English-language social media in Finland: Twitter data collection and analysis

Steven Coats, University of Oulu
31 August 2014 – 12th ESSSE Conference, SLANG27
Outline

1. Twitter as a ”Linguistic Landscape”
2. Methods
   – Compiling the Finland tweet data, processing the tweets, excluding non-English tweets, preparing the comparison corpus
3. Preliminary analysis
   – Rank/frequency profiles for lexical items, lexical clusters, collocations/n-grams
4. Preliminary conclusions
Twitter as a Linguistic Landscape?

• Linguistic landscapes typically describe the use of language in public physical spaces and thus provide insight into localized issues of sociolinguistic interest such as language use and identity, particularly in bi- and multilingual environments (Landry and Bourhis 1997, Shohamy and Gorter 2008)

• Linguistic landscapes are real physical spaces?
Twitter as a linguistic landscape?

BUT:

• ”It has to be determined what belongs to the linguistic landscape.” (Gorter 2006: 3)

• ”Almost all humans today live in a textually mediated world, and the texts which mediate and impact on our lives are by no means all fixed in (physical) space.” (Sebba 2010: 61)

• Online and virtual space constitute increasingly important domains of language use for much of the world’s population.

• An expanded linguistic landscape concept could incorporate online media

• Of particular interest would be media platforms that represent sites of bi- and multilingual interaction

• Twitter: ~500m users, ~340m tweets per day
Some recent Twitter research

<table>
<thead>
<tr>
<th>Field</th>
<th>Nr. articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Law</td>
<td>2</td>
</tr>
<tr>
<td>Physics</td>
<td>2</td>
</tr>
<tr>
<td>Mathematics</td>
<td>3</td>
</tr>
<tr>
<td>Geography</td>
<td>4</td>
</tr>
<tr>
<td>Sociology</td>
<td>10</td>
</tr>
<tr>
<td>Political Science</td>
<td>12</td>
</tr>
<tr>
<td>Medicine</td>
<td>15</td>
</tr>
<tr>
<td>Economics</td>
<td>29</td>
</tr>
<tr>
<td>Information science (including some language/linguistics studies)</td>
<td>37</td>
</tr>
<tr>
<td>Communications</td>
<td>50</td>
</tr>
<tr>
<td>Computer science</td>
<td>64</td>
</tr>
</tbody>
</table>

(Boyd 2014)
Some recent Twitter linguistics/NLP research

• Developing accurate Twitter-English translation and PoS tagging tools (Gimpel et al. 2011; main problem is the high degree of orthographic variation)

• Emoticon sociolinguistics :--DD :) (-;
  – Tweet lexical/emoticon frequencies and demographic variables e.g. gender (Bamman, Eisenstein and Schnoebelen 2014, Schnoebelen 2012)

• Twitter geographical variation/dialectology (Eisenstein et al. 2010)
This project: steps towards characterizing an emerging online English variety?

• To what extent are people in Finland using English in Twitter?

• Is English-language Twitter in Finland different from “normal” Twitter English? Steps:
  – Compile corpus of English-language tweets originating from Finland
  – Compare with corpus with no geographical restrictions
Compiling the data: Twitter API

- Better than native Twitter web interface (e.g. number, chronological range, geocoding, other features of accessible Tweets)
- Unlimited Twitter stream (“firehose”) is proprietary big data, only available to companies working in the “Twitter ecosystem”
- 1% stream available to all
Tweepy

• Interact with the Twitter API using Python
• Automate authentication process
• Customizable
Collecting tweets: Python script

```python
# -*- coding: utf-8 -*-
import sys
import tweepy
import unicodedata
import codecs

consumer_key = "y1xzUXUqmiR77A1qngULHq"
counter_secret = "w21nQcJqC302zC0keSodjC0ycz7lkMGKQjF52btweMK"
access_key = "26257923-h3z31XGHRm2b3aZunFyLy5yYsystukwXW9PcL0zLY7"
access_secret = "5jMoRh50G2z2027K0YjpyWNttspF6y250xtkScEKA"

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_key, access_secret)
api = tweepy.API(auth)

class CustomStreamListener(tweepy.StreamListener):
    def on_status(self, status):
        try:
            print(u"%s\n%s\n%s\n\n" % (status.text,
            status.author.screen_name,
            status.created_at,
            status.source, status.coordinates, status.place))
            with codecs.open('test1.txt', 'ab', 'utf-8') as f:
                newline = u'\n'
                linebreak = u'\n'
                mylist = (status.text, status.author.screen_name, status.created_at, status.source, status.coordinates, status.place)
                f.write (u\"%s\" % (mylist)+linebreak)
        except Exception, e:
            print >> sys.stderr, 'Encountered Exception:', e
            pass

    def on_error(self, status_code):
        print >> sys.stderr, 'Encountered error with status code!', status_code
        return True # Don't kill the stream

    def on_timeout(self):
        print >> sys.stderr, 'Timeout...
        return True # Don't kill the stream

sapi = tweepy.streaming.Stream(auth, CustomStreamListener())
sapi.filter(locations=[21, 60, 29, 70])
```
Limiting geographical range of collected tweets

