IB Mathematics SL

Regularization of Irregular Verbs:
When can I use the words *swimmed* and *knowed* correctly?

Table of Contents

Regularization of Irregular Verbs:
When can I use the words *swimmed* and *knowed* correctly?

Introduction

I moved to the United States seven years ago, not knowing how to speak English. It did not take long to learn the basics and become fluent. After all, I was *thrust* into a full immersion (or maybe the word should be *thrusted?*). Anyway, the English language continues to baffle me at times. For every grammar rule I thought I have down, there are several exceptions. The English language is unique in that it always finds a way to break its own rules. No matter how much I practice writing the words *strenght* or *lenght*, for instance, I could never get it right the first time. I believe there's no other language in the world that delights to confuse its learners with four consecutive consonants to make one sound (maybe besides German).

Another one of these seemingly easy rules in the Modern English language is the morphology of the past tense, a very important grammar aspect to learn no matter in any language. In my own language, Tagalog, add *na-* to the beginning of a verb. In English, simply add *–ed* to the end of the verb. I got that! After I finish *walking*, I can say: I *walked.* But there's a catch. This *–ed* rule is only obeyed by regular verbs. Irregular verbs do not *go* by this rule, they *went* using their own unique rules or even by no rule at all. No wonder English is difficult for new learners! Personally, the irregular verbs took some getting used to. There's even a rumor going around that English might be harder to learn than Chinese, and this might be one of the reasons.

However, what is dynamic and fascinating about language, any language, is that it evolves along with the people who use it. New words get added all the time, words are

borrowed from other parts of the world, and the rules change to simplify language, making communication easier. Imagine how hard it would be to learn first learn English if *thou* have to learn it Shakespearian Old-English style.

What does this have to do with mathematics? Language, spoken and written *words*, seems to be one of those areas of knowledge that exist in a different world than mathematics. It is a generalization that an English person is not a science and math person. However, MIT and Harvard mathematicians Erez Lieberman and Jean-Baptiste Michel found a way to link these two worlds together. Generally, irregular verbs in English simplify over time. The *–ed* words *thrived*, not *throve*. So Lieberman and Baptist-Michel were able to "quantify the dynamics of language evolution…[by studying] the regularization of English verbs over the past 1,200 years" (Lieberman 713). Regularization means that an irregular verb eventually acquires an *–ed* past tense.

Aim of the Exploration

Since mathematicians Lieberman and Baptiste-Michel have figured out the mathematics, that the rate of regularization of irregular verbs is "inversely proportional to the square root of their usage frequency" ("Predicting… n.pag), this exploration will utilize their findings to find the number of years the words *know* and *swim* will regularize. I picked these two words because I occasionally find myself messing them up. So, I was wondering when my mistakes would be the correct way to say the past tense of these words.

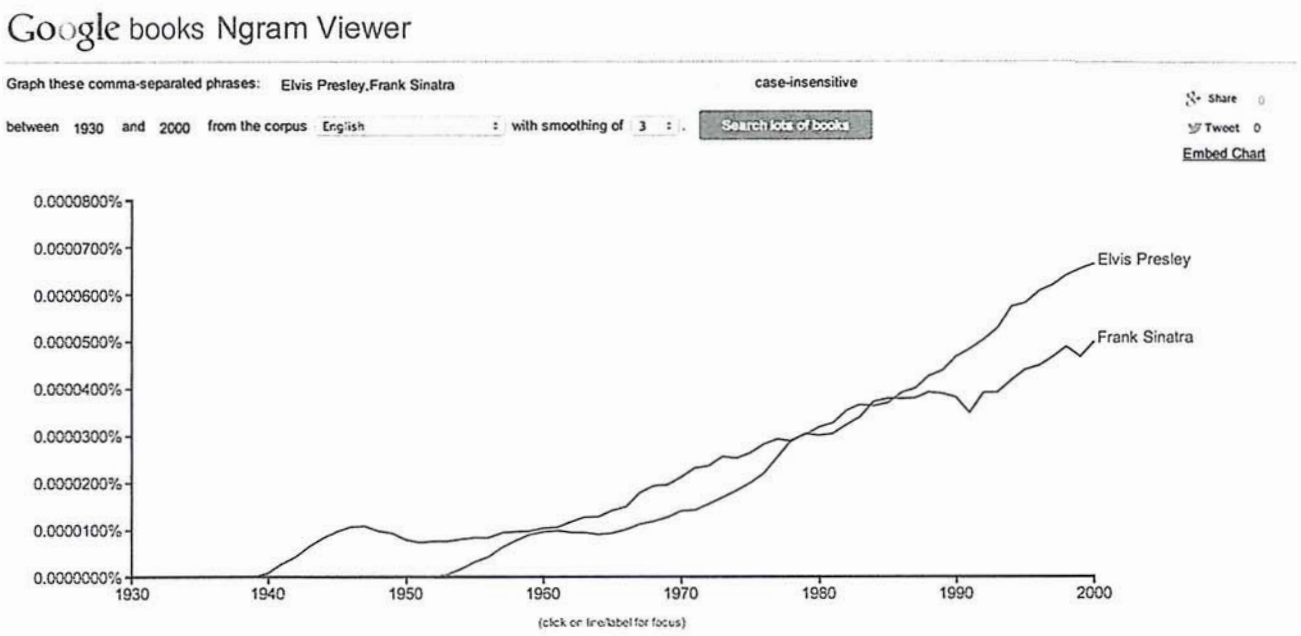
Rationale why this topic was chosen

The evolution of verbs seems like a tedious and uninteresting subject, but I beg to differ. I picked this topic because of two reasons. The first is the fact that although there

