Visualizing the relation between L2 proficiency and a learner’s native language
(work in progress)

Sowmya Vajjala
(joint work with Magdalena Wolska)

Iowa State University, USA

01 April 2016,
ALT Brownbag Colloquium Series
Talk outline

- Background
- Overview of the dataset and methods
- Some example visualizations
- Road ahead

Disclaimer: This is a work in progress. There is no published (or submitted) writeup yet.
A question to start ... 

Given a piece of English writing (say a paragraph), how many of you think you can harbor an informed guess about the native language of the author?
A question to start ...

What could be the native language of this author?
What could be the native language of this author?

People who think cars are useless may have forgotten a fact that the invention of cars really changed our lives very much. With the help of cars, our life becomes easier. Every morning, many people choose cars to drive to work. It is hard to imagine a world without car. I think people will use more and more cars. The second reason is that it needs a long time to invent a kind of new vehicle. Inventing a new thing is a long process, and need a lot of money and effort. If one day, a new vehicle was invented, it would take a long time to put it into practice. Twenty years are not sufficient for that long process.
Another Question

The author’s native language is Chinese. Now, a second question: On a scale of low-medium-high, what is the suitable proficiency level for this author?
Another Question

The author’s native language is Chinese. Now, a second question: On a scale of low-medium-high, what is the suitable proficiency level for this author? Medium, according to TOEFL raters.
Let us for the moment focus on the first question. What is the point of knowing L1 based on L2?
Let us for the moment focus on the first question. What is the point of knowing L1 based on L2?

Some answers: general curiosity, useful for SLA research, useful for NSA (ofcourse!! what is not useful??), challenging as a research problem and so on.
So what?

- Let us for the moment focus on the first question. What is the point of knowing L1 based on L2?
- Some answers: general curiosity, useful for SLA research, useful for NSA (of course!! what is not useful??), challenging as a research problem and so on.
- One interesting characteristic of the problem of L1 identification: It is a task that is difficult for human, but easy for a computer, unlike some popular machine learning tasks like say image recognition.
L1 identification and more

- Black-box predictive models already exist in research for native language identification.
- However, relation between L1 and L2 proficiency is not explored much in computational modeling in a way that can give qualitative insights.
- Our end goal is to computationally model this relationship in a way that is interpretable by humans.
- Visualizations are our first step to understand what to do to develop such models.
This talk in one question

- How far is visualization useful in forming hypotheses about the relationship between L1 and L2 proficiency?
Methods: Texts

- **TOEFL11 corpus of non-native English** (Blanchard et. al., 2013).
- Contains 11000 TOEFL essays, written by learners with 11 L1 backgrounds, and 3 proficiency levels (low, medium, high).
- We took only medium and high proficiency essays (8830 essays in total) as the low proficiency texts are unevenly distributed across L1s.
- 11 L1s: CHI, JPN, KOR, TEL, HIN, ARA, TUR, GER, ITA, SPA, FRE.
Methods: Features Studied

- 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 with Stanford Tagger.
Methods: Visualization Methods

Inspiration: Several papers from stylistics we read last year, which study authorship attribution. Visit the following URL for some references:
https://sites.google.com/site/computationalstylistics/papers-and-articles

We used:
- Bootstrap Consensus Trees
- Principal component analysis
Tools Used

- corpus pre-processing: combination of perl and command line scripting
- Visuals exclusively using the software package Stylo, written in R (Eder et.al., 2013) (no extra code written to generate visuals!)
- Visit: https://cran.r-project.org/web/packages/stylo/index.html for more details.
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.
The Measure of Similarity

We need some way of measuring the similarity, if we want to group texts together. We used Burrows’ Delta (Burrows, 2002), which is used in stylistics to differentiate between texts.

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 continued

If we have two documents D1 and D2, represented as $(z_1(D1), z_2(D1) \ldots z_n(D1))$ and $(z_1(D2), z_2(D2) \ldots z_n(D2))$ 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).
Why bootstrap? Why not use plain dendrogram clusters?

Figure: Cluster Dendrograms with Word tri-grams
Some Visual Observations with BCT-1

corpus_L1_level
Bootstrap Consensus Tree

100-1000 MFW 3-grams Culled @ 0%
Classic Delta distance Consensus 0.5
Some Visual Observations with BCT-2

corpus_L1_level
Bootstrap Consensus Tree

100-1000 MFC 3-grams Culled @ 0%
Classic Delta distance Consensus 0.5
Some Visual Observations with BCT-3
Some observations so far

- Grouping by word/character trigrams shows two clear clusters (Telugu-Hindi; Chinese-Japanese-Korean). The grouping continues to exist even at high L2 proficiency.
- Most of the other L1s also group together at both proficiencies, although the grouping with other L1s is not as strong.
- Grouping by POS trigrams shows four clusters (TEL-HIN, FRE-GER-ITA, CHI-JPN-KOR, ARA-SPA-TUR)
Principal Component Analysis (PCA)

- PCA is a statistical method used to reduce the dimensionality of data.
- It is also a technique used to find out the underlying structure in your data, by looking for a principal component that separates the data into groups of similar objects (texts).
- What it does: it forms "principal components" which are a combination of "features" (word, character or POS ngrams in our case), thereby reducing the number of features you have.
Some Visual Observations with PCA -1
Some Visual Observations with PCA -2

corpus_L1_level
Principal Components Analysis

PC1 (20.9%)
1000 MFC 3-grams Culled @ 0%
Correlation matrix

PC2 (18.1%)
Some Visual Observations with PCA -3
Some Visual Observations with PCA -4
Some Visual Observations with PCA -5
What we learnt so far

- Visualizations capture meaningful latent grouping of proficiencies and L1s.
- There are broadly 3 groupings (sometimes 4), and grouping of L1s seem to be closer to the sociolinguistic relationship between languages (e.g., Telugu-Hindi grouping together although they belong to different language families).
- Some L1 grouping can be seen even among high proficiency texts.
Road Ahead

- Remove prompt specific phrases properly and check if the same similarity relations still hold.
- Using these visual observations to form hypotheses that build machine learning models that distinguish native languages based on L2 writing and its proficiency.
- Eventual goal: to be able to show concrete patterns (n-grams in our case) that distinguish learners of different proficiencies and L1s, while also having good predictive models.
References I
