Project 2

Xiaonan Hu, Lee Stovall, Liqi Zhu

Part 1 ASJP Dataset

1. Introduction

The purpose of Part 1 is to find the overall status for those languages we choose based on the population that use the language. (http://www.vistawide.com/languages/language_families_statistics1.htm)

Language Family	**Speaker Percentage**
Indo-European	44.78%
Sino-Tibetan	22.28%
Niger-Congo	6.26%
Afro-Asiatic	5.93%
Austronesian	5.45%
Dravidian	3.87%
Japanese	2.16%
Austro-Asiatic	1.77%
Tai-Kadai	1.37%
Total	**93.87%**

We managed to download the dataset for each language family from ASJP database.

2. Method & Processing Data

Our thought is to compare the Distance data within each language family. One dataset is with the numbers and

the other is not.

```
#### Cygwin ####
egrep -v '\sone|\stwo' < Indo-European.txt > IEwonum.txt
egrep -v '\sone|\stwo' < Austro-Asiatic.txt > AuAwonum.txt
egrep -v '\sone|\stwo' < Austronesian.txt > Awonum.txt
egrep -v '\sone|\stwo' < Dravidian.txt > Dwounum.txt
egrep -v '\sone|\stwo' < Japanese.txt > Jwonum.txt
egrep -v '\sone|\stwo' < Niger-Congo.txt > NCwonum.txt
egrep -v '\sone|\stwo' < Sino-Tibetan.txt > STwnum.txt
egrep -v '\sone|\stwo' < Tai-Kadai.txt > TKwonum.txt
egrep -v '\sone|\stwo' < Afro-Asiatic.txt > AAwonum.txt
# Exclude numbers in dataset > Datasets without numbers.
```

With Programs for calculating ASJP distance matrices (http://asjp.clld.org/software). We managed to get the distance matrix for each language family.

```
#### Power Shell ####
 asjp62 < listss17.txt > output1.txt
 asjp62 < Indo-European.txt > IEall.txt
 asjp62 < Austro-Asiatic.txt > AuAall.txt
 asjp62 < Austronesian.txt > Aall.txt
 asjp62 < Dravidian.txt > Dall.txt
 asjp62 < Japanese.txt > Jall.txt
 asjp62 < Niger-Congo.txt > NCall.txt
 asjp62 < Tai-Kadai.txt > TKall.txt
 asjp62 < Sino-Tibetan.txt > STall.txt
 asjp62 < Afro-Asiatic.txt > AAall.txt
# Overall view of whole dataset and different language families with n
umbers.
 asjp62 < IEwonum.txt > IEall_won.txt
 asjp62 < AuAwonum.txt > AuAall_won.txt
 asjp62 < Awonum.txt > Aall_won.txt
 asjp62 < Dwounum.txt > Dall_won.txt
 asjp62 < Jwonum.txt > Jall_won.txt
 asjp62 < NCwonum.txt > NCall_won.txt
 asjp62 < TKwonum.txt > TKall_won.txt
 asjp62 < STwnum.txt > STall_won.txt
 asjp62 < AAwonum.txt > AAall_won.txt
# Overall view of whole dataset and different language families withou
t numbers.
```

```
cat IEall*.txt > IE.txt
```

