). This leaves us with 27 sense columns for X5 and 28 sense columns for X6.

## Transferring To Binary

The raw data cells have values varies form zero to one with a step of 0.1. These values were transferred into binary with ‘V’ denoting vote and ‘¬V’ denoting no vote so that the data can be processed similar to word collocation in X7[[3]](#footnote-2). For that end, different thresholds were used (0.3, 0.4 and 0.5). Any value is bigger or equal to the threshold is et to ‘V’, ‘¬V’ otherwise (table 1).

|  |  |
| --- | --- |
| Table 1. | |
| Before binary transformation | After binary transformation (**α = 0.3)** |
|  |  |

In other word, for α = 0.3, if 3 or more out of 10 annotators chose a certain lable fot a certain instance, we consider this label to be voted for. Otherwise (2 annotators or less, we ignored these votes.

## Data Grouping (Leveling Up)

The data were processed at two levels. The leaves level and at level two based on the X4 table. Thus, 12 data frames were created for further processing (2 data sets \* 2 levels \* 3 thresholds = 12)

## Constructing Contingency Tables

Doing a pairwise comparison between each possible sense pair, we construct 2X2 contingency table that has for cells (table 2).

|  |  |  |
| --- | --- | --- |
| Table 2. Contingency Table Structure | | |
|  | Sense 2 | |
| Sense 1 | V | ¬V |
| V | a | b |
| ¬V | c | d |

Where:

**a:** represent the count of instances where enough annotators voted for each sense for a single instance.

**b:** represent the count of instances where enough annotators voted for sense 1 but not enough annotators voted for sense 2.

**c:** represent the count of instances where enough annotators voted for sense 2 but not enough annotators voted for sense 1.

**d:** not enough votes for either sense 1 or sense 2

All contingency tables can be found in the following directory ‘3\_results/binary/contingency\_tables’

Note that all-zero value senses were dropped from further analyse. This means that there is one cell (at max ) with zero count for each row or column in the contingency table. Figure 1 shows a sample of the contingency table that is considered for the processing phase.

|  |
| --- |
| Fig 1. Sample of Accepted Contingency Tables for the Discogem Dataset at the Leaves Level with α = 0.3 |
|  |

By the end of the data processing, X10 sense pairs were analysed using different statistical methods.

# Data Analysis -Part One

## Chi-Squared Test and Fisher's Exact Test and P-Value

Chi-squared value was calculated for all the eligible pairs. However, when calculating Chi-squared value, the expected table is constructed. It is recommended that all the cells of the expected table to have a value bigger than 5 for Chi-squared to be calculated [[4]](#footnote-3). If any of the expected table cells is less than five (figure 2), other methods should be followed. In this report, Fisher's exact test was used whenever the expected table has a cell with a count less than five. Afterwards, P-Value is calculated from the Chi-squared value (figure 3) and Fisher's exact test (figure 4)

|  |
| --- |
| Fig 2. Example of an Expected Table with a Value Less than Five (contrast and synchronous, Discogem Dataset at the Leaves Level with α = 0.3) |
|  |

|  |
| --- |
| Fig 3. Chi-Squared P-Value for Discogem Dataset at the Leaves Level with α = 0.3 |
|  |
| Note: ‘fisher’ here means that the P-value was calculated based on Fisher’s exact test in another table |

|  |
| --- |
| Fig 4. Fisher Exact Test P-Value for Discogem Dataset at the Leaves Level with α = 0.3 |
|  |
| Note: ‘chi2 here means that the P-value was calculated based on . Chi-Squared test |

P-value less than 0.05[[5]](#footnote-4) indicate an association between the sense pair. But it does not indicate that if it is a positive association (the aim of this report) or a negative one[[6]](#footnote-5). And for that, the Odd Ration (OR) was calculated.

Figure 5 and 6 shows the P-values for same binary transformed data set at level two

|  |
| --- |
| Fig 5. Chi-Squared P-Value for Discogem Dataset at the Level Two with α = 0.3 |
|  |
| Note: ‘fisher’ here means that the P-value was calculated based on Fisher’s exact test in another table |

|  |
| --- |
| Fig 6. Fisher Exact Test P-Value for Discogem Dataset at the Level Two with α = 0.3 |
|  |
| Note: ‘chi2 here means that the P-value was calculated based on . Chi-Squared test |

## Odd Ration (OR)

Odd Ratio (OR) was calculated whenerver the P-value is less than 0.05. OR greater than 1 indicates a positive associations, and OR less than 1 indicates a negative association. Figure 5 shows the OR ratios for the pairs with a P-value less than 0.05. to address issues suzero counts or small cell frequencies in contingency tables, a ‘continuity correction’ by adding 0.5 was used to help stabilize the calculations and prevents division by zero errors or extreme values in cases where one or more cells have low counts[[7]](#footnote-6).

