# LING 410X: Language as Data

## Semester: Spring '18

Instructor: Sowmya Vajjala

Iowa State University, USA

10 Apr 2018

# Class Outline

- Assignment 5 and project initial submission - comments
- Discussion on Stylo
- Clustering - overview
- One example application of stylo for non-literature analysis

# Assignment 5 - some comments

- How long did it take for you to run the program as is? $\Rightarrow$ 15 min to 1 hour

# Assignment 5 - some comments

- How long did it take for you to run the program as is? $\Rightarrow$ 15 min to 1 hour
- With some customized pre-processing, you may see more meaningful topic clusters than default.
  - adding custom stopwords, increasing to 8-10 topics seem to make more sense.
  - note: stemming is less interpretable for humans.

# Assignment 5 - some comments

- How long did it take for you to run the program as is? $\Rightarrow$ 15 min to 1 hour
- With some customized pre-processing, you may see more meaningful topic clusters than default.
  - adding custom stopwords, increasing to 8-10 topics seem to make more sense.
  - note: stemming is less interpretable for humans.
- There were some strange characters in the topics, most frequent words. One possible solution: reading the text as utf8?

# Some other observations

- ngram models: May result in better modeling, but will increase vocabulary size and training time.

# Some other observations

- ngram models: May result in better modeling, but will increase vocabulary size and training time.
- topic models vs word clouds: topic models give more information than word clouds

# Some other observations

- ngram models: May result in better modeling, but will increase vocabulary size and training time.
- topic models vs word clouds: topic models give more information than word clouds

# Where are topic models useful?

- Creating topic distribution for research articles, political speeches, tweets etc
- Comparing topics covered in different newspapers from different regions of US.
- To detect “hottest” topics in a certain time frame
- Understanding system logs to know about bug patterns
- Knowing customer reactions to a product, Market demands over time etc.
- Find themes inside large collections of texts
- Using with non-text data (music recommendation, image classification etc)

# Your project ideas

- In general, ideas are good, and come from your own discplines.
- Some people wrote in detail (with alternative plans) and some wrote like 1 paragraph.
- More details is good for you

# Clarification regarding dead-week

Final exam for this course = Your submission of final project reports. Project presentations do not count as final exam.

"Final exams may not be given at a time other than that for which the exam is scheduled by the registrar. An instructor may not give a final exam prior to final exam week nor change the time of offering of the final examination as it appears in the final exam schedule. Permission to change the time for which an exam is scheduled may be given only by the dean of the college. If the instructor elects not to give a final exam in a course of two or more credits, the class is required to meet at the scheduled final exam period for other educational activity such as a review of the course or feedback on previous exams." Note that additional policies and scheduling processes apply to final exams that are given in the online testing center."

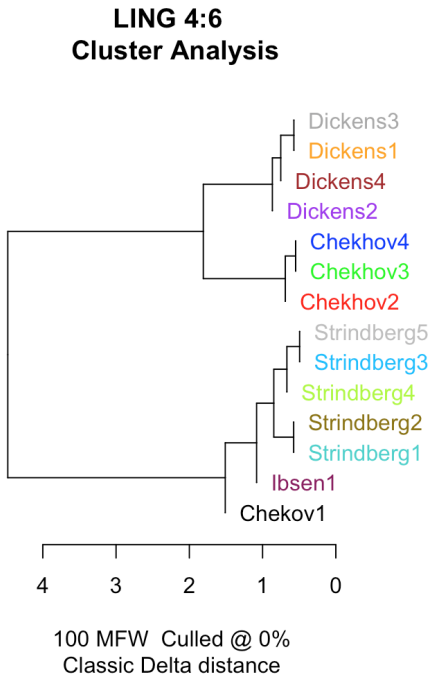
# Working with stylo - some observations