- Most tweets do not have geocoordinates, users can opt-in
- Summer 2013: Geolocation protocol changed in API (not just long/lat)
- Twitter stream access: 1% of all available tweets returned
- Tweets from Sweden, Norway and Russia returned as well
Output

- ~139,000 tweets
- ~1,152,000 words
- Tagged with geographic location (longitude and latitude)
- Finland English Corpus
Geographical distribution of Finland English Corpus material

Location of individual tweets harvested by script
Output: Comparison Corpus

• Global tweet database compiled 2008–9 at Texas A&M University, Texas, USA
• ~299,000 tweets
• ~3,610,000 words
• No geographical tags
• Comparison Corpus
Cleaning up the tweets: regex

• Removing #hashtags (including #rt retweets), @usernames, urls:
  
  ```
  ((mailto:\|(news|ht\f)tp(s?)\:\//\)/\{1\}\S+) ; @\w+; \#\w+
  ```

• Identifying and removing (some) automated tweets of limited linguistic interest
  
  Automated weather reports, automated hourly tweets announcing the time, e.g.
Processing: Language detection

- English, Finnish, Swedish, Russian, Norwegian, Danish, French, Chinese, German, Polish, Japanese, (etc.) tweets
- Language detection: chromium compact language detection, a Python binding of chromium (McCandless/Sites 2013)
- Utilizes probability matrices for character 4-grams to assign language
- Each individual tweet was tagged with the most likely language according to the assignation algorithm from the cld script
<table>
<thead>
<tr>
<th>Language</th>
<th>Comparison Corpus</th>
<th>Finland English Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of tweets</td>
<td>% of total</td>
</tr>
<tr>
<td>English</td>
<td>182244</td>
<td>60.9</td>
</tr>
<tr>
<td>Finnish</td>
<td>59</td>
<td>0.01</td>
</tr>
<tr>
<td>Swedish</td>
<td>138</td>
<td>0.05</td>
</tr>
<tr>
<td>Russian</td>
<td>952</td>
<td>0.3</td>
</tr>
<tr>
<td>French</td>
<td>924</td>
<td>0.3</td>
</tr>
<tr>
<td>Norwegian/Danish</td>
<td>213</td>
<td>0.07</td>
</tr>
<tr>
<td>Turkish</td>
<td>318</td>
<td>0.1</td>
</tr>
<tr>
<td>Japanese</td>
<td>31058</td>
<td>10.4</td>
</tr>
<tr>
<td>Mandarin</td>
<td>11580</td>
<td>3.9</td>
</tr>
<tr>
<td>German</td>
<td>1777</td>
<td>0.59</td>
</tr>
<tr>
<td>Spanish</td>
<td>9652</td>
<td>3.2</td>
</tr>
<tr>
<td>Unknown</td>
<td>40907</td>
<td>13.7</td>
</tr>
<tr>
<td>Others</td>
<td>19620</td>
<td>6.6</td>
</tr>
<tr>
<td><strong>Σ</strong></td>
<td><strong>299442</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>
Corpora basic statistics, after processing steps

Finland English Corpus

Sample size: \( N = 362351 \)
Vocabulary size: \( V = 32934 \)
Range of freq's: \( f = 1 \ldots 11775 \)
Mean / median: \(\mu = 11.00234, \ M = 1\)
Hapaxes etc.: \(V_1 = 20735, \ V_2 = 4137\)

