

## Linguistic Standards in the US using Twitter

by Richard Heimann

<p>Big Social Data is driven by a social aspect and ultimately analyzes data that could serve directly as or as a proxy for other more substantive variables. The Flesch-Kincaid index, which you may all be familiar with as a consequence of using Microsoft Word. It has for some time provided the readability index to documents. </p>

<p>The Guardian in February 2013 used the Flesch-Kincaid index to track the reading level of every state of the union address and noted how the linguist standard of the <a href="http://www.guardian.co.uk/world/interactive/2013/feb/12/state-of-the-union-reading-level" target="new">presidential address </a> has declined.</p>

<p>The Readability Ease Index is the average sentence length weighted then subtracted from the average number of syllables per word. The output generally ranges from 0 - 100. To provide examples the Reader's Digest magazine has a readability index of about 65, Time magazine scores about 52, an average 6th grade student age 11 has written assignments at a readability score of 60–70, and the Harvard Law Review has a general readability score in the low 30s. </p>

<p>The highest (easiest) readability score possible is around 120 (meaning every sentence consisting of only two one-syllable words). The score does not have a theoretical lower bound. It is possible to make the score as low as you want by arbitrarily including words with many syllables. In Twitter this could easily happen; as an example a tweet where LOL is repeated to the max character limit of 140 would possess subsequent indices well below 0. </p>

<p>This sentence, for example, taken as a reading passage unto itself, has a readability score of about thirty-three. The sentence, "The Australian platypus is seemingly a hybrid of a mammal and reptilian creature" is a 24.4 as it has 26 syllables and 13 words. One particularly long sentence about sharks in chapter 64 of Moby-Dick has a readability score of -146.77. </p>

<p>The index is inversely related to its linguistic sophistication. A high score is easier to read or put different poorly written. An example of a low score or a tweet written with high sophistication [Table 1] is as follows, "this gas situation is absolutely ridiculous" and written at an 11th grade level and has a mean centered value well below zero. The It is parsimonious and more dense with syllables on average than other tweets. The location of the Tweet is Mahwah NJ, located about 20 miles outside of New York City (NYC).</p>

<p>The tweet [Table 2] "down here in beach bout to shut this down wit & feeling the vibe s" is written at a 4th grade level and has a mean centered value well above zero. This is an example of a high score or a Tweet written with low sophistication. It has but one non-monosyllable word. The location of the tweet is Myrtle Beach SC.</p>

**Table 1**

Clean Text	<b>“this gas situation is absolutely ridiculous.”</b>
Language	english
Latitude	41.0862
Longitude	-74.1520
Kincaid	14.3
Flesch-Kincaid (Mean Centered)	-76.273849
Leesbaarheid Grade	11

**Table 2**

Clean Text	<b>“down here in beach bout to shut this down wit &amp; feeling the vibe s.”</b>
Language	english
Latitude	33.68709
Longitude	-78.88915
Kincaid	3.5
Flesch-Kincaid (Mean Centered)	20.42615
Leesbaarheid Grade	4

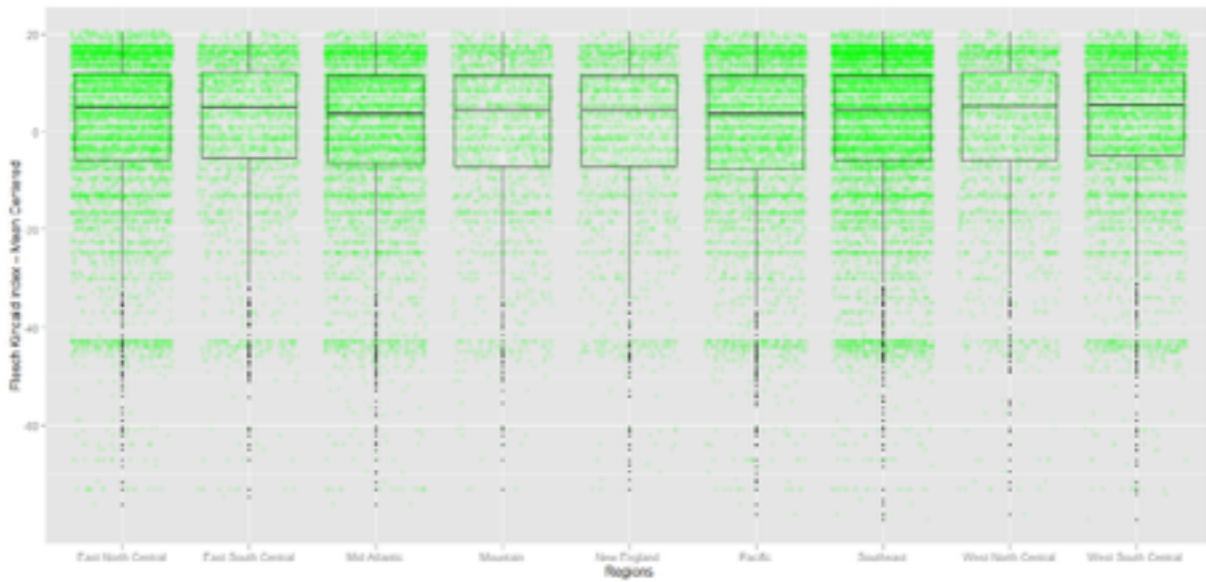
By mean centering the data, that is subtracting the global mean from each region we can quickly identify deviation from the global mean. The Mid-Atlantic, Mountain, New England, and Pacific are all below the global mean whereas East North Central, East South Central, Southeast, West North Central, and West South Central are all above. You can also quickly see that the Pacific and the West South Central regions deviate most in their respective direction from the global mean.