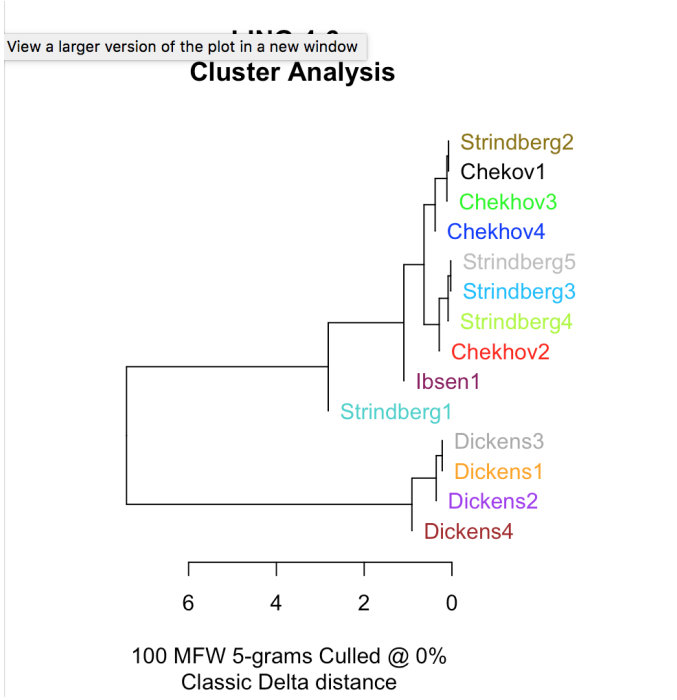
# Using stylo for comparing writing style - 1

- Using bigrams showed based results - Dickens and Strindberg had one cluster each. Chekhov 2–4 was one cluster and Ibsen1 and Chekhov1 came as another cluster.
- Setting ngram as 2 or changing MFW setting did not show any change.
- Classic delta distance perform better than Euclidean distance and cosine distance for this text data.

# Using stylo for comparing writing style - 2

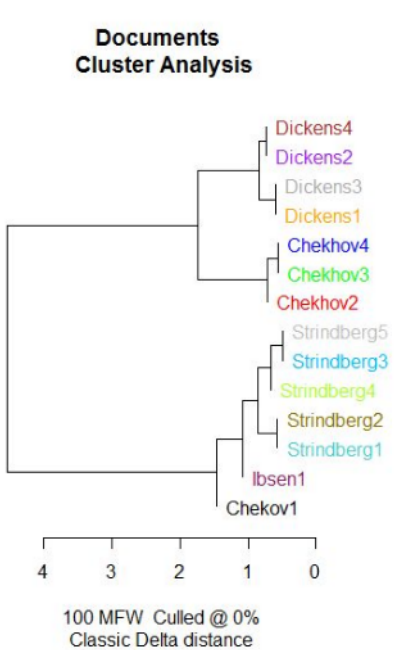
ngrams 1 vs ngrams 5 - very different clusters.

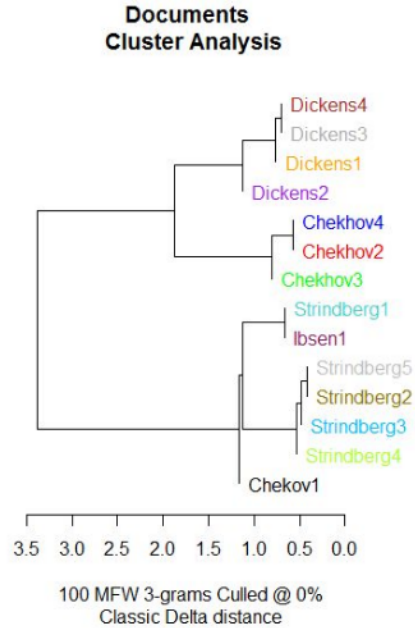


# Using stylo for comparing writing style - 3

ngrams 1 vs ngrams 3.




choosing 1000 MFW showed best cluster.

# Using Stylo for comparing writing style - 4

- Cluster analysis changes with different settings (MFW, ngrams etc)
- Bootstrap consensus trees, another cluster visualization method, is perhaps a better way.

# Clustering - overview

# Process

- Let us say I have a collection of texts.
- I choose to represent the collection of texts as a feature vector (say, taking 500 most frequent words in the corpus as my features).
- Now, if i can represent each document like this, I should be able to compare two documents based on how far away their feature vectors are from each other.
- If I start with considering each document as its own cluster, and step by step, group documents that are "close" to each other, you end up with one large cluster for entire data

(Consider a space of only two words as features - to visualize the notion of document similarity)

