Twitter Alignment Results and Measure Comparison

Gabe and Jake

July 29, 2015

Overview

This write-up presents basic tests of alignment to power and of alignment in general on Twitter. We find significant positive alignment in general, as well as alignment to power based on verification status using our measure of alignment (log-odds with the not-A baseline). We find positive alignment in general using the DNM measure, but no alignment to power (in fact, we see divergence from power, though this is likely a spurious sparsity result). We then analyze the dimensions on which these measure differ and show that smoothing and baseline choices have relatively small effects and the difference between our results appear to be mainly driven by the switch from subtractive to logarithmic calculation.

Alignment (and alignment to power) on the two main measures

[Unechoed block of code to load results.]

```
##
Read 28.9% of 345603 rows
Read 63.7% of 345603 rows
Read 98.4% of 345603 rows
Read 345603 rows and 20 (of 20) columns from 0.048 GB file in 00:00:05
```

A brief overview of the dataset The dataset tested here is based on twtr.py in commit `c91d11563cf6f805d80348189f14dadae4cb206d` in the alignment project. There were 345603 speaker-replier-marker triplets, with 48 markers chosen from the most frequent tokens in the Twitter dataset, including punctuation. For virtually all of the data discussed here, we exclude triplets where the speaker has less than 5 messages with the marker or less than 5 messages without the marker (this reduces noise in the estimates of the two probabilities in the alignment calculation); this leaves 16911 triplets.

First things first, we'd like to get a sense of how the alignmnt measure decisions affect the outcomes. We're currently considering three variables in the calculation of alignment:

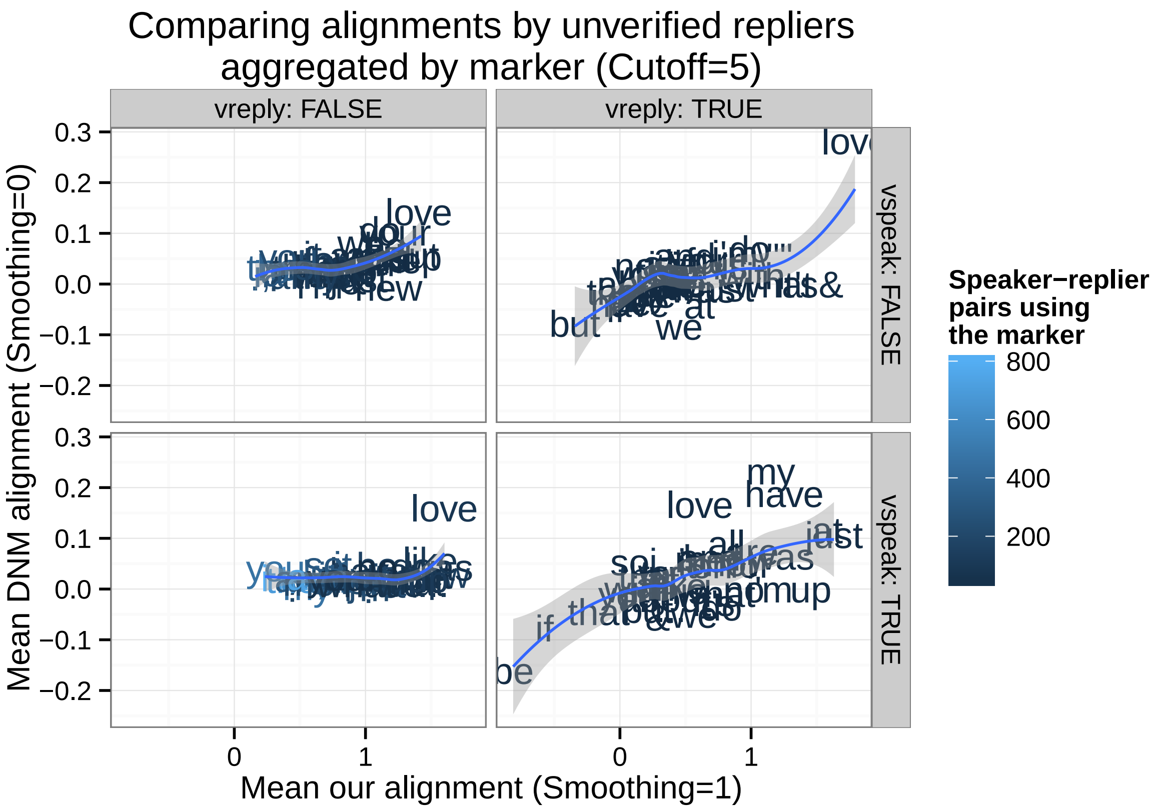
1. Baseline measure ($p(B)$ vs. $p(B|\neg A)$)
2. Logarithmic vs. non-logarithmic difference
3. Smoothing

Our proposed metric uses the $p(B|\neg A)$ baseline, in log space, with +1 smoothing. The DNM metric that has been standardly used has the $p(B)$ baseline, in non-log space, with no smoothing. We'll look at each of the influence of each of these factors in turn.

Why do we want to change the measure in the first place? Our toy example work found some problems with the original metric. For low-frequency markers, it limits the amount of positive alignment that can be detected (alignment is at most $p(B|A)$ since $p(B) \geq 0$). This is an especially noticeable problem with the short messages on Twitter and in child-directed/produced speech. Moving to log-space removes this upper limit. The original metric also loses discriminative power for moderate- and high-frequency markers, as $p(B) = p(A)p(B|A) + (1 - p(A))p(B|\neg A)$, which linearly approaches $p(B|A)$ as the marker frequency increases. Changing the baseline to $p(B|\neg A)$ addresses this problem.

We find substantially different distributions and results depending on the alignment measure. We'll start out by just looking at the two measures of interest (ours vs. DNM's), and later look at the full set of 8 possible measures to identify how each chocie affects our conclusions. Plotting these two core measures against each other by marker (and split by verification status, since we think that affects the means of the markers), we see that DNM shows substantially reduced and less discriminative values. Only when the range of alignments is very large (as in the verified replier cases) do we see substantial correlation between the two.

```
d2 <- d %>%
  filter((ba+nba)>=cutoff,(bna+nbna)>=cutoff) %>%
  group_by(vspeak,vreply,category) %>%
  summarize(convs=n(),Ours=mean(lo1),DNM=mean(sd0))

ggplot(d2,aes(x=Ours,y=DNM,color=convs,label=category)) + geom_text() + geom_smooth(method="loess") +fa
```



(Throughout this write-up, we are only considering speaker-replier-marker triplets where the speaker says the marker at least 5 times and the speaker doesn't say the marker at least 5 times; this is similar to the 10-instance cutoff DNM et al used.) The correlations by speaker-replier verification:

```
d2 %>% group_by(vspeak,vreply) %>% summarize(correlation=cor(Ours,DNM))
```

```
## Source: local data table [4 x 3]
## Groups: vspeak
##
##   vspeak vreply correlation
## 1   TRUE  FALSE   0.2430334
## 2  FALSE  FALSE   0.5193319
## 3   TRUE   TRUE   0.6298634
## 4  FALSE   TRUE   0.5690604
```

So there's clearly correlation; these measures aren't measuring completely different things. The main difference is that our measure appears to have somewhat better resolution, for the reasons argued above, on the fairly small positive alignment effects that are predicted under accommodation theories. And we do see in these plots that the alignment estimates are consistently above zero, especially with our alignment estimates. Just to confirm that this is significant in both measures by a t-test:

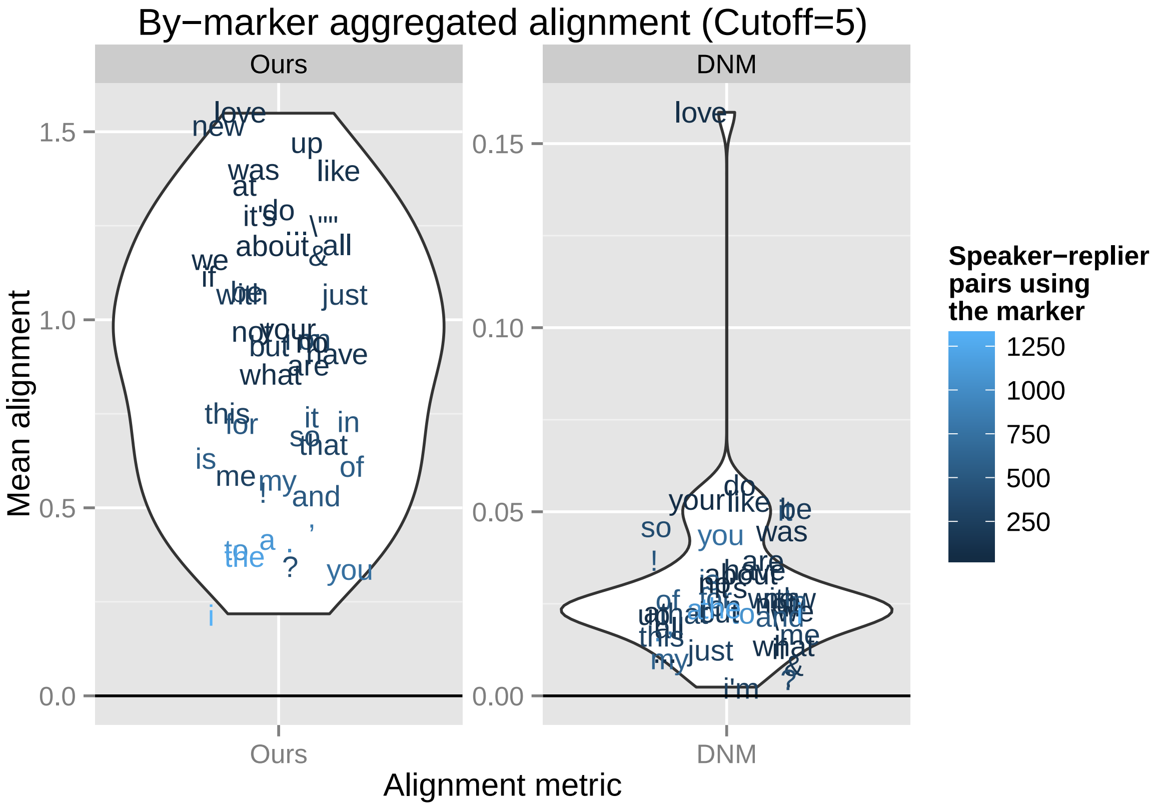
```
d2 <- d %>%
  filter((ba+nba)>=cutoff,(bna+nbna)>=cutoff) %>%
  group_by(category) %>%
  summarize(convs=n(),Ours=mean(lo1),DNM=mean(sd0))

t.test(d2$Ours)
```

```
##
##  One Sample t-test
##
## data:  d2$Ours
## t = 17.35, df = 47, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.7895502 0.9966628
## sample estimates:
## mean of x
## 0.8931065
```

```
t.test(d2$DNM)
```

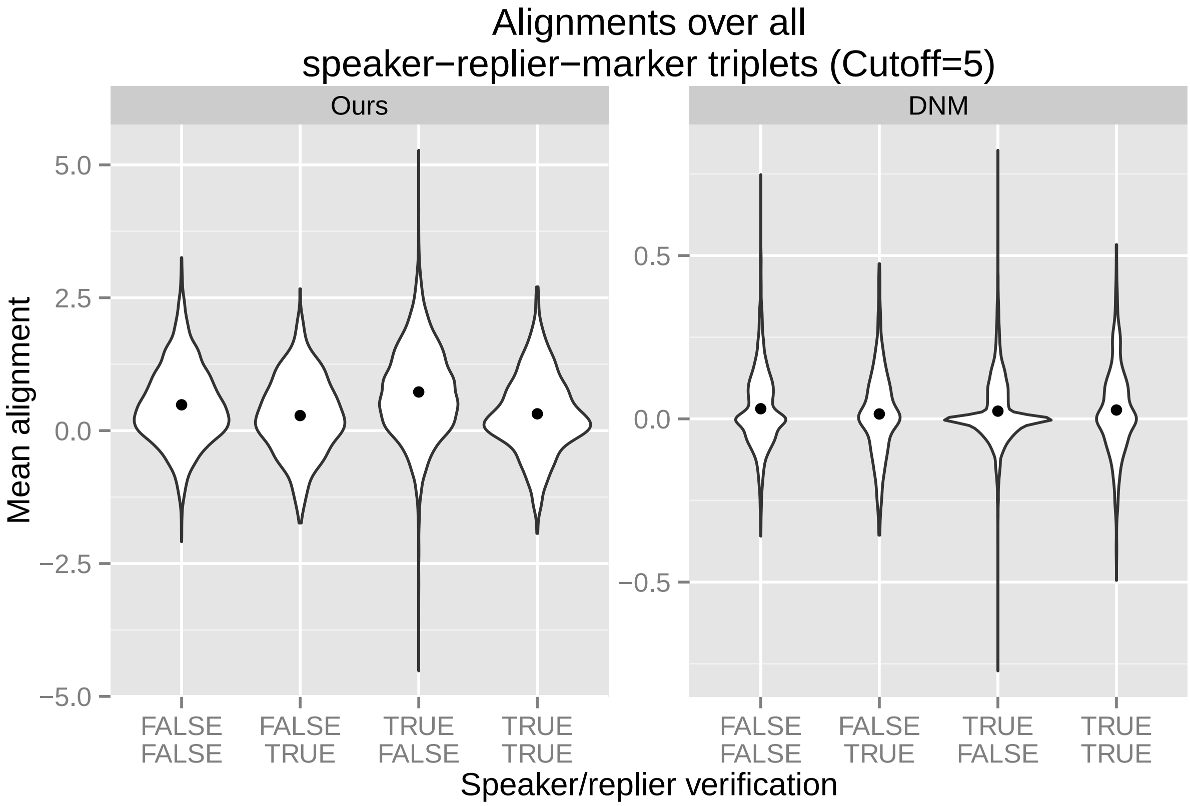
```
##
##  One Sample t-test
##
## data:  d2$DNM
## t = 8.6211, df = 47, p-value = 3.063e-11
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.02222362 0.03575227
## sample estimates:
##  mean of x
## 0.02898795
```

```
d2 <- d2 %>%
  gather(alignment,mean,Ours,DNM)

ggplot(d2,aes(x=alignment,y=mean,color=convs,label=category)) + geom_violin() + geom_text(position=posi
```



Now let's turn to our main research question: alignment to power. We find clearer effects under our log-odds measure than under the DNM measure.