Comparison Corpus

Sample size: \( N = 2536815 \)
Vocabulary size: \( V = 306332 \)
Range of freq's: \( f = 1 \ldots 77829 \)
Mean / median: \(\mu = 8.28126, \ M = 1\)
Hapaxes etc.: \(V_1 = 255988, \ V_2 = 14991\)
Analysis: Rank/frequency profiles and vocabulary growth measures

<table>
<thead>
<tr>
<th>k</th>
<th>f</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11777</td>
<td>i</td>
</tr>
<tr>
<td>2</td>
<td>8241</td>
<td>to</td>
</tr>
<tr>
<td>3</td>
<td>8017</td>
<td>the</td>
</tr>
<tr>
<td>4</td>
<td>6261</td>
<td>and</td>
</tr>
<tr>
<td>5</td>
<td>5664</td>
<td>a</td>
</tr>
<tr>
<td>6</td>
<td>5213</td>
<td>you</td>
</tr>
<tr>
<td>7</td>
<td>4622</td>
<td>my</td>
</tr>
<tr>
<td>8</td>
<td>4146</td>
<td>im</td>
</tr>
<tr>
<td>9</td>
<td>4080</td>
<td>in</td>
</tr>
<tr>
<td>10</td>
<td>3995</td>
<td>is</td>
</tr>
<tr>
<td>11</td>
<td>3598</td>
<td>of</td>
</tr>
<tr>
<td>12</td>
<td>3569</td>
<td>it</td>
</tr>
<tr>
<td>13</td>
<td>3149</td>
<td>for</td>
</tr>
<tr>
<td>14</td>
<td>3116</td>
<td>me</td>
</tr>
<tr>
<td>15</td>
<td>3084</td>
<td>at</td>
</tr>
<tr>
<td>16</td>
<td>2913</td>
<td>so</td>
</tr>
<tr>
<td>17</td>
<td>2812</td>
<td>that</td>
</tr>
<tr>
<td>18</td>
<td>2356</td>
<td>on</td>
</tr>
<tr>
<td>19</td>
<td>2325</td>
<td>this</td>
</tr>
<tr>
<td>20</td>
<td>2249</td>
<td>just</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>k</th>
<th>f</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80282</td>
<td>the</td>
</tr>
<tr>
<td>2</td>
<td>70000</td>
<td>to</td>
</tr>
<tr>
<td>3</td>
<td>57199</td>
<td>i</td>
</tr>
<tr>
<td>4</td>
<td>51071</td>
<td>a</td>
</tr>
<tr>
<td>5</td>
<td>36912</td>
<td>and</td>
</tr>
<tr>
<td>6</td>
<td>31890</td>
<td>is</td>
</tr>
<tr>
<td>7</td>
<td>31063</td>
<td>of</td>
</tr>
<tr>
<td>8</td>
<td>29301</td>
<td>for</td>
</tr>
<tr>
<td>9</td>
<td>28979</td>
<td>in</td>
</tr>
<tr>
<td>10</td>
<td>26432</td>
<td>my</td>
</tr>
<tr>
<td>11</td>
<td>24476</td>
<td>you</td>
</tr>
<tr>
<td>12</td>
<td>23840</td>
<td>on</td>
</tr>
<tr>
<td>13</td>
<td>23145</td>
<td>it</td>
</tr>
<tr>
<td>14</td>
<td>17967</td>
<td>that</td>
</tr>
<tr>
<td>15</td>
<td>15292</td>
<td>with</td>
</tr>
<tr>
<td>16</td>
<td>15035</td>
<td>at</td>
</tr>
<tr>
<td>17</td>
<td>13220</td>
<td>just</td>
</tr>
<tr>
<td>18</td>
<td>13151</td>
<td>have</td>
</tr>
<tr>
<td>19</td>
<td>11978</td>
<td>be</td>
</tr>
<tr>
<td>20</td>
<td>11872</td>
<td>this</td>
</tr>
</tbody>
</table>

20 most frequent terms, Finland English Corpus and Comparison Corpus
• Rank-frequency profile exhibits familiar Zipfian shape

• What about terms that occur infrequently?
Hapax legomenon:
a word that occurs once in a text/corpus

Dis legomenon:
a word that occurs twice in a text/corpus

Etc.