# The Measure of Closeness

- We need some way of measuring the closeness or similarity, if we want to group texts together. One such measure Burrows' Delta (Burrows, 2002).
- What is that?
  - Let us assume a collection of texts, and we want to study their stylistic differences, using $n$ most frequent words in the collection.
  - Now, each Document $D$ can be represented by a vector: ($f_1$(D), $f_2$(D) ... $f_n$(D)) of the frequency of occurrence of all frequent words.
  - If we scale the feature vector, to normalize the values, each $f_i$(D) for a document D becomes equal to a new value $z_i$(D) which is defined as: ($f_i$(D) - $\mu$)/$\sigma$
    (where $\mu$ is the mean of the distribution of $f_i$ across all documents in the corpus)
  - So, the new representation for a document is: D = ($z_1$(D), $z_2$(D) ... $z_n$(D))

# Burrows Delta

If we have two documents D1 and D2, represented as ($z_1$(D1), $z_2$(D1) ... $z_n$(D1)) and ($z_1$(D2), $z_2$(D2) ... $z_n$(D2)) respectively, Burrows Delta is given by: $\Delta_B = |z(D1) - z(D2)|$ (i.e., this is the "Manhattan Distance" between these two normalized feature vectors).

# Different formulae

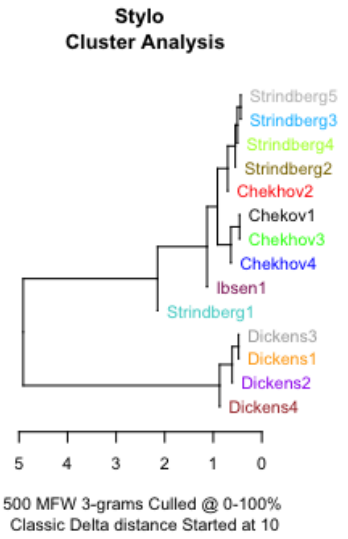
- Different variations of Delta exist, based on how the normalized representations for documents are formed, and what measure of distance is used on the normalized representations.
- Examples: Argamon's Delta (square of the $\Delta_B$ value in the previous slide), Eder's Delta etc.
- Calculating distances without the normalization step using Manhattan distance, Euclidean distance, Canberra distance etc.

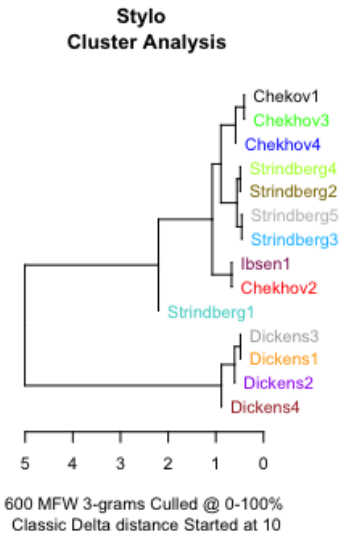
# Problem with clusters

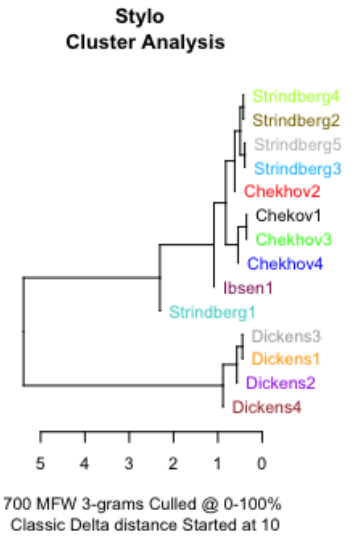
- sensitive to the initial choices (e.g., distance measure, number of features)
- any change for those settings may result in drastically different clusters.
- ... so, inconsistent.

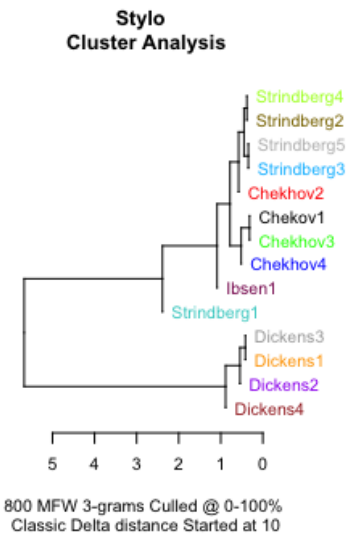
# Overcoming this problem with clusters

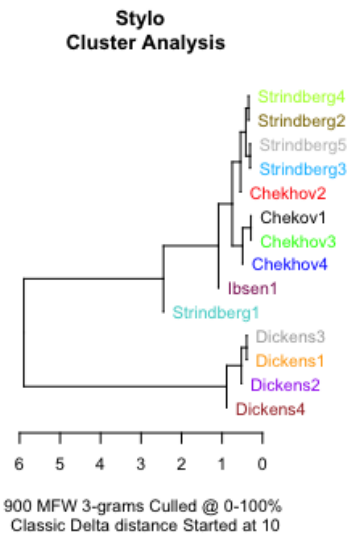
Build multiple clusters and compare.