|  |
| --- |
| Fig 7. OR Values for Discogem Dataset at the Leaves Level with α = 0.3 |
|  |
| Note: Cells in green indicates positive correlation, while cells in red indicates negative correlation |

Note that ‘NA’ values indiacate the cases where the dominator of the Yule’s Q was zero and cannot be calcuated

Based on OR values, we can conclude the positively associated Sene pairs (figure 6)

|  |
| --- |
| Fig 8. Positively Associated Sense Pairs Based on OR Values for Discogem Dataset at the Leaves Level with α = 0.3 |
|  |

Figure 9 shows the OR values for same binary transformed data set at level two while Figure 10 lists the positive associated sense pair on that level

|  |
| --- |
| Fig 9. OR Values for Discogem Dataset at the Level Two with α = 0.3 |
|  |

|  |
| --- |
| Fig 8. Positively Associated Sense Pairs Based on OR Values for Discogem Dataset at the Level Two with α = 0.3 |
|  |

Thus, first iteration of the decision pipeline is as shown in figure 9.

|  |
| --- |
| Fig 9. Decision Pipeline, First Iteration (Based on OR) |
| A diagram of a flowchart  Description automatically generated |

## Additional Measure, Yule’s Q

Yule's Q is a measure of association for 2x2 contingency tables, ranging from -1 to 1. A value of 1 indicates a perfect positive association, -1 indicates a perfect negative association, and 0 signifies no association between the variables. A value of 0.5 indicates a moderate positive association, a value of -0.5 indicates a negative one. Figure 10 depicts Yule’Q values for for Discogem Dataset at the Leaves Level with α = 0.3

Explain NA

|  |
| --- |
| Fig 10. Yule’sQ Positive and Negative Sense Pair Associations Highlighted in Green and Red Respectively (for Discogem Dataset at the Leaves Level with α = 0.3) |
|  |

Figure 11 lists the positively associated sense pairs. While figure 12 shows the negatively associated ones. This is considering -0.5 to be the limit value of moderate negative association.

|  |
| --- |
| Fig 11. Positively Associated Sense Pairs Based on Yule’s Q Values for Discogem Dataset at the Leaves Level with α = 0.3 |
|  |

|  |
| --- |
| Fig 12. Negatively Associated Sense Pairs Based on Yule’s Q Values for Discogem Dataset at the Leaves Level with α = 0.3 |
|  |

Note that Yule’s Q values of 0.8 and above (-0.8 and below) denotes ‘strong’ positive (negative) association, while a value of 0.5 (-0.5) denotes a ‘moderate’ positive (negative) association.

Figures 13, 14 and 15 shows the same results for Level Two

|  |
| --- |
| Fig 13. Yule’sQ Positive and Negative Sense Pair Associations Highlighted in Green and Red Respectively (for Discogem Dataset at Level Two with α = 0.3) |
|  |

|  |
| --- |
| Fig 14. Positively Associated Sense Pairs Based on Yule’s Q Values for Discogem Dataset at Level Two with α = 0.3 |
|  |

|  |
| --- |
| Fig 15. Negatively Associated Sense Pairs Based on Yule’s Q Values for Discogem Dataset at Level Two with α = 0.3 |
|  |

## Accounting for Yule ’Q results in the OR results

OR and Yule’s Q results in somehow similar but not the same results. Comparing figure 8 and figure 11 results in overlapping for some sense pairs (fig 16)

|  |
| --- |
| Fig 16 |
| Positively associated pairs based on OR (for discogem, leaves level at 0.3): |
|  |
| Positively associated pairs based on Yule’s Q (for discogem, leaves level at 0.3): |
|  |
| Overlapped sense pairs: |
| Succession | arg2-as-detail  Arg1-as-denier | arge2-as-denier  Similarity | conjunction |

For the purpose of this report, we will continue with the Chi-Squared/Fisher’s exact test OR based positively association pairs. We will add an annotation of (YQ+) just to show that certain sense pairs were also deemed as positively associated based on Yule’s Q test. Thus, the collection of positively associated sense pairs become as shown in Fig 17:

|  |
| --- |
| Fig 17. |
| synchronous | precedence  succession | arg2-as-detail YQ+  result | arg1-as-detail  arg1-as-denier | arg2-as-denier YQ+  arg2-as-denier | contrast  similarity | conjunction YQ+  arg2-as-instance | arg2-as-detail |

Yule’s Q also yields the set of sense pairs that are negatively associated (figure 12) which is the opposite and a disprove of what we are looking for. To account for Yule’s Q based negative associations, these set of sense pairs should be excluded from the collection of positively associated sense pairs. After examination, no sense pair were excluded.