\[ V(m, N) = \sum_{i=1}^{V(N)} I[f(i, N) = m] \]

the number of types with frequency \( m \) in a sample of \( N \) tokens (Baayen 2001: 8)
Vocabulary growth

- Type-token ratio and vocabulary growth rate (rate at which new words are added to the corpus) can indicate lexical richness.
- \( V(1,N)/V \) of Finland English Corpus grows somewhat more slowly than does that of Comparison Corpus initially, but then increases.
- But: measures such as type-token ratio are strongly influenced by corpus size (Baayen 2001: 24ff.)
Lexicon: clustering

• Clustering shows which words are likely to appear together in a tweet
• A matrix is created in which axes are all word types and all individual tweets.
• Using least squares method (Ward’s), determine geometric distance between all matrix elements.
• Remove word types that have extremely low frequencies
• R text mining package (tm)
Lexicon: clustering algorithms

Term clustering, Finland English Corpus, 98% sparsity

adistMatrix
hclust ("daisy", "ward")
Lexicon: clustering algorithms

- Remove common English function words, including articles, conjunctions, pronouns, demonstratives, numbers, modal verbs

[1]  "a"
[6]  "ago"
[11]  "also"
[16]  "an"
[21]  "anything"
[26]  "as"
[31]  "been"
[36]  "beside"
[41]  "both"
[46]  "can"n't
[51]  "do"
[56]  "don't"
[61]  "eighteenth"
[66]  "every"
[71]  "every"th
[76]  "except"
[81]  "fifteenth"
[86]  "five"
[91]  "fourteen"
[96]  "get"
[101]  "about"
[106]  "all"
[111]  "although"
[116]  "and"
[121]  "anywhere"
[126]  "at"
[131]  "before"
[136]  "between"
[141]  "each"
[146]  "could"
[151]  "does"
[156]  "down"
[161]  "eighth"
[166]  "every"th
[171]  "everybody"
[176]  "far"
[181]  "fifth"
[186]  "for"
[191]  "fourteenth"
[196]  "gets"
[201]  "above"
[206]  "almost"
[211]  "always"
[216]  "another"
[221]  "are"
[226]  "back"
[231]  "being"
[236]  "beyond"
[241]  "but"
[246]  "can't"
[251]  "does"
[256]  "doesn't"
[261]  "during"
[266]  "eightieth"
[271]  "enough"
[276]  "everyone"
[281]  "few"
[286]  "fifty"
[291]  "fortieth"
[296]  "fourth"
[301]  "getting"
[306]  "after"
[311]  "along"
[316]  "am"
[321]  "an"y
[326]  "aren't"
[331]  "else"
[336]  "below"
[341]  "billion"
[346]  "by"
[351]  "can't"
[356]  "could"
[361]  "can't"
[366]  "did"
[371]  "done"
[376]  "eight"
[381]  "eighteen"
[386]  "eighty"
[391]  "even"
[396]  "every"th
[401]  "everywhere"
[406]  "everyone"
[411]  "every"th
[416]  "everyone"
[421]  "everyone"
[426]  "everyone"
[431]  "everyone"
[436]  "everyone"
[441]  "everyone"
[446]  "everyone"
[451]  "everyone"
[456]  "everyone"
[461]  "everyone"
[466]  "everyone"
[471]  "everyone"
[476]  "everyone"
[481]  "everyone"
[486]  "everyone"
[491]  "everyone"
[496]  "everyone"
[501]  "everyone"
[506]  "everyone"
[511]  "everyone"
[516]  "everyone"
[521]  "everyone"
[526]  "everyone"
[531]  "everyone"
[536]  "everyone"
[541]  "everyone"
[546]  "everyone"
[551]  "everyone"
[556]  "everyone"
[561]  "everyone"
[566]  "everyone"
[571]  "everyone"
[576]  "everyone"
[581]  "everyone"
[586]  "everyone"
[591]  "everyone"
[596]  "everyone"
[601]  "everyone"
[606]  "everyone"
[611]  "everyone"
[616]  "everyone"
[621]  "everyone"
[626]  "everyone"
[631]  "everyone"
[636]  "everyone"
[641]  "everyone"
[646]  "everyone"
[651]  "everyone"
Lexicon: clustering algorithms

Term clustering, Finland English Corpus, 98% sparsity, function words removed

31 August 2014
ESSE, SLANG27

Steven Coats: English-Language Social Media in Finland
Lexicon: clustering algorithms