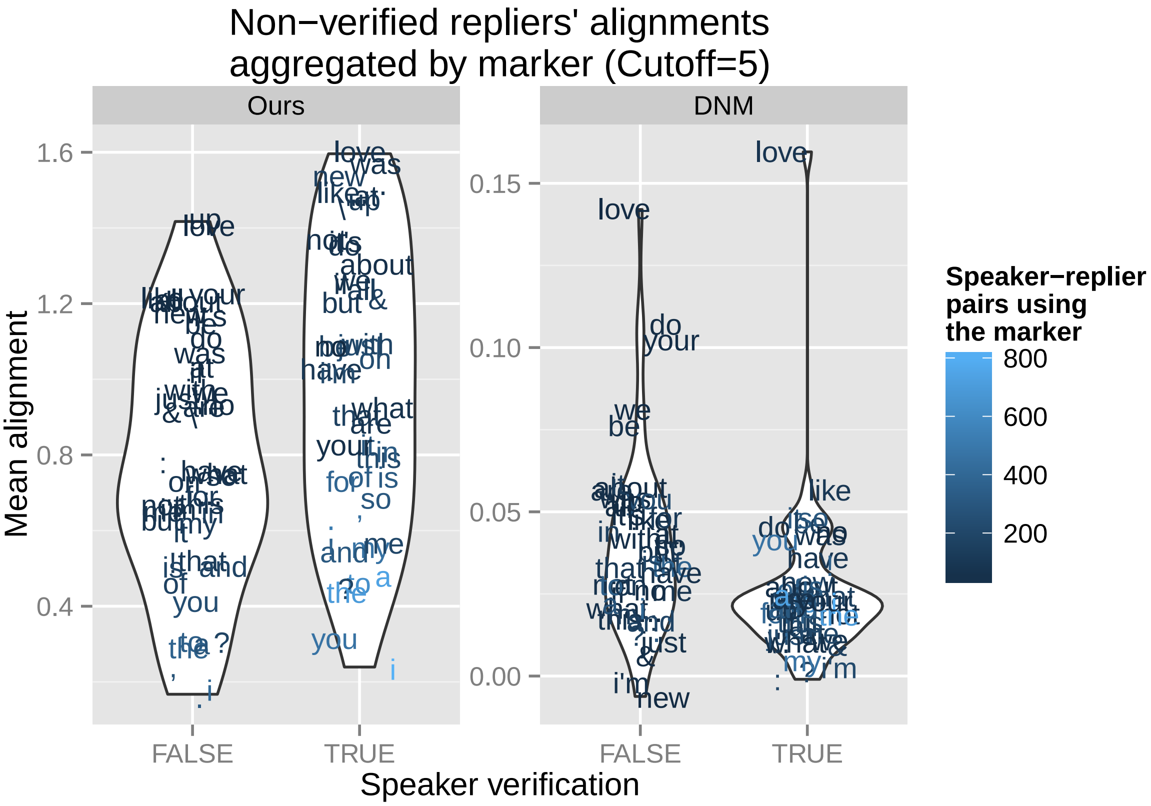
```
d2 <- d %>%
  filter((ba+nba)>=cutoff,(bna+nbna)>=cutoff) %>%
  mutate(Ours=lo1,DNM=sd0) %>%
  gather(alignment,mean,Ours,DNM)


ggplot(d2,aes(x=paste(vspeak,vreply,sep="\n"),y=mean,label=category)) + geom_violin() + stat_summary(ge
```



Our metric has clearer differences in the means, and better looking distributions as well. The weird peaks in the DNM violins at and just under 0 are due to the lack of smoothing; especially for rarer markers, there will be many times that a replier just never uses the marker. If the replier *never* uses it in conversation with the speaker, the DNM alignment has to be zero (as $p(B|A) = p(B) = 0$), whereas due to the smoothing on our model, the alignment may be non-zero if there is a difference in the number of instances of A and not A. Furthermore, the DNM alignment has a peak of small negative alignments due to the sampling of rare events. Because $p(BA) << p(B\neg A)$, we will fairly often get cases where we'll draw one or two instances of $B\neg A$ but none of BA, which means that $p(B)$ will be small but positive, whereas $p(B|A)$ will be zero. This happens even if B and A are truly independent, leading to the spurious negative bulge we see in the DNM violins, and is probably the source of the supposed negative alignment to power in the DNM results.

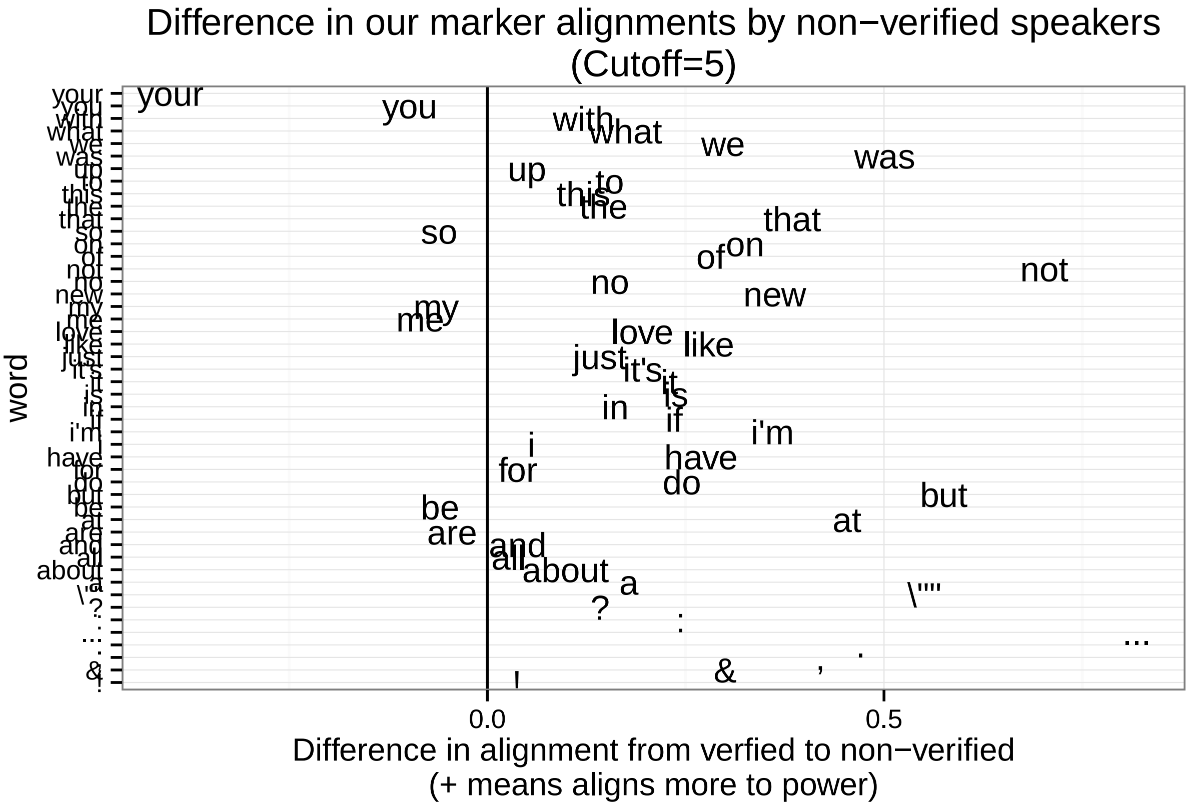
Let's look at the results just for unverified repliers, as we have more data for them and have a clear expectation of increased alignment to verified speakers.

```
d2 <- d %>%
  filter(vreply==F,(ba+nba)>=cutoff,(bna+nbna)>=cutoff) %>%
  group_by(vspeak,category) %>%
  summarize(convs=n(),Ours=mean(lo1),DNM=mean(sd0)) %>%
  gather(alignment,mean,Ours,DNM)

ggplot(d2,aes(x=vspeak,y=mean,color=convs,label=category)) + geom_violin() + geom_text(position=positio
```



And let's test the by-marker alignment to power (the difference in alignment values when responding to a verified vs. unverified speaker):

