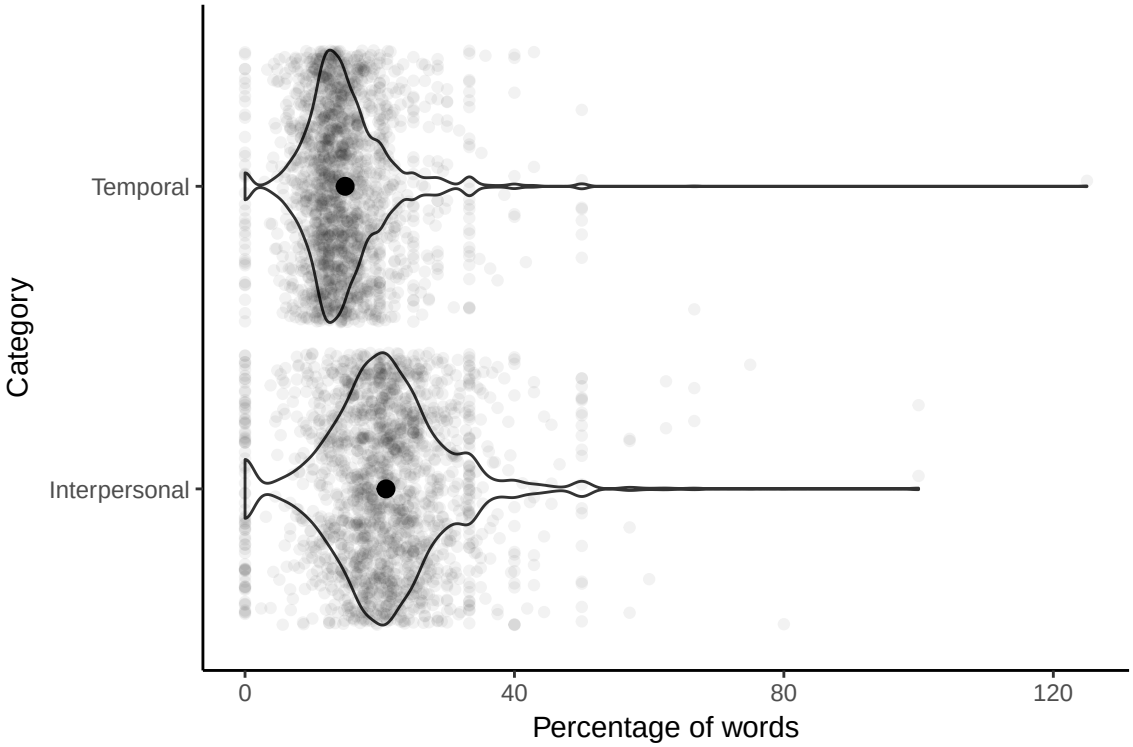
NLP vignette

*Ian Hussey**

29 October 2017

Interpersonal vs temporal

```
# dependencies

library(NLP)
library(tidyverse)
library(effsize)

# process data

tidy_data <- tidy_parcels(data = reddit_suicide_data)

categorized_data <- categorize_parcels(data = tidy_data, dictionary = relations)

# plot

subset <- categorized_data %>%
  filter(category %in% c("Interpersonal", "Temporal"))

plot_percentages(subset)
```



*Ghent University. Email: ian.hussey@ugent.be

```r
# analyse

t.test(percent ~ category,
       data = subset,
       paired = FALSE)
```

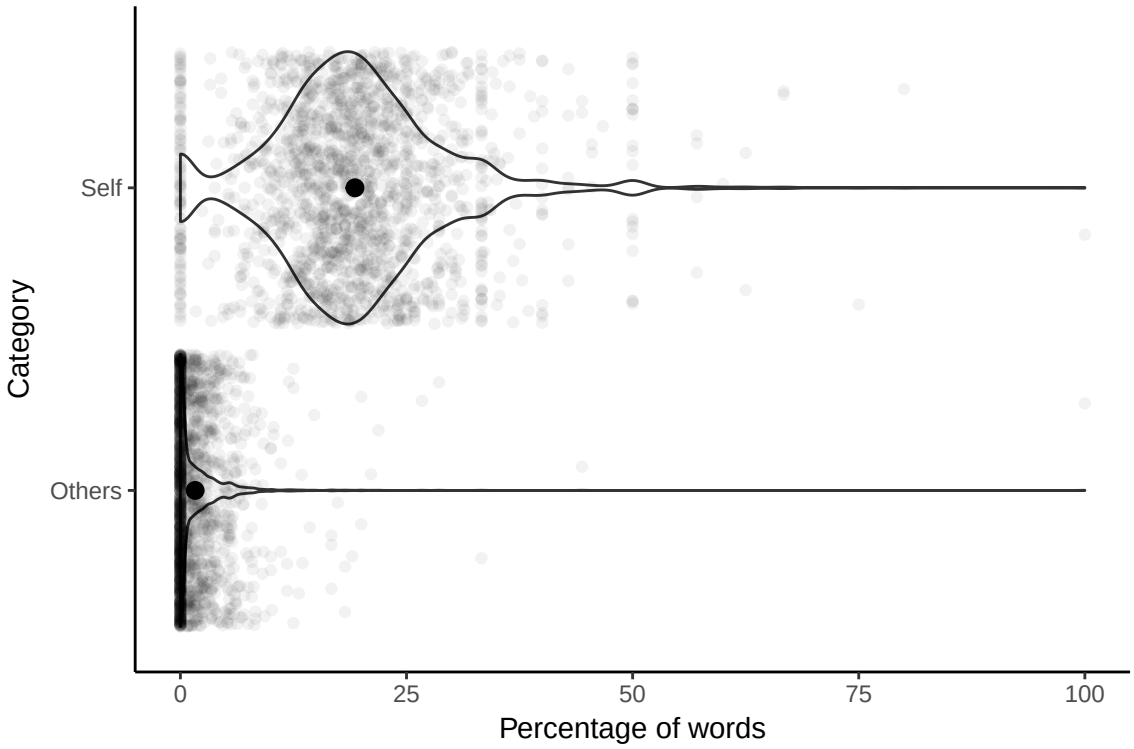
```
##
##  Welch Two Sample t-test
##
## data:  percent by category
## t = 21.219, df = 3228.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  5.502421 6.622818
## sample estimates:
## mean in group Interpersonal      mean in group Temporal
##                    20.92940                    14.86678
```

```r
cohen.d(formula = percent ~ category,
        data = subset,
        paired = FALSE)
```

```
##
## Cohen's d
##
## d estimate: 0.7114699 (medium)
## 95 percent confidence interval:
##       inf       sup
## 0.6436830 0.7792567
```