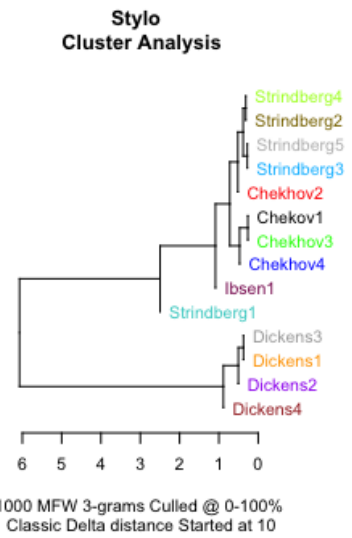


# Bootstrap Consensus Trees (BCT)

- Idea:
  1. Build multiple hierarchical clusters of the corpus following a bootstrapping procedure, using a sample of data each time.
  2. Eventually, keep only those groupings that are seen at least X% of times in the process.
- So, if I choose X as 50%, it means that in my final grouping, I choose only those groups whose elements are seen at least 50% of the times in the dendrograms I constructed.
- This is an alternative to building a single hierarchical cluster, as this decides its grouping after looking multiple clusters.

# BCT in action

https://sites.google.com/site/computationalstylistics/projects/lee_vs_capote

# Rolling Delta

- Burrow's delta method, extended to visualize stylistic shifts within a single text.

# Rolling Delta

- Burrow's delta method, extended to visualize stylistic shifts within a single text.
- Idea:
  - Break up each text into equally sized, partially overlapping samples using two parameters - window size, step size.
  - If we specify a window size? of 5,000 and a step size of 100, the first sample of a text contains1-5000 words of the text, second sample contains: 101-5101 words, and so on.
  - Let us say I want to take only the n most frequent words as my "features" for doing this.

# Rolling Delta - continued

- For each "reference text", we compute a "centroid" vector of size n where each element represents the mean frequency for a word in all those samples of a text.
- We also keep track of the standard deviations for these n words in this reference text.
- When we get a new test document, we split that into samples as before, and use the following formula to get the "delta" of each sample:
  $\Delta(C, W) = \Sigma_{i=1}^{n} \frac{1}{\sigma_i(C)} |\mu_i(C) - f_i(W)|$
- This is then plotted along with the reference text.

# Rolling Delta - student projects

https://sites.google.com/site/computationalstylistics/projects/testing-rolling-delta

# Other Functions in Stylo

- text classification
- principal component analysis
- ...