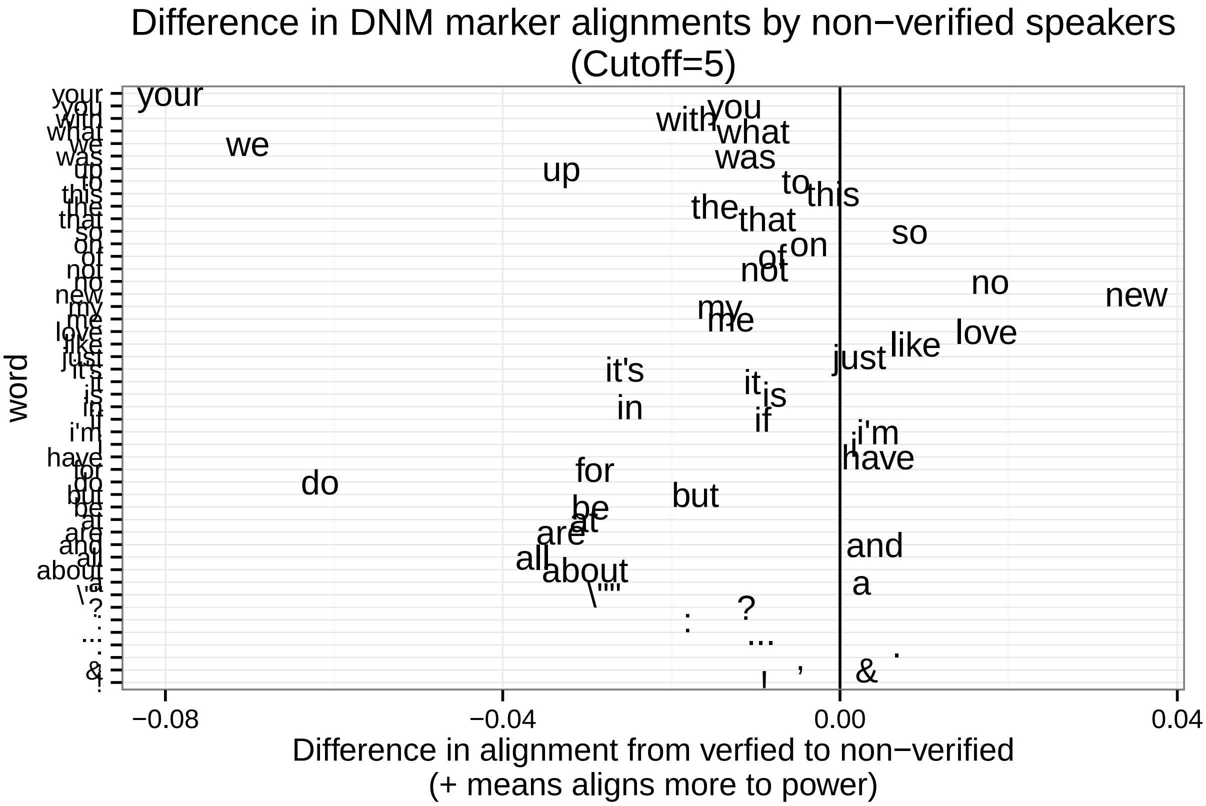
```
d2 <- d %>%
  filter(vreply==F,(ba+nba)>=cutoff,(bna+nbna)>=cutoff) %>%
  group_by(vspeak,category) %>%
  summarize(mean=mean(lo1)) %>%
  spread(vspeak,mean,fill=NA) %>%
  transmute(category=category,tvalue=`TRUE`,fvalue=`FALSE`)

ggplot(d2,aes(x=tvalue-fvalue,y=category)) + geom_vline(xintercept=0) + geom_text(aes(label=category),s
```



```
t.test(d2$tvalue,d2$fvalue)
```

```
##
##  Welch Two Sample t-test
##
## data:  d2$tvalue and d2$fvalue
## t = 2.8645, df = 92.441, p-value = 0.005169
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.06364297 0.35136737
## sample estimates:
## mean of x mean of y
## 0.9776452 0.7701401
```

```
d2 <- d %>%
  filter(vreply==F,(ba+nba)>=cutoff,(bna+nbna)>=cutoff) %>%
  group_by(vspeak,category) %>%
  summarize(mean=mean(sd0)) %>%
  spread(vspeak,mean,fill=NA) %>%
  transmute(category=category,tvalue=`TRUE`,fvalue=`FALSE`)

ggplot(d2,aes(x=tvalue-fvalue,y=category)) + geom_vline(xintercept=0) + geom_text(aes(label=category),s
```



```r
t.test(d2$tvalue,d2$fvalue)
```

```
##
##  Welch Two Sample t-test
##
## data:  d2$tvalue and d2$fvalue
## t = -2.4506, df = 91.778, p-value = 0.01616
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.023407872 -0.002450408
## sample estimates:
##  mean of x  mean of y
## 0.02591320 0.03884234
```

Our alignment measure is positive for most markers, with many of the negative power-based alignments occurring with personal pronouns, which often show negative alignments in previous work. We also see an overall significant positive power-based alignment based on a paired t-test. The DNM alignments, on the other hand, are more evenly divided between positive and negative power-based alignments, without a clear pattern in the negative alignment words. There is also a significant *negative* power-based alignment, which I don't think any of the previous literature would predict. This is probably due to sparsity in drawing rare markers, as discussed above. (Note that DNM's paper may not have been substantially affected by this negative bulge because they used marker categories, which have higher base-rates.)

So, in summary, we can argue that our alignment measure is more appropriate for distributions like those we expect (mostly positive by-marker alignments, with low- to moderate-frequency markers) than the DNM alignment measure. Looking at the distribution of alignment estimates, the two measures are correlated but ours seems to better distinguish small positive alignments from zero (see the DNM vs. ours scatterplot) and looks less noisy (see the violin plot). Thanks to these advantages, our measure is able to identify a power-based aligment that the DNM measure does not.

Other measures (Or: why are we so different from DNM?)

Above, we identified three differences between our measure and the DNM measure. We want to better understand exactly how each of these choices affects the results, though, with an eye to whether our arguments for preferring our measure are actually valid on real data. This section compares the eight logical possibilities for metrics (baselines x log/non-log x smoothing).

```
d <- df[,list(ba=ba,nba=nba,bna=bna,nbna=nbna,
              vspeak=verifiedSpeaker,vreply=verifiedReplier,
              sid=speakerId,rid=replierId,category=category,
              pyalign=alignment,reciprocity=reciprocity,dnm=dnmalignment),]

d2 <- d %>%
  mutate(valid=((ba+nba)>=cutoff&(bna+nbna)>=cutoff)) %>%
  transmute(valid=valid,
            nl0=smoothalign(d,0,"logodds"),nl1=smoothalign(d,1,"logodds"),
            ns0=smoothalign(d,0,"subodds"),ns1=smoothalign(d,1,"subodds"),
            al0=smoothalign(d,0,"logdnm"),al1=smoothalign(d,1,"logdnm"),
            as0=smoothalign(d,0,"subdnm"),as1=smoothalign(d,1,"subdnm")) %>%
  filter(valid==T,is.finite(nl0),is.finite(al0))
d2$valid <- NULL

t <- cor(d2)
round(t[c(1,5,2,6,3,7,4,8),c(1,5,2,6,3,7,4,8)]*100)/100 #re-order columns for easier comparison
```

```
##      nl0  al0  nl1  al1  ns0  as0  ns1  as1
## nl0 1.00 0.97 0.97 0.91 0.79 0.83 0.85 0.88
## al0 0.97 1.00 0.98 0.97 0.71 0.78 0.79 0.86
## nl1 0.97 0.98 1.00 0.97 0.73 0.78 0.82 0.87
## al1 0.91 0.97 0.97 1.00 0.59 0.68 0.71 0.81
## ns0 0.79 0.71 0.73 0.59 1.00 0.97 0.98 0.91
## as0 0.83 0.78 0.78 0.68 0.97 1.00 0.97 0.96
## ns1 0.85 0.79 0.82 0.71 0.98 0.97 1.00 0.96
## as1 0.88 0.86 0.87 0.81 0.91 0.96 0.96 1.00
```

Above is the correlation table based on all speaker-replier-marker triplets that with at least 5 instances of the speaker saying the marker, 5 instances of them not saying the marker, and a finite alignment score on all measures. The three-character column labels represent the choices on the three dimensions: baseline (*a* vs. *n*ota), difference (*l*og vs. *s*ubtract), smoothing (0/1). Our alignment is “nl0”; DNM’s is “as1”

This correlation table essentially consists of two blocks; first the four logarithmic measures, then the four subtractive measures. Changing the baseline and/or smoothing has very little effect on the correlation ($0.91 - 0.99$ for all pairs). Changing from logarthmic to subtractive space, though, can greatly decrease the correlation ($0.78 - 0.82$ from just changing space).