Table 3

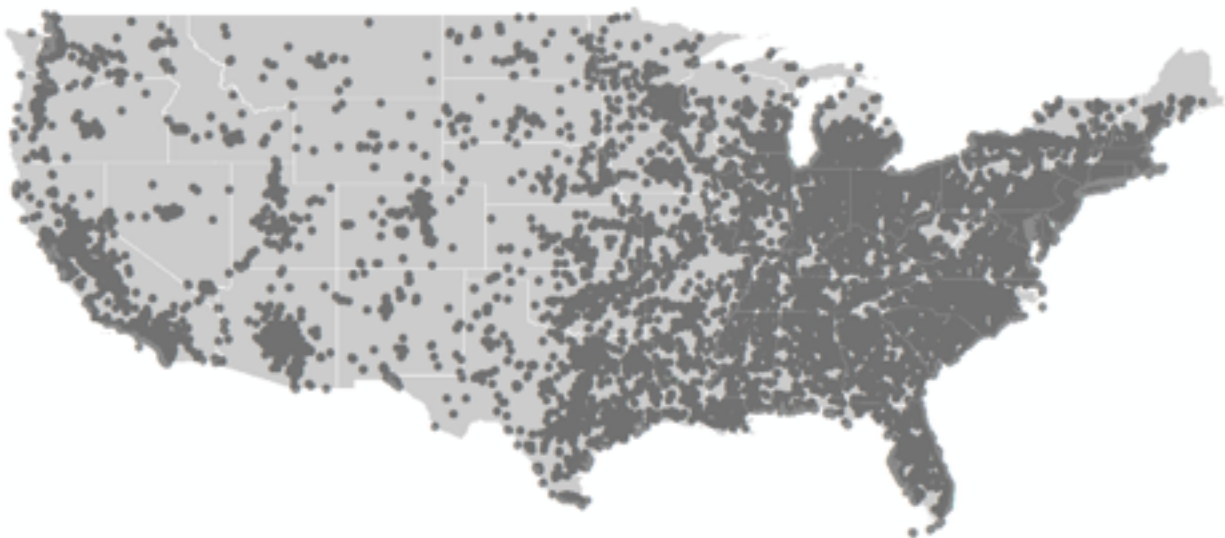
Region	mean	SD	0%	25%	50%	75%	100%	data:n
East North Central	0.6193	16.514	-76.274	-5.77	4.93	11.92	20.426	7579
East South Central	0.6314	16.576	-74.673	-5.27	4.93	12.23	20.426	3028
Mid-Atlantic	-0.1988	16.590	-76.273	-6.47	3.73	11.43	20.426	6278
Mountain	-0.1212	16.586	-73.174	-7.00	4.32	11.43	20.426	2452
New England	-0.1837	16.864	-73.174	-7.00	4.32	11.43	20.426	2392
Pacific	-0.8560	17.276	-78.274	-7.78	3.72	11.43	20.426	5390
Southeast	0.1469	16.730	-79.373	-5.78	4.32	11.43	20.426	10022
West North Central	0.6010	16.385	-78.274	-5.78	5.22	12.23	20.426	2781
West South Central	0.8323	16.386	-79.273	-4.77	5.33	12.12	20.426	5572

Another way of exploring [Graph 1] the data are box plots by region with underlying scatter plots. We see much of the same information captured by the summary statistics but the addition of the jitter allows us to get a sense of the distribution.