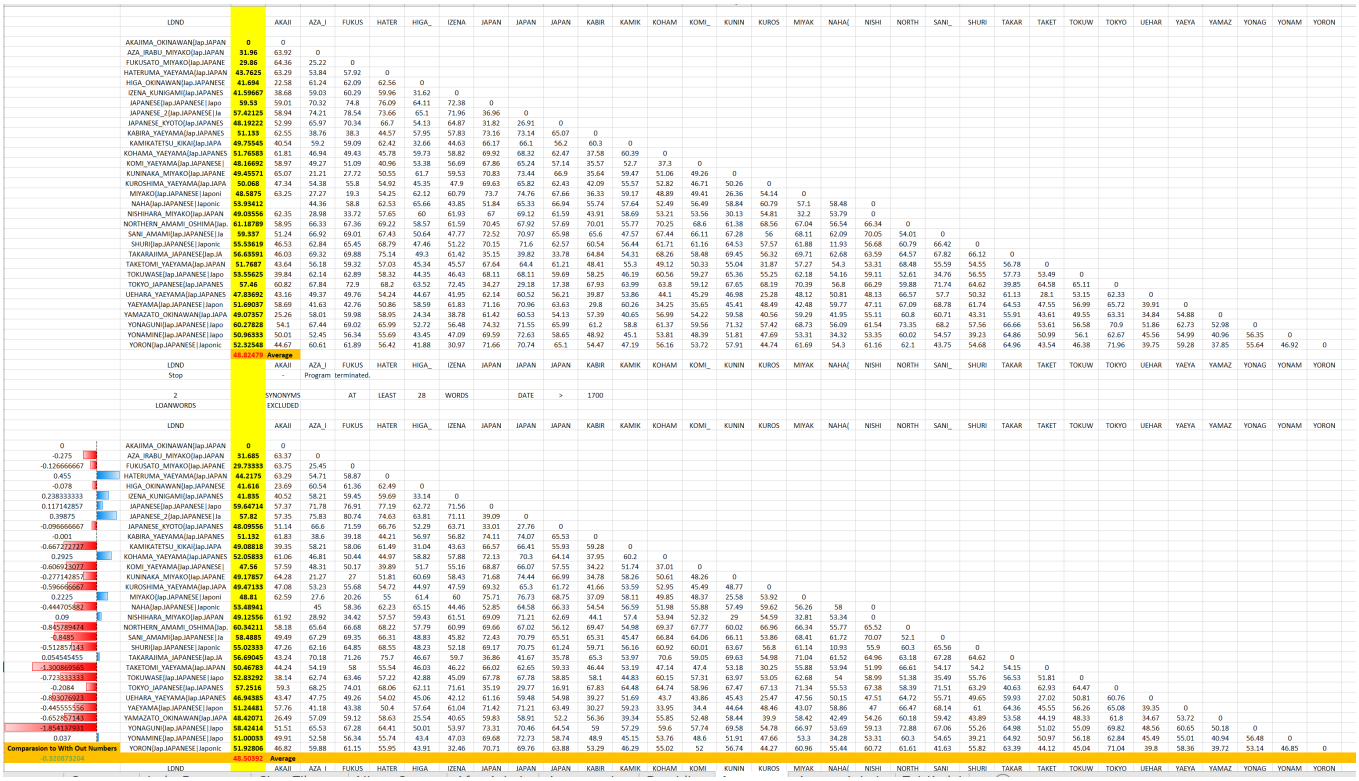
```
cat AuAall*.txt > AuA.txt
cat Aall*.txt > A.txt
cat Dall*.txt > D.txt
cat Jall*.txt > J.txt
cat NCall*.txt > NC.txt
cat TKall*.txt > TK.txt
cat STall*.txt > ST.txt
cat AAall*.txt > AA.txt

sed 's/ \+/,/g' IE.txt > IE.csv
sed 's/ \+/,/g' AuA.txt > AuA.csv
sed 's/ \+/,/g' A.txt > A.csv
sed 's/ \+/,/g' D.txt > D.csv
sed 's/ \+/,/g' J.txt > J.csv
sed 's/ \+/,/g' NC.txt > NC.csv
sed 's/ \+/,/g' TK.txt > TK.csv
sed 's/ \+/,/g' ST.txt > ST.csv
sed 's/ \+/,/g' AA.txt > AA.csv
# Arrange data into CSV file
```

With the combination and calculation of the .CSV files, we are able to compare the ASJP Distance of each language famil with and without numbers.

3. Results

Example of Japanese



As showed above, values are mesured ASJP Distance data. We calculated the average standard for each family language family and then compare the Distance matrix with numbers and the one without.
Then we repeat the process for the other language families.
The results of all language families are followed:

	Speaker Percentage	ASJP Distance With Numbers	ASJP Distance Without Numbers	Difference
Indo-European	44.78%	85.56	85.84	0.280495
Sino-Tibetan	22.28%	88.27	88.46	0.191502
Niger-Congo	6.26%	91.43	91.32	-0.1074
Afro-Asiatic	5.93%	91.32	91.16	-0.15255
Austronesian	5.45%	85.31	85.40	0.089109
Dravidian	3.87%	85.31	57.28	-28.0276
Japanese	2.16%	48.82	48.50	-0.32087
Austro-Asiatic	1.77%	88.80	81.34	-7.4635
Tai-Kadai	1.37%	75.99	75.61	-0.37409
Average	10.43%	82.31	78.32	-3.99
Total	93.87%	740.80	704.92	-35.8849

4.Conclusion

As shown above, the first thing we discovered is that language families have different features. Japanese was obviously the language family containing the most similar languages since ASJP Distance is the smallest. While all the other language families show different level of variance.
About the influence of numbers, Dravidian's Distance drops significantly, which means the number changes a lot in Dracvidian Languages. Thus to say all the language famil with the positive numbers of Difference means number changes less than other words. And Negative number indicates number changes more.

As a conclusion, in Indo-European, Sina-Tibetan, Austronesian language families, numbers changes less. While in other language families numbers may change more than average.