Term clustering, Comparison Corpus, 97% sparsity

bdistMatrix
hclust (", "ward")
Term clustering, Comparison Corpus, 98% sparsity, function words removed

adistMatrix
hclust (*, "ward")
Collocations in the Finland English Corpus: bi- and trigrams

1. "('im', 'at')", "1499"
2. "('in', 'the')", "688"
3. "('i', 'have')", "677"
4. "('i', 'dona')", "664"
5. "('i', 'love')", "630"
6. "('follow', 'me')", "558"
7. "('i', 'just')", "546"
8. "('to', 'be')", "526"
9. "('and', 'i')", "519"
10. "('going', 'to')", "492"
11. "('helsinki', 'w')", "477"
12. "('of', 'the')", "473"
13. "('2', 'others')", "442"
14. "('i', 'cant')", "431"
15. "('but', 'i')", "416"
16. "('this', 'is')", "404"
17. "('for', 'the')", "388"
18. "('on', 'the')", "387"
19. "('i', 'am')", "379"
20. "('i', 'was')", "372"

1. "('w', '2', 'others')", "441"
2. "('w', '3', 'others')", "255"
3. "('i', 'love', 'you')", "210"
4. "('was', 'out', 'running')", "175"
5. "('i', 'want', 'to')", "162"
6. "('w', '4', 'others')", "147"
7. "('helsinki', 'w', '2')", "145"
8. "('i', 'have', 'to')", "143"
9. "('hel', 'vanta', 'w')", "134"
10. "('at', 'hel', 'vanta')", "133"
11. "('im', 'at', 'hel')", "133"
12. "('im', 'going', 'to')", "120"
13. "('i', 'dona', 'know')", "116"
14. "('im', 'at', 'kotilinnake')", "115"
15. "('w', '5', 'others')", "101"
16. "('follow', 'me', 'i')", "100"
17. "('come', 'to', 'finland')", "99"
18. "('others', 'im', 'at')", "95"
19. "('helsinki', 'w', '3')", "89"
20. "('i', 'need', 'to')", "87"
Collocations in the Finland English Corpus: 4-grams

1. "('helsinki', 'w', '2', 'others')", "145"
2. "('at', 'hel', 'vantaan', 'w')", "133"
3. "('im', 'at', 'hel', 'vantaan')", "133"
4. "('helsinki', 'w', '3', 'others')", "89"
5. "('as', 'the', 'mayor', 'of')", "67"
6. "('just', 'posted', 'a', 'photo')", "67"
7. "('come', 'to', 'finland', 'ill')", "64"
8. "('finland', 'ill', 'wait', 'for')", "64"
9. "('hope', 'someday', 'youll', 'come')", "64"
10. "('someday', 'youll', 'come', 'to')", "64"
11. "('to', 'finland', 'ill', 'wait')", "64"
12. "('youll', 'come', 'to', 'finland')", "64"
13. "('love', 'you', 'so', 'much')", "55"
14. "('follow', 'me', 'i', 'need')", "54"
15. "('i', 'love', 'you', 'so')", "53"
16. "('w', '2', 'others', 'pic')", "52"
17. "('helsinki', 'w', '4', 'others')", "51"
18. "('i', 'dont', 'want', 'to')", "51"
19. "('have', 'a', 'great', 'time')", "43"
20. "('hope', 'you', 'have', 'a')", "43"
Collocations in the Comparison Corpus: bi- and trigrams

1: ("in", "the")
2: ("listening", "to")
3: ("i", "am")
4: ("on", "the")
5: ("to", "the")
6: ("of", "the")
7: ("for", "the")
8: ("going", "to")
9: ("to", "be")
10: ("i", "have")
11: ("at", "the")
12: ("for", "a")
13: ("i", "think")
14: ("to", "get")
15: ("have", "a")
16: ("is", "a")
17: ("have", "to")
18: ("with", "the")
19: ("to", "go")
20: ("in", "a")

