

TRANSLATOR Hits for generated opposites

KBC

3/15/2021

Comparison with currently-used opposition relation

The reference data: Bill can add details. He pulled out pairs of concepts that are currently annotated as negatives in ... maybe just PATO, maybe all of the ontologies that Mike said to look at. There are 215 pairs in total, of which 213 are from PATO.

I ran that data through a script called pullIDPairs.pl, which pulls out just the IDs of the pairs, discarding the terms. It also does some sorting across the entire set of pairs, which is necessary for comparing them to other sources of opposites (in this instance, my script).

My data: I ran my script findAntonyms.pl on PATO version X.X. Then I used the pullIDPairs.pl script (see above) to pull out just the IDs and do the necessary sorting.

NB: I might need to rerun my script on PATO—I think that the version of the script that I used produces duplicates $\text{opp}(x, y)$ and $\text{opp}(y, x)$, which is fixed in later versions. I think/hope :-)

Finally, I used the UNIX utility `comm -12 reference.file my.file` to find the overlaps, if there be any. ##
Validating generated opposites against a corpus

Data: I took terms from PATO and added oppositional prefixes. Some occur natively in PATO already, some don't. Then I did a quoted search of PubMed Central for all of them—non-negated, “native” negated, and generated.

Question: what happens if we try to validate generated negatives by seeing whether or not they actually get used?

Comparison with Chris's paper as of 2021-03-31

NB: I think I need to redo these—not sure his output is correctly sorted for the overlap-finding utility that I used. This means that the finding of no overlaps might be wrong.

1. Wrote a script that converts Chris's id-id-logical-textual format to id-id. Required more mucking about than one might have expected, but: whatever.
2. For Chris's textual-evidence-only, logical-evidence-only, and both-kinds-of-evidence pairs, I used the UNIX command `comm` (thanks, Bill!) to find the size of the union of what he found with what I found.

The y-axis in these graphs is scaled to the total number of pairs found by all three of Chris's methods combined: 438.

The number for K (what we are getting as of today) is constant across all three graphs at 105 pairs.

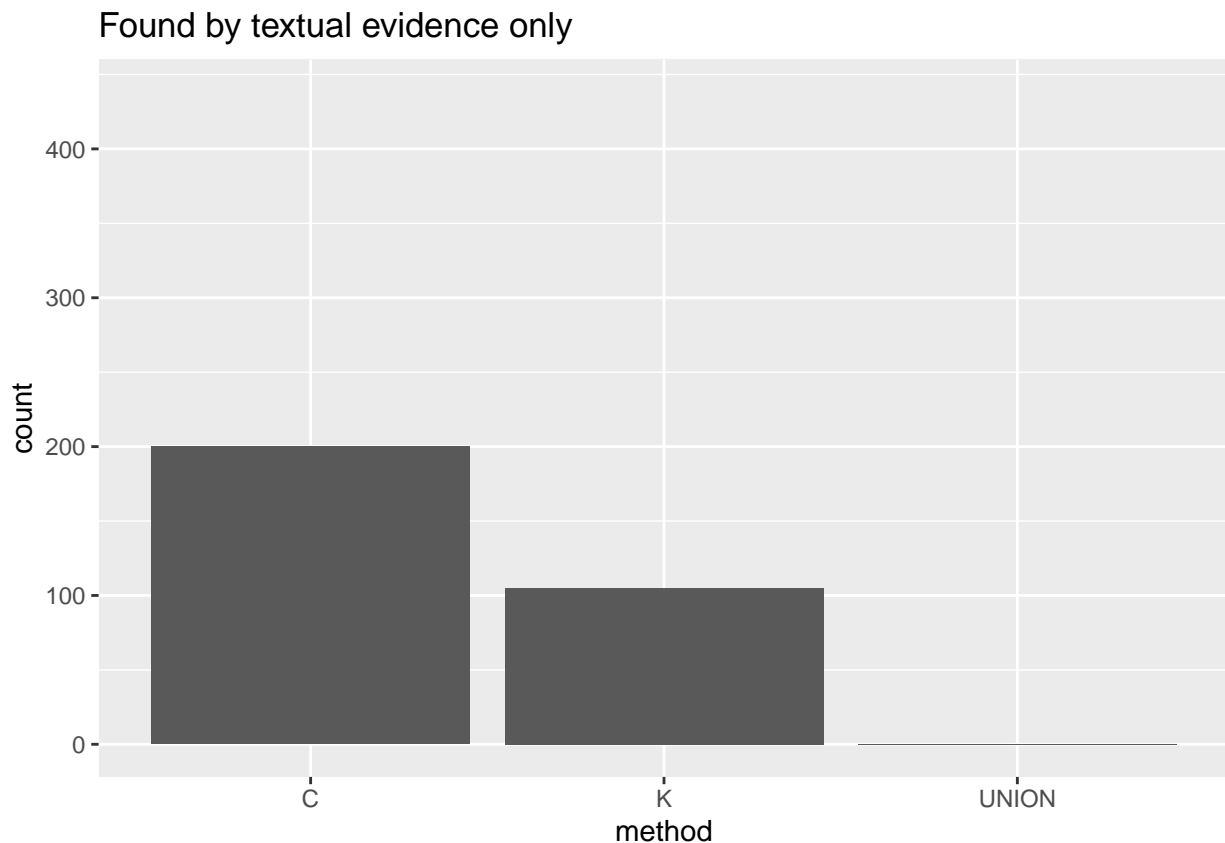
```
we.found.textual <- 105
chris.found.textual <- 200
overlap.textual <- 0
```

```
found.by <- as_tibble(rbind(
  c("TEXT", "K", 105),
  c("TEXT", "C", 200),
  c("TEXT", "UNION", 0)
))
```

```
## Warning: The 'x' argument of 'as_tibble.matrix()' must have column names if '.name_repair' is omitted.
## Using compatibility '.name_repair'.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_warnings()' to see where this warning was generated.
```