A similar, although less stark, pattern is seen if we look at rank-based correlation. Spearman’s rho shows equally high correlation on baseline/smoothing changes, with less of a decrease on space changes. Kendall’s tau shows a noticeable decrease from baseline/smoothing changes, but still has the strongest decreases from space changes.

```
t <- cor(d2,method="spearman")
round(t[c(1,5,2,6,3,7,4,8),c(1,5,2,6,3,7,4,8)]*100)/100
```

```
##      nl0  al0  nl1  al1  ns0  as0  ns1  as1
## nl0 1.00 0.98 0.97 0.92 0.87 0.90 0.91 0.92
## al0 0.98 1.00 0.98 0.96 0.83 0.88 0.89 0.93
## nl1 0.97 0.98 1.00 0.98 0.80 0.85 0.88 0.92
## al1 0.92 0.96 0.98 1.00 0.70 0.77 0.80 0.88
## ns0 0.87 0.83 0.80 0.70 1.00 0.99 0.97 0.92
## as0 0.90 0.88 0.85 0.77 0.99 1.00 0.98 0.96
## ns1 0.91 0.89 0.88 0.80 0.97 0.98 1.00 0.97
## as1 0.92 0.93 0.92 0.88 0.92 0.96 0.97 1.00
```

```
t <- cor(d2,method="kendall")
round(t[c(1,5,2,6,3,7,4,8),c(1,5,2,6,3,7,4,8)]*100)/100
```

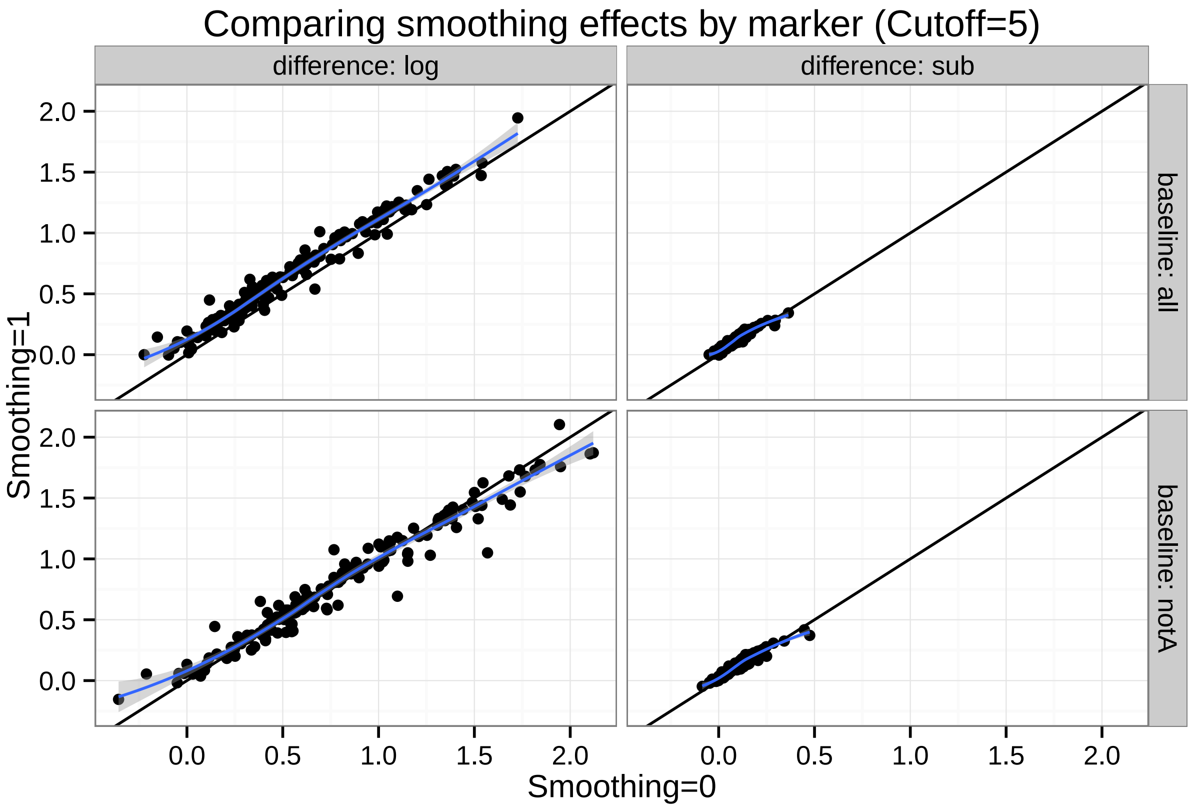
```
##      nl0  al0  nl1  al1  ns0  as0  ns1  as1
## nl0 1.00 0.90 0.86 0.75 0.71 0.74 0.75 0.76
## al0 0.90 1.00 0.89 0.83 0.65 0.72 0.71 0.77
## nl1 0.86 0.89 1.00 0.88 0.62 0.67 0.72 0.77
## al1 0.75 0.83 0.88 1.00 0.52 0.59 0.63 0.72
## ns0 0.71 0.65 0.62 0.52 1.00 0.92 0.86 0.76
## as0 0.74 0.72 0.67 0.59 0.92 1.00 0.90 0.83
## ns1 0.75 0.71 0.72 0.63 0.86 0.90 1.00 0.88
## as1 0.76 0.77 0.77 0.72 0.76 0.83 0.88 1.00
```

That gives us a sense of the general agreement between the alignment measures. Let's look at a little closer at the relationship between the specific values with a by-marker-and-verification scatterplot.

```
d2 <- d %>%
  mutate(valid=((ba+nba)>=cutoff&(bna+nbna)>=cutoff)) %>%
  transmute(valid=valid,vspeak=vspeak,vreply=vreply,category=category,
            notA.log.0=smoothalign(d,0,"logodds"),notA.log.1=smoothalign(d,1,"logodds"),
            notA.sub.0=smoothalign(d,0,"subodds"),notA.sub.1=smoothalign(d,1,"subodds"),
            all.log.0=smoothalign(d,0,"logdnm"),all.log.1=smoothalign(d,1,"logdnm"),
            all.sub.0=smoothalign(d,0,"subdnm"),all.sub.1=smoothalign(d,1,"subdnm")) %>%
  filter(valid==T,is.finite(notA.log.0),is.finite(all.log.0)) %>%
  select(-valid) %>%
  group_by(vspeak,vreply,category) %>%
  summarise_each(funs(mean)) %>%
  gather(atype,alignment,-vspeak,-vreply,-category) %>%
  separate(atype,into=c("baseline","difference","smoothing"),sep="\\.")
```