[Graph 1]

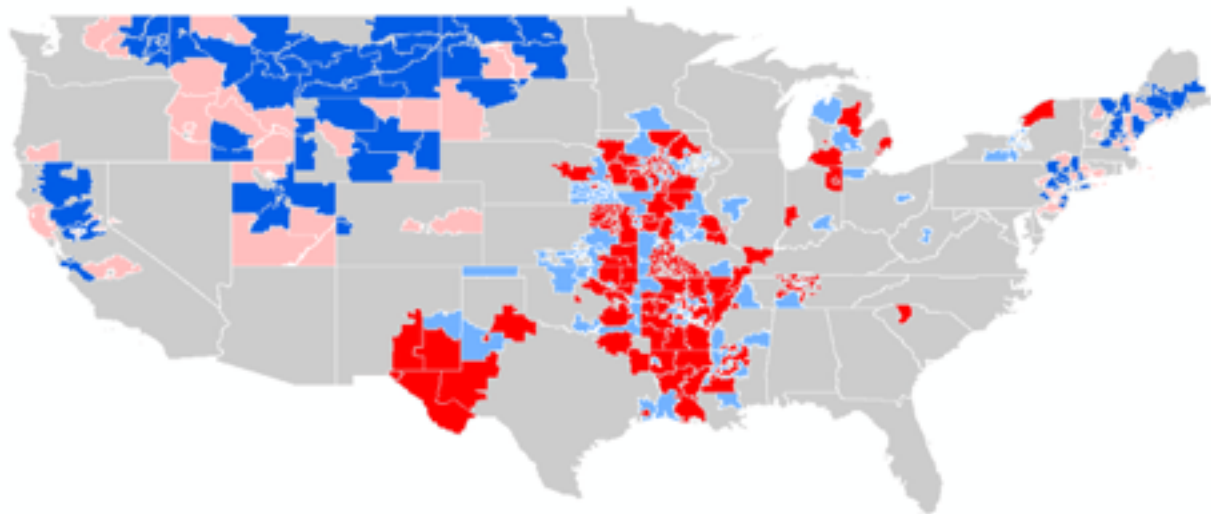


The map [1] merely shows the post processed data after thresholding. Notice that even with just about 48,000 observations the pattern recognition is difficult due in part to coincident points in space and perhaps support for quantitative methods of pattern recognition and discovery.



Map 1

Using the Moran's I Statistic for spatial autocorrelation and Local Indicators of Spatial Autocorrelation (LISA) for classification and local hypothesis testing we can examine both spatial dependency and spatial heterogeneity. Notice the large spatial clusters representative of spatial dependency in the north and south with smaller regimes in the northeast and west. These are high values surrounded by high values in the southwest and in the heartland. Also noticeable are the low indices surrounded by other low indices in the north, centered around Montana and on the coasts namely the NYC Metropolitan area and San Jose/SF area.







pseudo p-value < 0.05  
data: 862 (3-digit Zip Codes)

**Map 2**

There are also numerous more localized relationships not clear from this map. However, in addition to the smooth quality of the analysis as noted by high values surrounded by high values and low values surrounded by low values there are also some interesting rough qualities characteristic of spatial outliers or high values surrounded by low values and low values surrounded by high values. For example, Columbus OH, Ithaca NY, and Gassaway WV are all low values surrounded by high values - meaning writing at a more sophisticated level than its neighbors and meeting statistical significance.

By performing a spatial inner join with major cities, in this case cities with more than 300,000 people and the LISA classifications we can identify large cities and their sophistication in crafting Tweets. The following are the only cities that meet that criteria.

High, High	[n=77]		= El Paso, Oklahoma City, Omaha, Detroit,
Low, Low	[n=74]		Memphis
Low, High	[n=53]		= NYC & San Jose #nerds
High, Low	[n=55]		= Sacramento

El Paso, Oklahoma City, Omaha, Detroit, and Memphis all have statistically significant high values surrounded by high values (HH). NYC and San Jose are low values surrounded by low values (LL). Sacramento is a low value surrounded by otherwise high values (LH) and Wichita, Kansas City, Tulsa, and Nashville are all high flesch-kincaid indices surrounded by low flesch-kincaid indices (HL). These indices are inversely related with writing ability and linguistic standards; high values are low writing ability and vice versa low values are high writing ability. One might conclude among other things that NYC and San Jose write with high linguistic standards.

The LISA categories are statistically significant with a pseudo p-value < 0.05. Pseudo p-values are a computational approach to inference and proves to be a nice data reduction technique. Our original dataset of 3-digit zip codes is reduced from 862 observations to just 259 where **all other** observations are **not** statistically significant in the patterning of the kincaid index or just 30% of the original dataset.

Github: <https://github.com/rheimann>

Slideshare: <http://www.slideshare.net/rheimann04/big-social-data-the-spatial-turn-in-big-data>