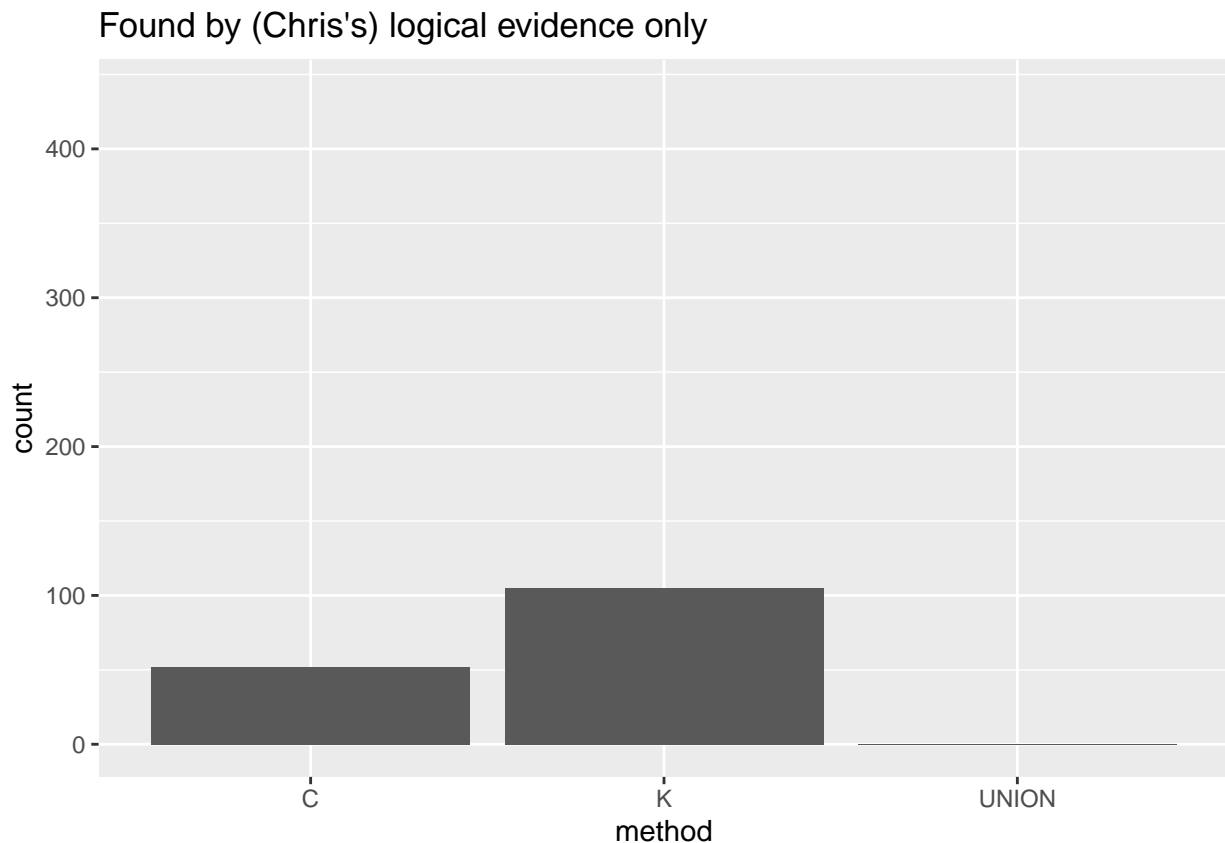
```
# housekeeping: column names and types
colnames(found.by) <- c("evidence.from", "method", "count")
found.by$evidence.from <- as.factor(found.by$evidence.from)
found.by$method <- as.factor(found.by$method)
found.by$count <- as.integer(found.by$count)

ggplot(data = found.by, mapping = aes(x = method, y = count)) +
  geom_bar(stat = "identity") +
  ylim(0, 438) +
  labs(title = "Found by textual evidence only")
```



```
found.by <- as_tibble(rbind(
  c("TEXT", "K", 105),
  c("LOGICAL", "C", 52),
  c("N/A", "UNION", 0)
))
# housekeeping: column names and types
colnames(found.by) <- c("evidence.from", "method", "count")
found.by$evidence.from <- as.factor(found.by$evidence.from)
found.by$method <- as.factor(found.by$method)
found.by$count <- as.integer(found.by$count)