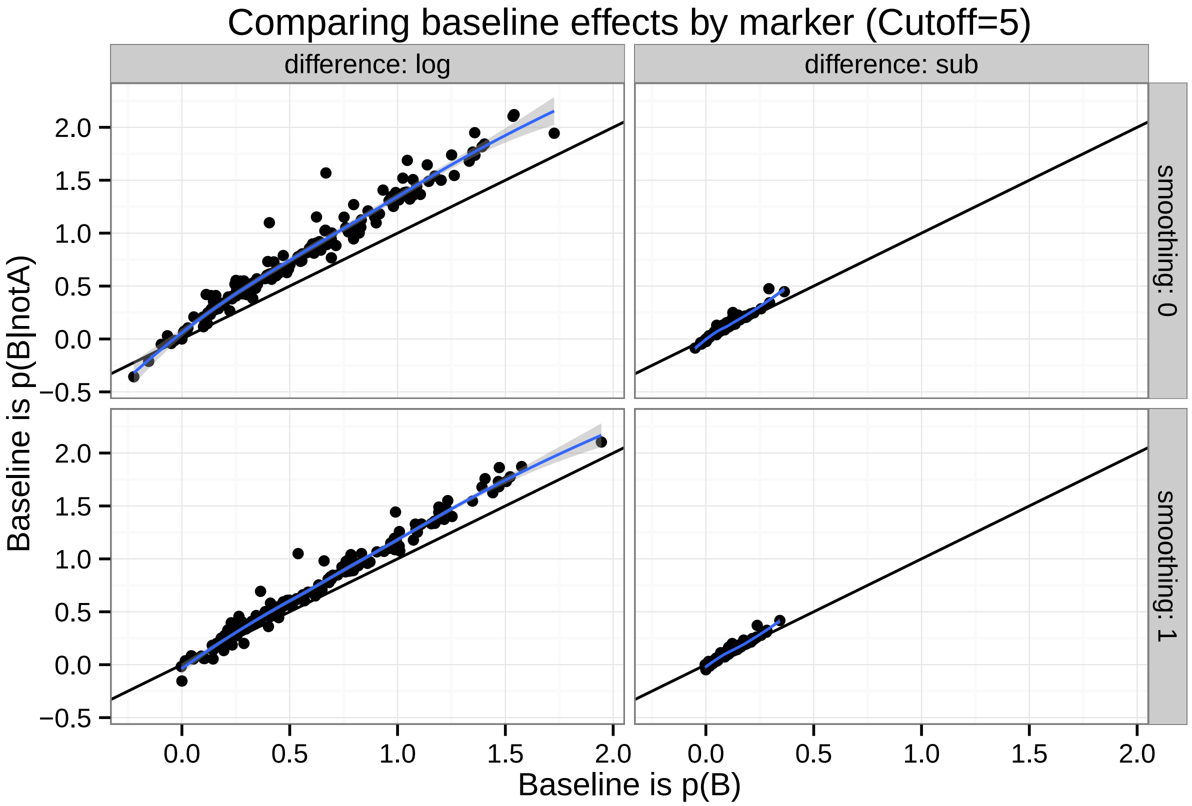
Comparing the effects of smoothing first. Smoothing has little effect in most cases, as excepted from the high correlation. One concern: when the difference is log and the baseline is all instances of B ($\log p(B|A) - \log p(B)$), there is a small constant difference between the smoothing values. We do not use this measure in any of our calculations at present, and shouldn't in the future.

```
d3 <- d2 %>%
  spread(smoothing,alignment,fill=NA) %>%
  transmute(vspeak=vspeak,vreply=vreply,category=category,baseline=baseline,difference=difference,a=`0`
ggplot(d3,aes(y=b,x=a)) + geom_abline(xintercept=0,slope=1) + geom_point() + geom_smooth(method="loess"
  facet_grid(baseline ~ difference,labeller = label_both) +
  labs(title=paste("Comparing smoothing effects by marker (Cutoff=",cutoff,")",sep=''),y="Smoothing=1",
```



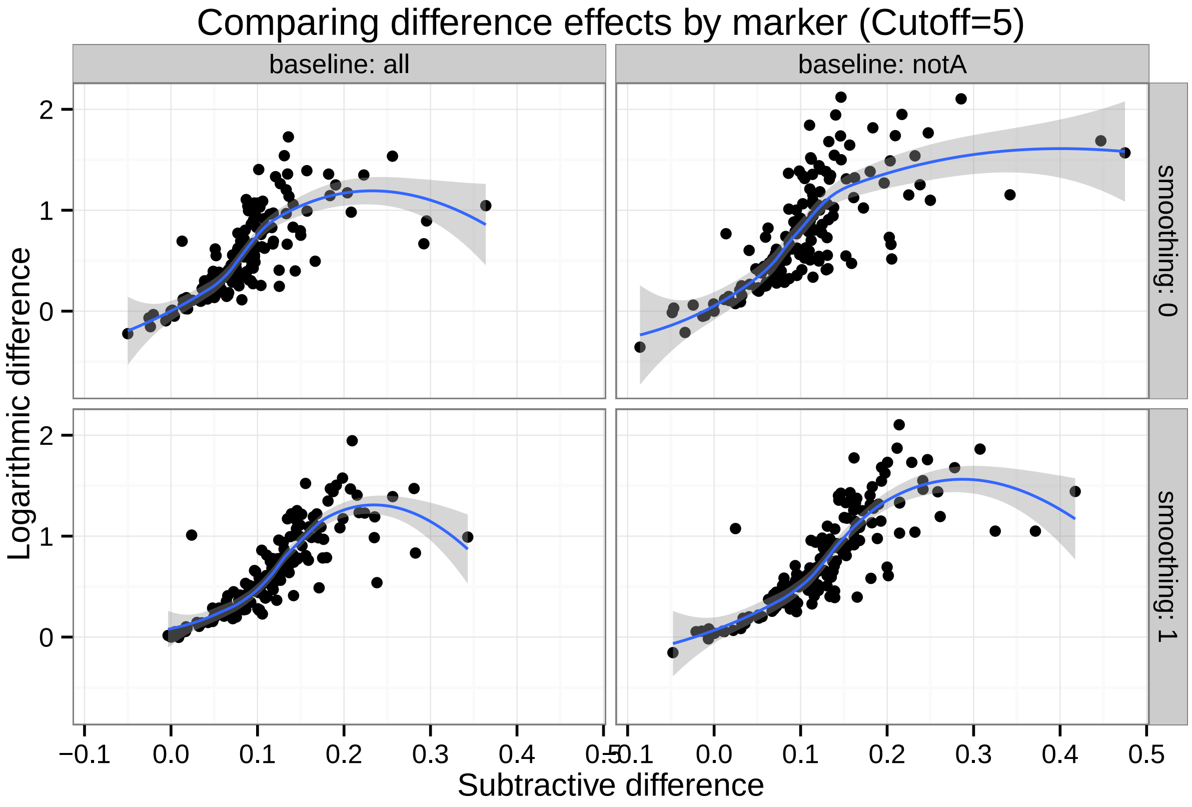
Comparing the effects of the baseline next. This does have a noticeable effect in the log case, as we expect. (In the subtractive measures, baseline choice does not have a substantial effect.) Specifically, measures with baseline $p(B|\neg A)$ have greater absolute differences from zero than measures with a baseline of $p(B)$. Furthermore, the difference in alignment measures is approximately constant, except for alignments near zero. The measures all appear to agree on zero alignment, regardless of baseline.

```
d3 <- d2 %>%
  spread(baseline,alignment,fill=NA)

ggplot(d3,aes(y=notA,x=all)) + geom_abline(xintercept=0,slope=1) + geom_point() + geom_smooth(method="l
  facet_grid(smoothing ~ difference,labeller = label_both) +
  labs(title=paste("Comparing baseline effects by marker (Cutoff=",cutoff,")",sep=''),y="Baseline is p(
```



Lastly, let's compare the effects of the logarithmic vs subtractive differences. This had the largest effect on correlation, and we can see why in these plots; there is a sigmoidal relationship between the logarithmic and subtractive differences. Interestingly, the sigmoid is centered near a logarithmic difference of 0.5 and a subtractive difference of 0.1. It's not immediately clear why this would be the case, but it suggests that relative to the logarithmic calculations, the subtractive calculations magnify small and large alignments while shrinking moderate positive alignments. This may further explain why the logarithmic measure shows power-based alignment that the subtractive measure did not, given the large number of data points in that moderate positive alignment region.

```
d3 <- d2 %>%
  spread(difference,alignment,fill=NA)

ggplot(d3,aes(y=log,x=sub)) + geom_point() + geom_smooth(method="loess") +
  facet_grid(smoothing ~ baseline,labeller = label_both) +
  labs(title=paste("Comparing difference effects by marker (Cutoff=",cutoff,")",sep=''),y="Logarithmic d
```