Interpersonal self vs others

```r
# plot

subset <- categorized_data %>%
  filter(category %in% c("Self", "Others"))

plot_percentages(data = subset)
```




```
# analyse

t.test(percent ~ category,
       data = subset,
       paired = FALSE)
```

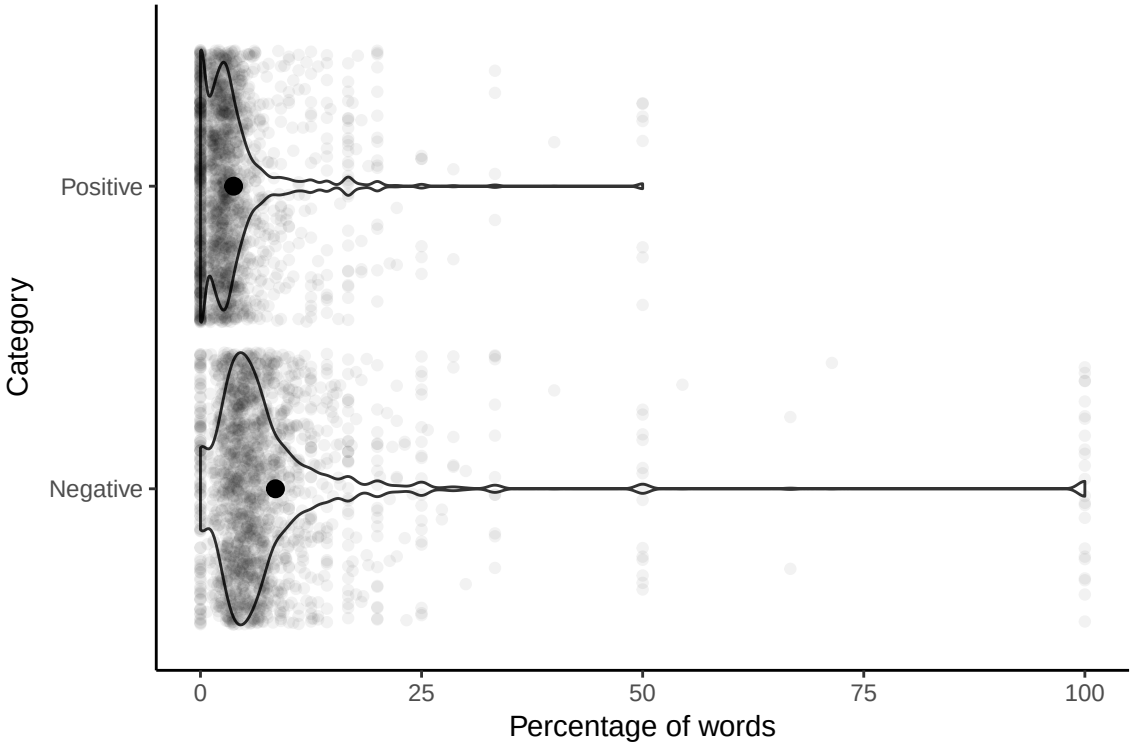
```
##
##  Welch Two Sample t-test
##
## data:  percent by category
## t = -70.651, df = 2309.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.16078 -17.17986
## sample estimates:
## mean in group Others   mean in group Self
##             1.629399            19.299719
```

```
cohen.d(formula = percent ~ category,
        data = subset,
        paired = FALSE)
```

```
##
## Cohen's d
##
## d estimate: 2.368883 (large)
## 95 percent confidence interval:
##      inf      sup
## 2.283133 2.454633
```

Positive vs negative valence/sentiment

```
# process data

tidy_data <- tidy_parcels(data = reddit_suicide_data)

categorized_data <- categorize_parcels(data = tidy_data, dictionary = valence)

# plot

plot_percentages(categorized_data)
```




```
# analyse

t.test(percent ~ category,
       data = categorized_data,
       paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  percent by category
## t = 14.126, df = 2349.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4.075130 5.388888
## sample estimates:
## mean in group Negative mean in group Positive
##               8.477835               3.745826
```

```
cohen.d(formula = percent ~ category,
```

```
        data = categorized_data,
        paired = FALSE)
```

```
##
## Cohen's d
##
## d estimate: 0.4793437 (small)
## 95 percent confidence interval:
##       inf       sup
## 0.4118656 0.5468219
```