Trees and Tweets: Mining Billions to Understand Human Migration and Regional Linguistic Variation

Jack Grieve
Centre for Forensic Linguistics
Aston University

30 April, 2014
JISC London
Research Goals: Year 1

Map patterns of regional linguistic variation in modern British and American English based on a multi-billion corpus of geocode tweets.

Map historical and modern human migration patterns in the UK and the US based on millions of family trees.
Research Goals: Year 2

Examine how the patterns of regional linguistic variation identified in the Twitter corpora are related to historical and modern migration patterns, in addition to Twitter communication patterns.
Project Partners

1. **Twitter Dialect Analysis**: Jack Grieve, Centre for Forensic Linguistics, School of Languages and Social Sciences, Aston University

2. **Family Tree Migration Analysis**: Diansheng Guo, Spatial Data Mining and Visual Analytics Lab, Department of Geography, University of South Carolina
Deliverables

Journal articles, conference presentations, and press releases.

Data and tools for researchers.

Websites allowing the public and researchers to map the emergence of new words and family tree migration patterns.
Dialect Analysis

Because we have access to such a large corpus of geocoded and timestamped tweets, we will be able to focus on lexical variation.

In particular, for the first time we will investigate how new words spread geographically on a large scale, both within and between the two countries.
Identifying New Words

Emerging words will be identified in three ways:

1. Current lexicographic research
2. Data mining online dictionaries (e.g. Urban Dictionary, Wiktionary)
3. Automatic identification based on the corpus
Automatically Identifying New Words

By looking for words whose frequency increases dramatically over the course of the year, possible new words can be identified (semi-)automatically.

For example, *unmeet* and *thotties* were identified as new words using this method in preliminary analyses.
Wish there was a way to "unmeet" certain people
(3/1/2013)

If I could just unmeet some people (3/1/2013)

Sometimes its so hard and i just want to unmeet you
#butreally (14/1/2013)
Unmeet
February
Thotties

These thotties is out here (3/1/2013)

why are my bestfriends thotties following me on instagram… (3/1/2013)

@RockingWithJAYs @Cat_Alena yall just dont understand dem thotties be calling me name (26/1/2013)
Common Patterns

Once a large number of new words have been found, we will identify common patterns of lexical spread and compare how words spread from different origins.

These patterns will then be compared to the communication and migration patterns identified by the American team.
Trees and Tweets: 
Mining Billions to Understand Human Migration and Regional Linguistic Variation

Jack Grieve
Centre for Forensic Linguistics
Aston University

30 April, 2014
JISC London