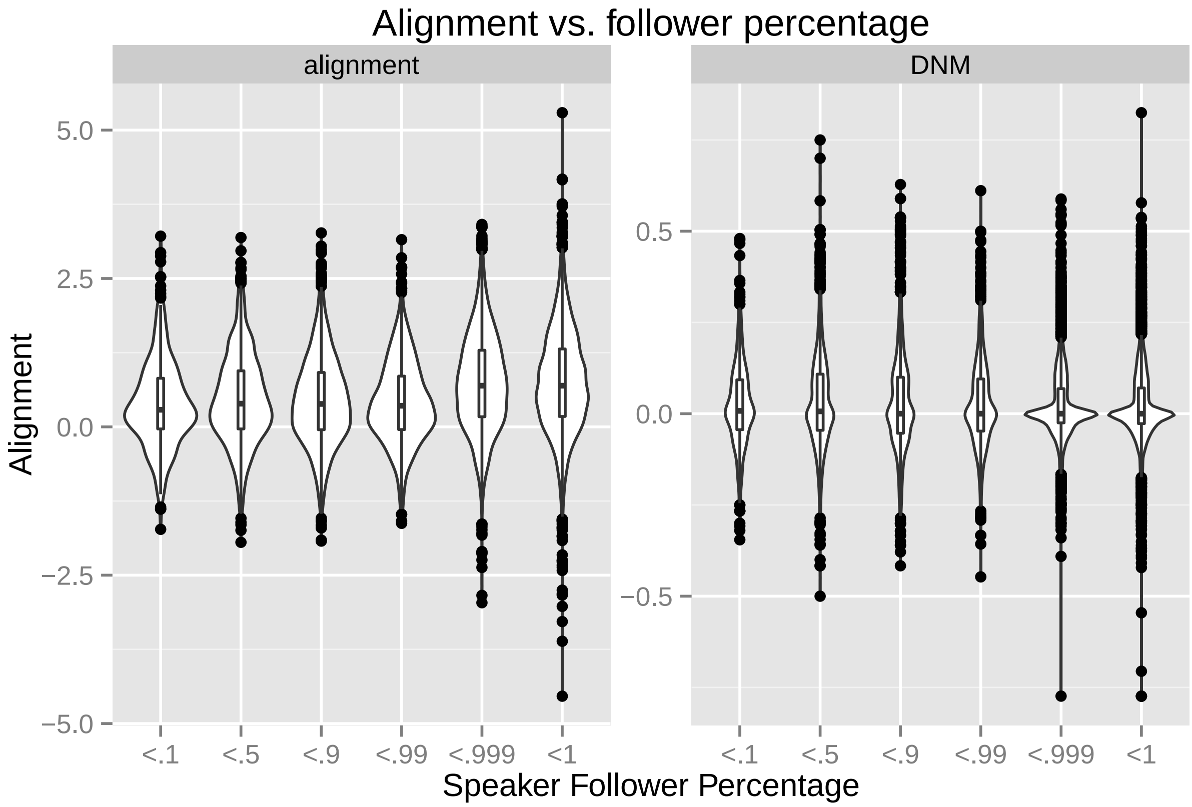


Other measures of power

One last analysis: are our results due to our definition of power as verified status? DNM's work on Twitter didn't find alignment to power, but they also didn't have verification as an available variable (as verification was instituted after their paper was written, I believe.). There are also many different kinds of power, and maybe the type of power that verification represents (usually real-world, Twitter-independent significance) behaves differently from Twitter-internal power proxies like follower ratio.

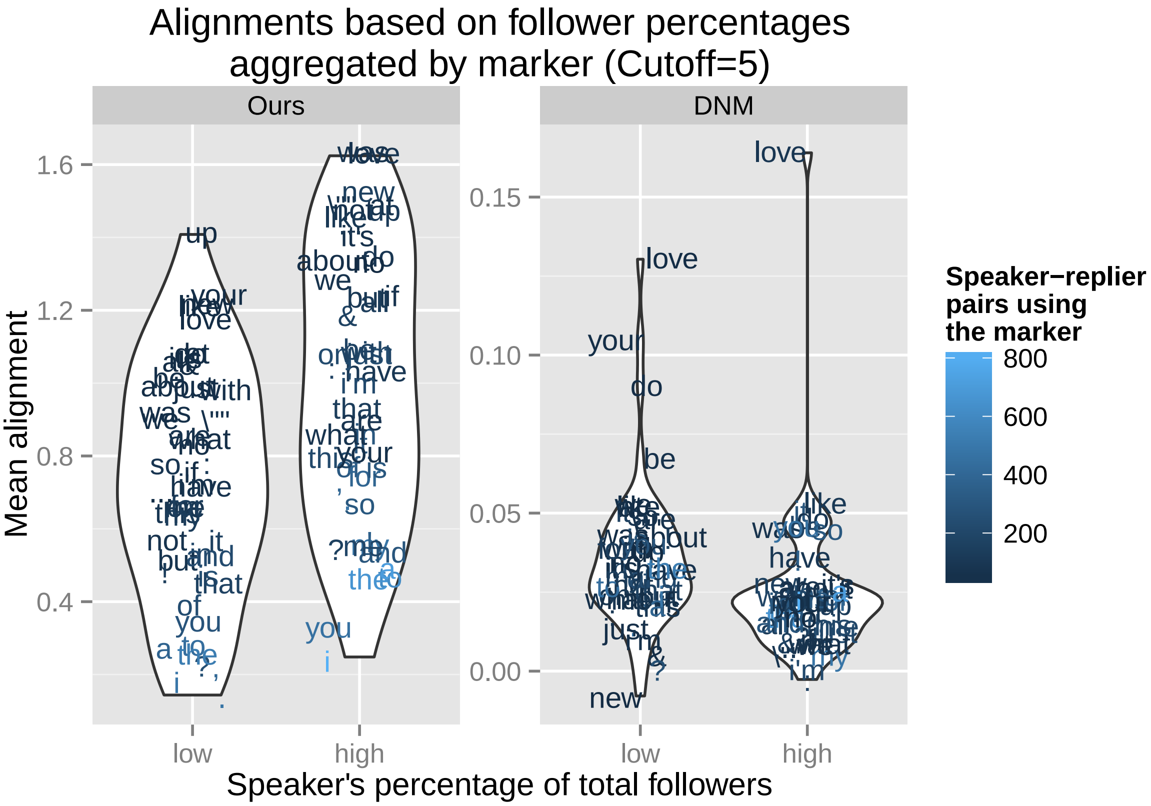
We test on speaker follower percentage: the number of followers that the speaker has divided by the sum of the number of followers of the replier and the number of followers of the speaker. High values indicate the speaker is followed by more people than the replier, with 0.5 indicating an even split. We expect high follower percentages to correlate with Twitter power. First, violin plots of the binned speaker follower percentage on alignment:

```
d <- df[,list(ba=ba,nba=nba,bna=bna,nbna=nbna,
              vspeak=verifiedSpeaker,vreply=verifiedReplier,
              sid=speakerId,rid=replierId,category=category,
              pyalign=alignment,reciprocity=reciprocity,dnmalignment=dnmalignment,
              percentDiff=percentDiff),]


d2 <- d %>%
  filter((ba+nba)>=cutoff,(bna+nbna)>=cutoff) %>%
  group_by(sid,rid) %>%
  summarize(convs=n(), alignment=pyalign, DNM=dnmalignment, percentDiff=percentDiff) %>%
  gather(alignmentType,alignmentValue,c(alignment, DNM))

d2$followerBins <- cut(d2$percentDiff, breaks=c(0,0.1,0.5,0.9, 0.99, 0.999, 1), labels=c("<.1","<.5", "

ggplot(d2,aes(x=followerBins,y=alignmentValue)) + geom_violin() + labs(title="Alignment vs. follower pe
```