1: ("lastfm", "listening", "to")
2: ("listening", "to", "the")
3: ("am", "listening", "to")
4: ("i", "am", "listening")
5: ("new", "blog", "post")
6: ("team", "hijack", "playing")
7: ("going", "to", "be")
8: ("thanks", "for", "the")
9: ("i", "have", "a")
10: ("i", "have", "to")
11: ("to", "go", "to")
12: ("is", "going", "to")
13: ("i", "want", "to")
14: ("i", "need", "to")
15: ("a", "lot", "of")
16: ("im", "listening", "to")
17: ("looking", "forward", "to")
18: ("i", "think", "i")
19: ("im", "going", "to")
20: ("getting", "ready", "to")
### Collocations in the Comparison Corpus: 4-grams

<table>
<thead>
<tr>
<th></th>
<th>Collocation</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>('i', 'am', 'listening', 'to')</td>
<td>1137</td>
</tr>
<tr>
<td>2</td>
<td>('heavy', 'traffic', 'on', 'the')</td>
<td>419</td>
</tr>
<tr>
<td>3</td>
<td>('posted', 'by', 'espncoms', 'james')</td>
<td>413</td>
</tr>
<tr>
<td>4</td>
<td>('by', 'espncoms', 'james', 'walker')</td>
<td>398</td>
</tr>
<tr>
<td>5</td>
<td>('team', 'hyjak', 'playing', 'xbox')</td>
<td>367</td>
</tr>
<tr>
<td>6</td>
<td>('hyjak', 'playing', 'xbox', '360')</td>
<td>357</td>
</tr>
<tr>
<td>7</td>
<td>('fiddling', 'with', 'my', 'blog')</td>
<td>273</td>
</tr>
<tr>
<td>8</td>
<td>('with', 'my', 'blog', 'post')</td>
<td>273</td>
</tr>
<tr>
<td>9</td>
<td>('is', 'going', 'to', 'be')</td>
<td>208</td>
</tr>
<tr>
<td>10</td>
<td>('playing', 'xbox', '360', 'dashboard')</td>
<td>199</td>
</tr>
<tr>
<td>11</td>
<td>('between', 'the', 'junctions', 'with')</td>
<td>183</td>
</tr>
<tr>
<td>12</td>
<td>('still', 'writing', 'the', 'new')</td>
<td>182</td>
</tr>
<tr>
<td>13</td>
<td>('the', 'new', 'blog', 'post')</td>
<td>182</td>
</tr>
<tr>
<td>14</td>
<td>('writing', 'the', 'new', 'blog')</td>
<td>182</td>
</tr>
<tr>
<td>15</td>
<td>('playing', 'xbox', '360', 'dashboard')</td>
<td>177</td>
</tr>
<tr>
<td>16</td>
<td>('am', 'listening', 'to', 'the')</td>
<td>173</td>
</tr>
<tr>
<td>17</td>
<td>('for', 'the', 'first', 'time')</td>
<td>173</td>
</tr>
<tr>
<td>18</td>
<td>('dashboard', 'watching', 'a', 'video')</td>
<td>159</td>
</tr>
<tr>
<td>19</td>
<td>('360', 'dashboard', 'watching', 'a')</td>
<td>154</td>
</tr>
<tr>
<td>20</td>
<td>('xbox', '360', 'dashboard', 'watching')</td>
<td>154</td>
</tr>
</tbody>
</table>
Issues and further directions

• Corpus compilation
  – Automatic language detection script produces many ”Unknowns”
  – Correction for orthographic variation needed
  – Removing ”noise” in terms of automated tweets

• Corpus analysis
  – Characterization of discourse function of n-grams
  – Correlation of geocoded information with various text or non-text variables (n-gram structure, country, province/region)
  – Comparison with other English language corpora (e.g. Davies’ GloWBe)
Preliminary conclusions

• Finnish Twitter users seem to use English more than any other language
• Lexical richness in English-language tweets in Finland is not significantly different from that of tweets from elsewhere
• English-language Twitter use in Finland reflects Finland-specific topicality in terms of lexis
  – Local communicative orientation despite globalized nature of the medium
• Collocations or lexical bundles seem to be dominated by tweets that may or may not represent human users
• The online linguistic landscape constituted by Twitter in Finland “may serve important informational and symbolic functions as a marker of the relative power and status of the linguistic communities inhabiting the territory” (Landry and Bourhis 1997: 23): Language shift?
Literature

  Dordrecht: Kluwer.


• NLTK: http://nltk.org/


• Tweepy: https://github.com/tweepy/tweepy