# Beyond stylometry - one application with stylo

# in what other scenario can Stylo visualizations be useful?

## my project from 2016

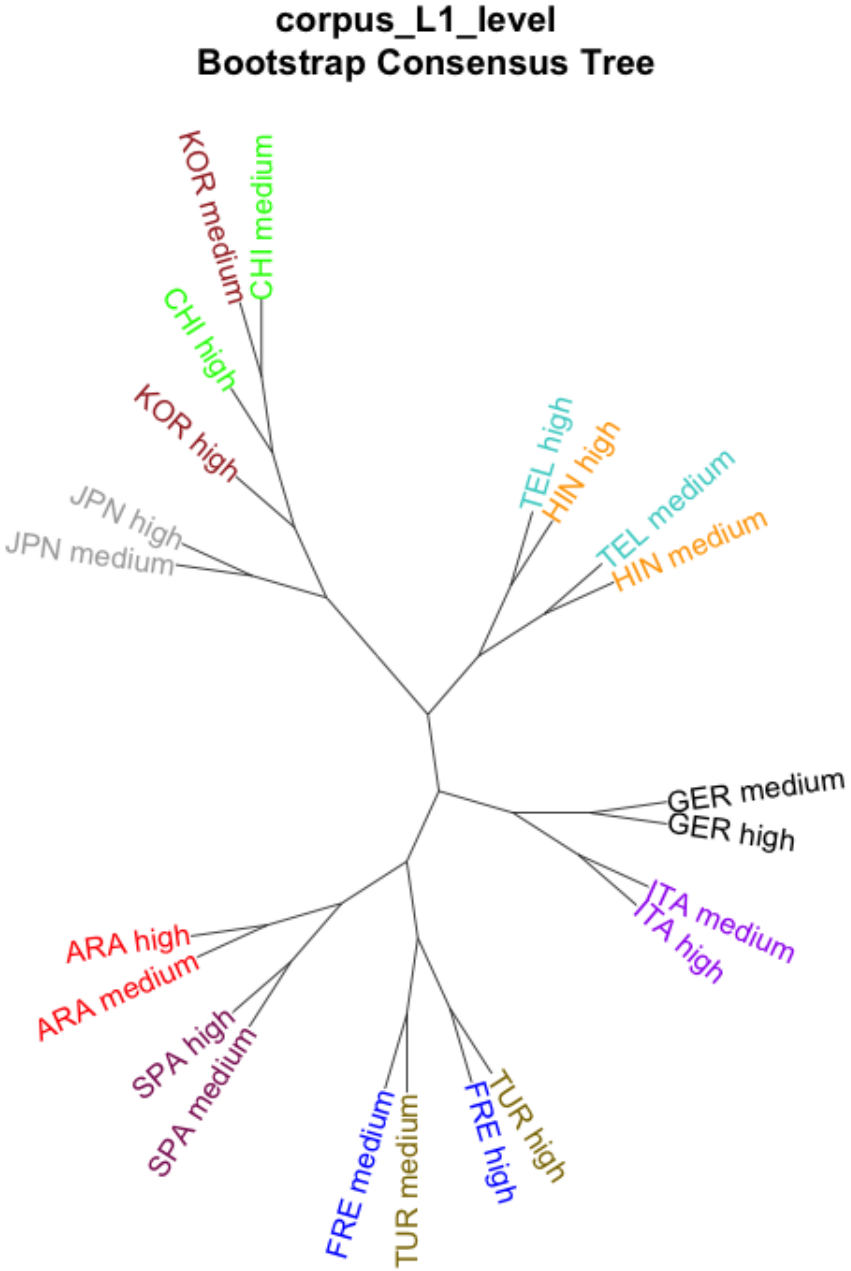
- Data we had:
  - TOEFL essays written by people with different native language backgrounds (L1) in English (L2).
  - Size: 8830 essays in total, 11 L1s, 2 proficiency levels (medium, high)
  - 11 L1s: CHI, JPN, KOR, TEL, HIN, ARA, TUR, GER, ITA, SPA, FRE.
- Question we explored: ow far is visualization useful in forming hypotheses about the relationship between L1 and L2 proficiency?