Similar analysis is done for Level Two (fig 18)

|  |
| --- |
| Fig 18. |
|  |

Based on the above mentioned discussion, the decision pipeline is modified as seen in fig 19

|  |
| --- |
| Fig 19. Decision Pipeline, Second Iteration (Based on OR & Yule’s Q) |
|  |

# Data Analysis – Part Two

## Closer Look At Our Contingency Tables

So far, the statistical methods used in the first part of the data analysis considers all of the four cells in the contingency table. However, looking closer at the contingency table one can find the followings:

1. Vote-vote cells represent the case there has been some confusion about the sense (in a sense pair) that exists in a single instance. For α = 0.3, It reads as following: six annotators at least have their votes split into to senses for the same instance; 3 annotators at least voted for sense one, and another three annotators at least voted for sense two. This type of sense pair is what this report is aiming to find with statistical significant.
2. Vote-¬Vote cells represent the case where there hasn't been confusion about the sense (in a sense pair) that exists in a single instance. For α = 0.3, 3 or more annotaters voted for sense one while two to none annotorrs voted for sense two. It can be said that the annotors were able to distinguish between these two sense for that specific instance. This is the opposite of what the report is trying to find.
3. Vote-¬Vote cells is similar to Vote-¬Vote
4. ¬Vote--¬Vote does not tell information about the sense pair in hand, the annottaotrs voted for other sense there could have been confusion or no confusion between another sense pair.

Thus in the following section uses other methods that do not take into account the ¬Vote-¬Vote cells

### Pointwise Mutual Information PMI

Explain NA

|  |
| --- |
| Fig . PMI values for Discogem Dataset at Level Two with α = 0.3 |
|  |

|  |
| --- |
| Fig . Negatively Associated Sense Pairs Based on PMI Values for Discogem Dataset at Level Two with α = 0.3 |
|  |

Comparing the PMI based negatively associated with the positively associaeted pairs of figure 17, we notice two overlaps pairs (result | arg1-as-detail, arg2-as-denier| contrast, and arg2-as-instance | arg2-as-detail ). We exclude these pairs and we end up with the suriving postiviel associated sense pair fig XX

|  |
| --- |
| Fig. survived positively associated sense pair |
| synchronous | precedence  succession | arg2-as-detail YQ+  arg1-as-denier | arg2-as-denier YQ+  similarity | conjunction YQ+ |

Based on the above mentioned discussion, the decision pipeline is modified as seen in fig

|  |
| --- |
| Fig . Decision Pipeline, Third and Final Iteration (Based on OR, Yule’s Q and PMI) |
|  |

1. ## Deeper Look at the Survived Positively Associated Pairs

   Considering the contingency tables for the four survived positively associated sense pairs for the discogem data set at the Leaves Level and being transformed into binary data at 0.3 threshold (fig) we find the following:

   |  |  |
   | --- | --- |
   |  |  |
   | precedence V ¬V  synchronous  V 13 49  ¬V 589 5854 | arg2-as-detail V ¬V  succession  V 6 6  ¬V 1410 5083 |
   | arg2-as-denier V ¬V  arg1-as-denier  V 23 98  ¬V 330 6054 | conjunction V ¬V  similarity  V 60 20  ¬V 2247 4178 |

   The majority of the counts is for the no\_vote-no\_vote cells, the count for vote-vote is minimum or second to minimum. However, the vote-vote cells are the ones that actually tell us that the annotators have been confused to choose a single sense that represetnes an instance. For example, let us take the cpnjunction | similarity sense pair. One could read it as follows: Out of 2327 (20+60+2247) instances where they appeared in the data (weather separately or together), there has been 60 cases where the annotators’ vote was split between these senses, and 2267 (2247+20) cases where the annotatoes favoured one ovet the other. This means that when these pair of instances appears in our data, the annotaotes confused between these senses 2.5% (60/2327) of the times, and they were able to favour one over another 97.5% ((2247+20)/2327) of the time. Thus one could challenge the findings of our results in the previous sections. Considering that our threshold is 0.3, in the 2.5% of the instances that the two senses appeared, six or more annotators have their votes split between the senses of this pair for a each of these instance (three or more voted for sense 1, and three or more voted for sense 2). On the other hand, in the 97.5% of the instances, the annototoes were able to favor one over the other (three or more annototares voted for one of these sense, and two or less voted fo the other one).

   Based on the discussion above, one could propose an indicator to quantify how many times the annotator where confused about which sense to choose out of a pair of senses for the same instance

   Proposed indicator

   In the case of complete confusion between the senses

   |  |
   | --- |
   |  |
   |  |

   |  |
   | --- |
   |  |
   |  |

   [↑](#endnote-ref-1)
2. [↑](#footnote-ref-1)
3. [↑](#footnote-ref-2)
4. Add a reference to support this point [↑](#footnote-ref-3)
5. Add resource to support this claim [↑](#footnote-ref-4)
6. Add resource to support this [↑](#footnote-ref-5)
7. Add Resource [↑](#footnote-ref-6)