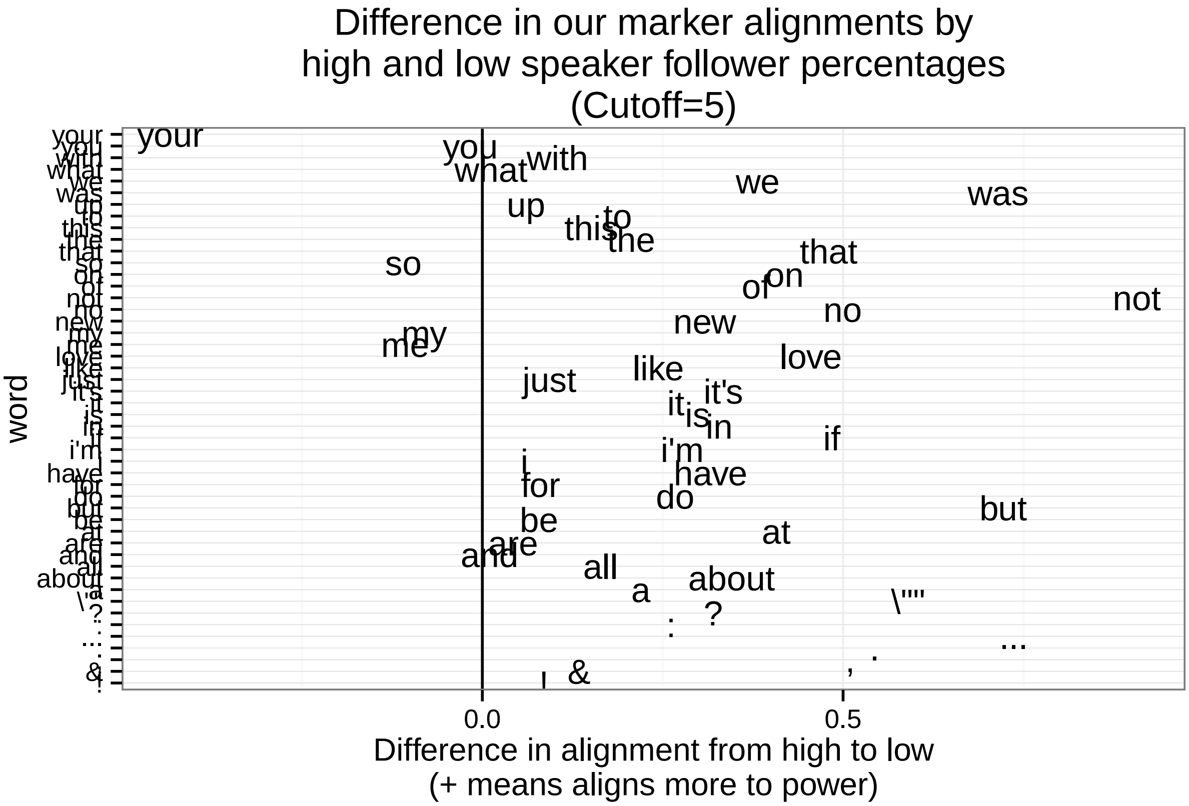


For both alignment measures, there is a categorical change at the .99 follower percentage mark (the first four vs. last two bins). We're going to binarize into two bins with the cutoff at .99, equivalent to the speaker having nearly 100 times as many followers. As before, we'll do by-marker aggregation and t-tests on the alignment to power on both meaasures of alignment.

```
d2 <- d %>%
  filter((ba+nba)>=cutoff,(bna+nbna)>=cutoff) %>%
  mutate(fb=cut(percentDiff, breaks=c(0,0.99,1), labels=c("low", "high"))) %>%
  group_by(category,fb) %>%
  #filter(convs<50) %>%
  summarize(convs=n(),Ours=mean(pyalign),DNM=mean(dnmalignment)) %>%
  gather(alignment,mean,Ours,DNM)

ggplot(d2,aes(x=fb,y=mean,color=convs,label=category)) + geom_violin() + geom_text(position=position_ji
```



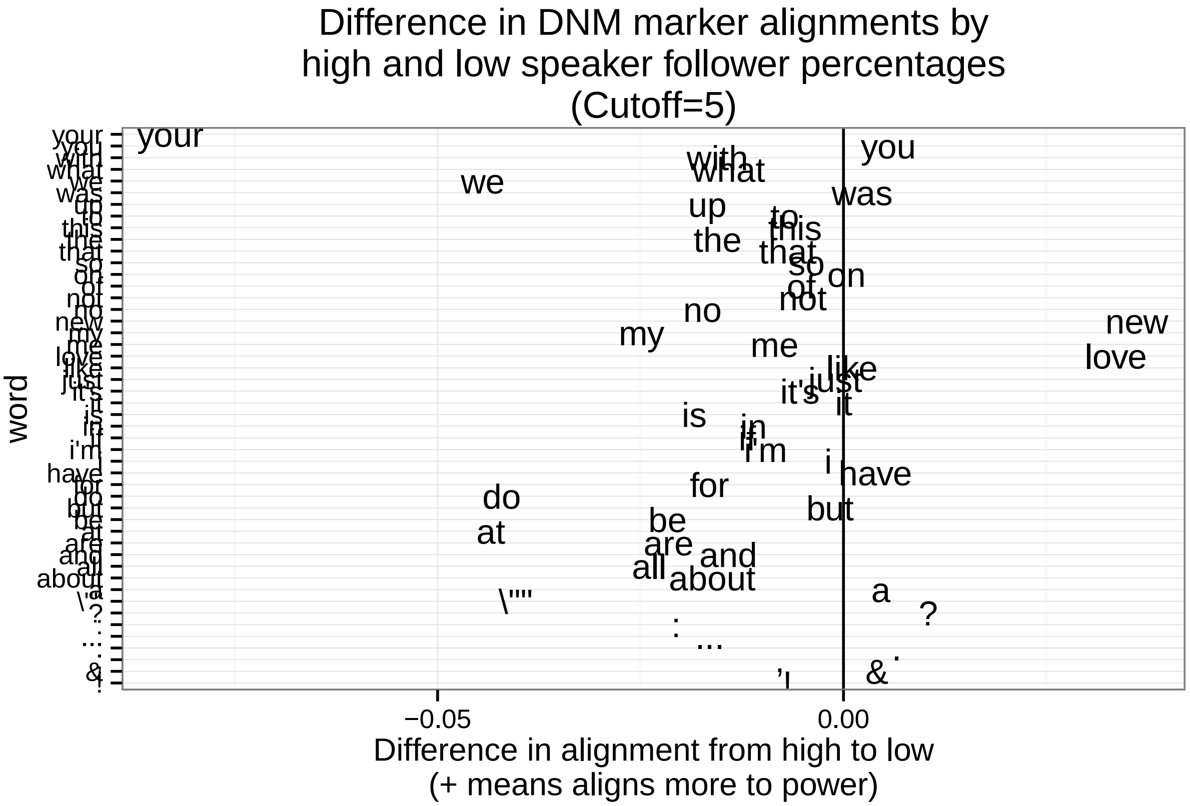
```
d3 <- d2 %>%
  select(-convs) %>%
  filter(alignment=='Ours') %>%
  group_by(category,alignment) %>%
  spread(fb,mean,fill=NA)

ggplot(d3,aes(x=high-low,y=category)) + geom_vline(xintercept=0) + geom_text(aes(label=category),size=4
```



```
t.test(d3$high,d3$low)
```

```
##
##  Welch Two Sample t-test
##
## data:  d3$high and d3$low
## t = 3.7013, df = 91.723, p-value = 0.0003657
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1229921 0.4078396
## sample estimates:
## mean of x mean of y
## 0.9944370 0.7290211
```

```
d3 <- d2 %>%
  select(-convs) %>%
  filter(alignment=='DNM') %>%
  group_by(category,alignment) %>%
  spread(fb,mean,fill=NA)

ggplot(d3,aes(x=high-low,y=category)) + geom_vline(xintercept=0) + geom_text(aes(label=category),size=4
```



```
t.test(d3$high,d3$low)
```

```
##
##  Welch Two Sample t-test
##
## data:  d3$high and d3$low
## t = -2.1703, df = 93.991, p-value = 0.0325
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0207035637 -0.0009205377
## sample estimates:
##  mean of x  mean of y
## 0.02507520 0.03588726
```

Again, we see a significant positive alignment to power in our measure and a significant but weak alignment against power in DNM's. Of course, the speaker follower percentage is highly correlated with verified status, so this is not a big surprise, but it's good to know that the measures of internal and external power agree.