# Methods: used Stylo

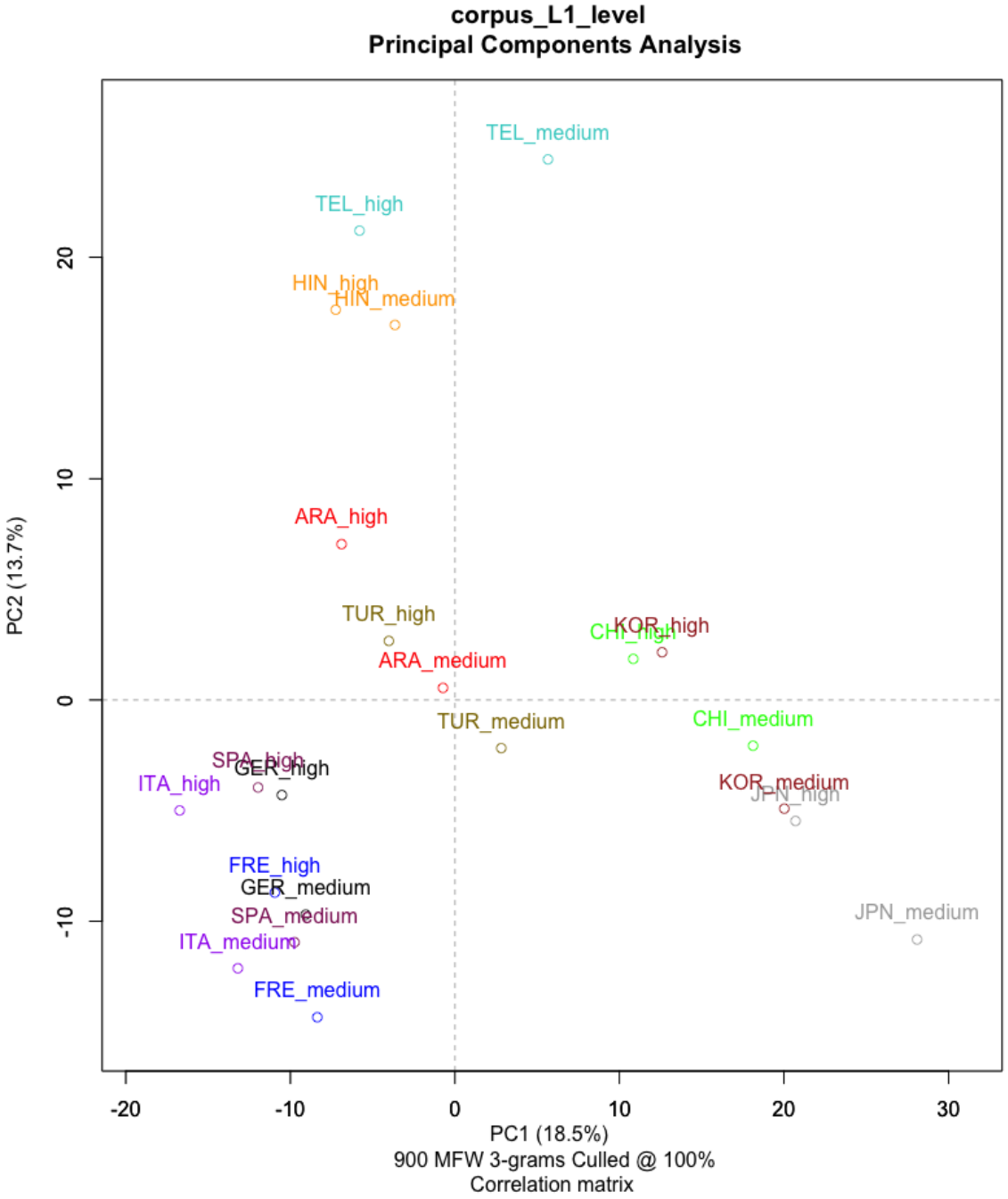
- Word, Character and POS n-grams (uni, bi, tri grams)
- Experimented with different frequencies (100 Most frequent n-grams to 1000 most frequent n-grams)
- Experimented with culling (how many n-grams in the final ones considered appear in how many text categories)

Pre-processing: only lowercasing and tokenizing. POS tagging was done with Stanford Tagger.