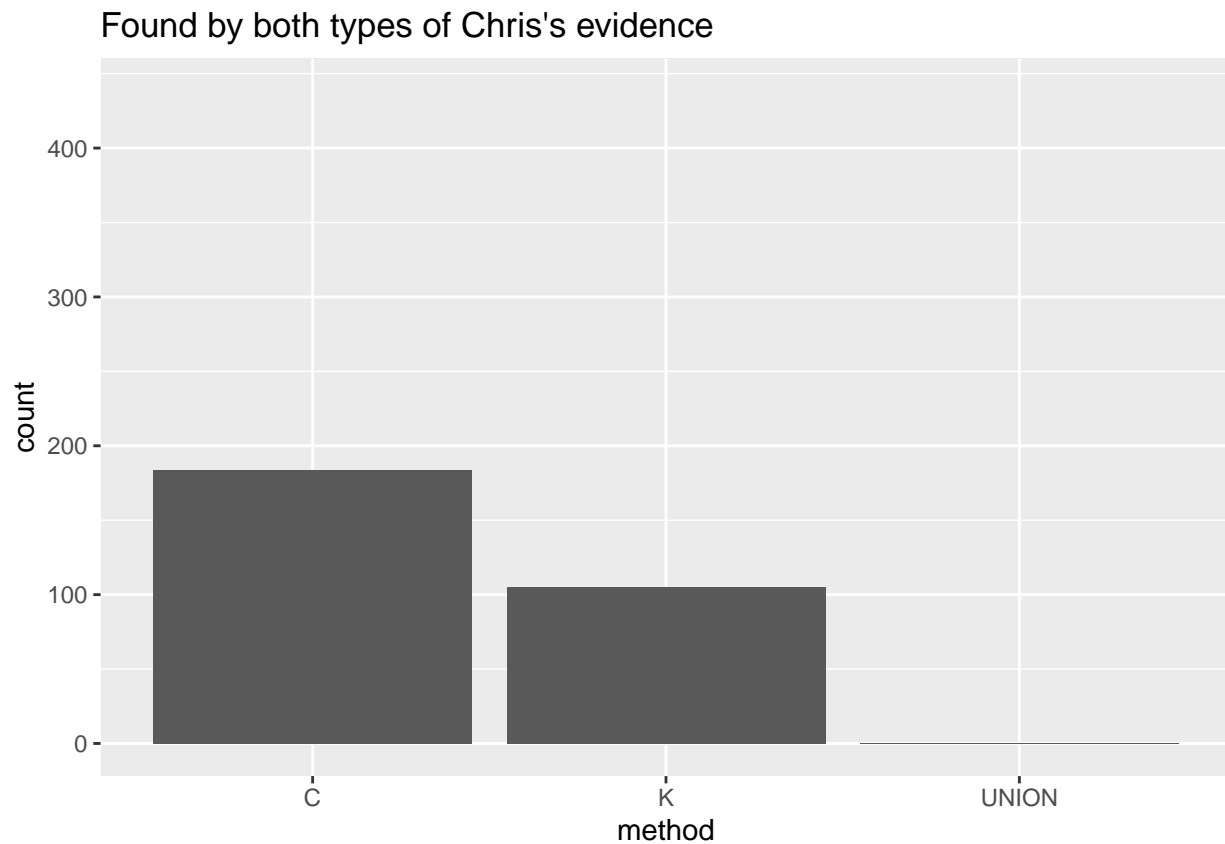
ggplot(data = found.by, mapping = aes(x = method, y = count)) +
  geom_bar(stat = "identity") +
  ylim(0, 438) +
  labs(title = "Found by (Chris's) logical evidence only")
```



```
found.by <- as_tibble(rbind(
  c("TEXT", "K", 105),
  c("BOTH", "C", 184),
  c("N/A", "UNION", 0)
))
# housekeeping: column names and types
colnames(found.by) <- c("evidence.from", "method", "count")
found.by$evidence.from <- as.factor(found.by$evidence.from)
found.by$method <- as.factor(found.by$method)
```

```
found.by$count <- as.integer(found.by$count)

ggplot(data = found.by, mapping = aes(x = method, y = count)) +
  geom_bar(stat = "identity") +
  ylim(0, 438) +
  labs(title = "Found by both types of Chris's evidence")
```



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.