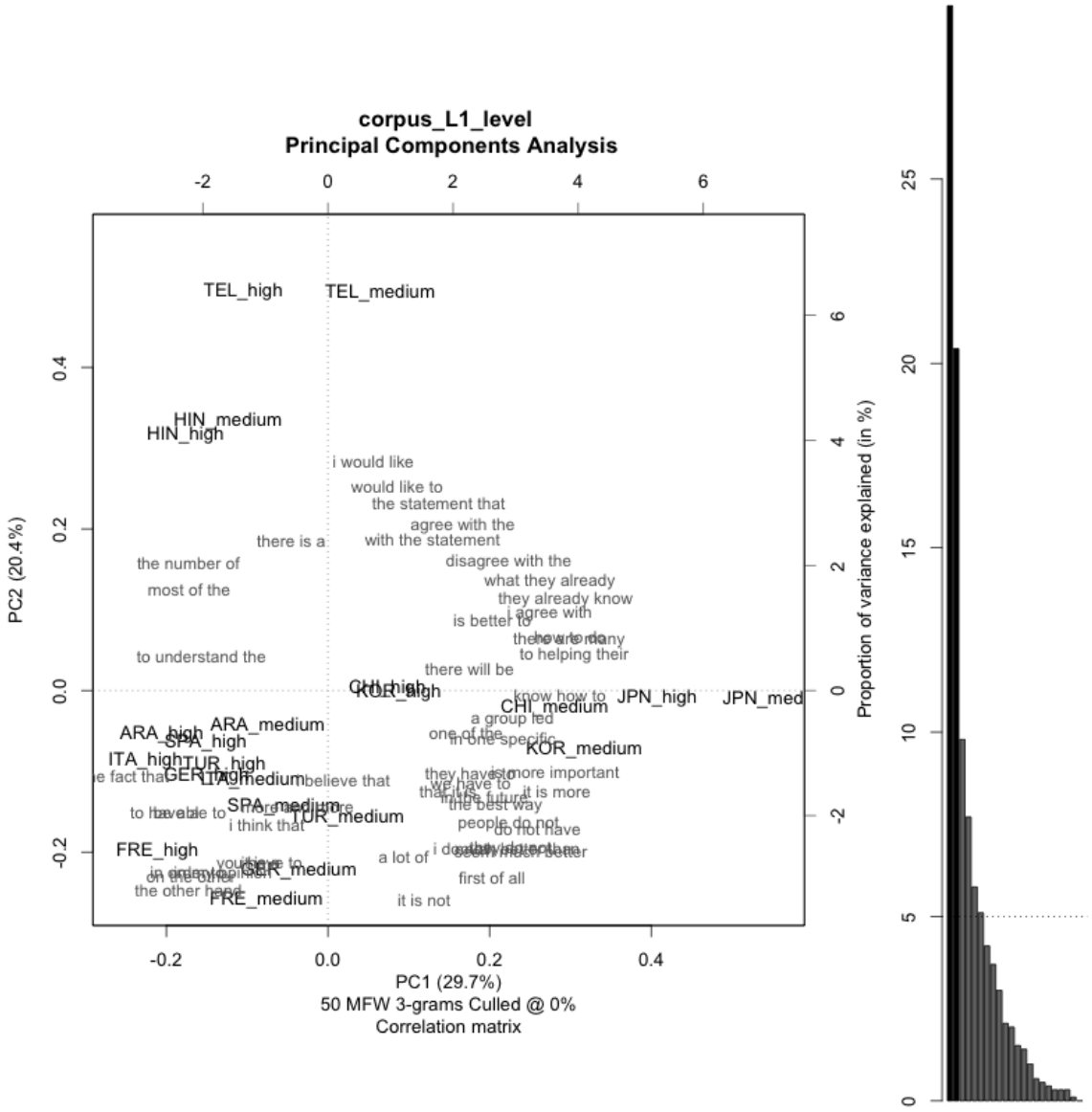
# Bootstrap Consensus Tree



# PCA-1


corpus_L1_level
Principal Components Analysis
TEL_medium
TEL_high
HIN_high
HIN_medium
ARA_high
TUR_high
KOR_high
CHI_high
ARA_medium
TUR_medium
CHI_medium
SPA_high
GER_high
ITA_high
KOR_medium
JPN_high
FRE_high
GER_medium
SPA_medium
ITA_medium
JPN_medium
FRE_medium
PC2 (13.7%)
PC1 (18.5%)
900 MFW 3-grams Culled @ 100%
Correlation matrix
20
10
0
-10
-20
-10
0
10
20
30

# PCA-2


corpus_L1_level
Principal Components Analysis
PC1 (29.7%)
PC2 (20.4%)
50 MFW 3-grams Culled @ 0%
Correlation matrix
TEL_high
TEL_medium
HIN_medium
HIN_high
ARA_high
ARA_medium
SPA_high
ITA_high
TUR_high
GER_high
ITA_medium
SPA_medium
TUR_medium
FRE_high
GER_medium
FRE_medium
KOR_high
JPN_high
JPN_med
CHI_medium
KOR_medium
i would like
would like to
the statement that
agree with the
with the statement
there is a
the number of
most of the
disagree with the
what they already
they already know
i agree with
is better to
to understand the
there will be
to helping their
know how to
a group fed
one of the
in one specific
they have to
it is more
the best way
people do not
do not have
a lot of
first of all
it is not
Proportion of variance explained (in %)
Principal components

# Next Class

- Hands on practice with Stylo, Wordcloud, Topic models etc (if stylo works by then!)
- Attendance question for today: Think of some scenarios where stylo can be a useful library for the kind of data you want to analyze in your disciplines?