are only "less than 3% of modern verbs [that] are irregular, the *ten most common verbs* are all irregular (be, have, do, go, say, can, will, see, take, get)" (Lieberman 713). So, interestingly, for new speakers of the English language, some of the most necessary words are the hardest to keep in track of. From experience, I know that language barriers can be formidable challenges to communication.

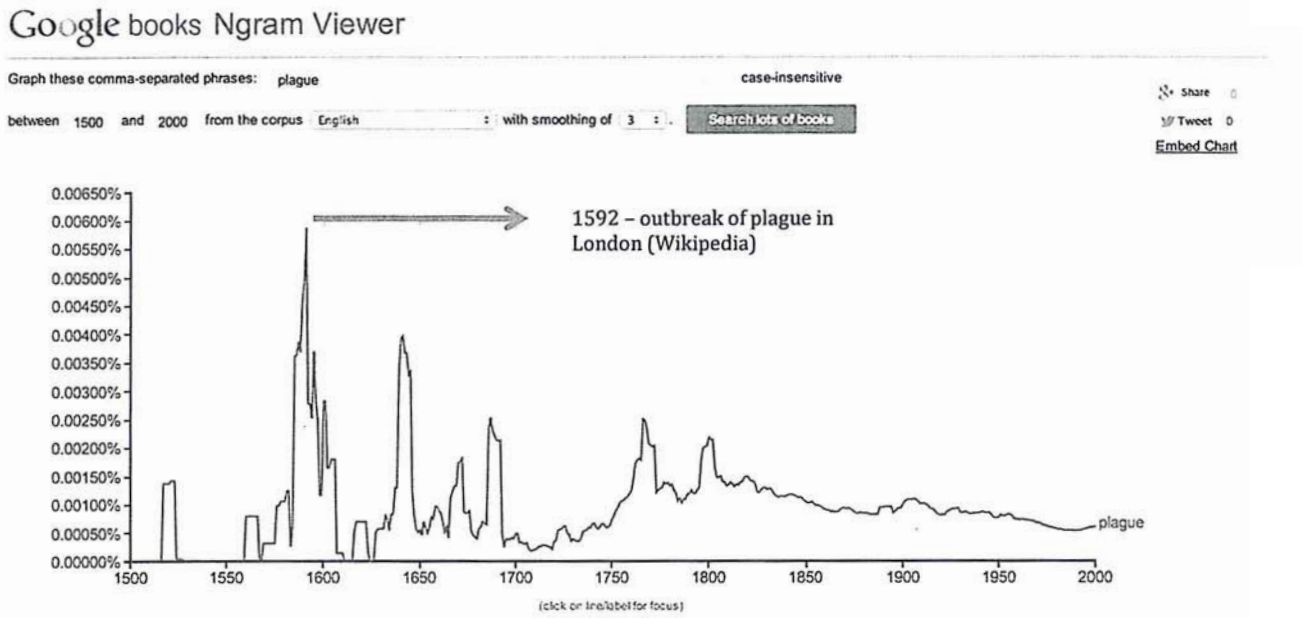
The second reason why I chose this topic is because of how amazing and extensive the method of how Lieberman and Baptiste-Michel studied the regularization of irregular verbs. They teamed with Google Books, who is trying to digitize all the books ever published and who have digitized 15 million books as of 2011 ("*What We Learned from 5 Million Books")*. These are 15 million books worth a span of human history and culture and about 12% of books ever published (Michel 1). Now, since it is impossible to be able to read all those 15 million books to analyze trends in history and culture over time ("*What We Learned from 5 Million Books")*, Google developed the Google Ngram, which they have made accessible to the public. For an entered ngram, a word or word phrase, "it displays a graph showing how [much] those phrases have occurred in a corpus of books over the selected years" ("Google Ngram Viewer"). Trends can be *mathematically* analyzed with these graphs. Using this method Lieberman and Baptiste- Michel study culture, or as they call it, **Culturonomics**, the "application of massive-scale data collection analysis to the study of human culture" ("*What We Learned from 5 Million Books")*. Of course, you can't just use books to fully analyze human culture. Newspapers, manuscripts, maps, artwork, songs, spoken legends, and others should be included, (Michel 181) but *millions* of books is still enough to give a piece of the picture of human culture.

The public accessible Google Ngram Viewer's capacity for years is 1500-2008. That's an impressive five centuries of books! As a demostration, I inputted *Frank Sinatra* and *Elvis Presley* into Google Ngram Viewer:



The vertical-axis shows that of all the bigrams (specific two-word combination) in Google's digitized books written in English, these are the percentages that are "Elvis Presley" and "Frank Sinatra." Obviously, the horizontal-axis shows the time in years. With this, it can be mathematically prove who is more popular in a given year!

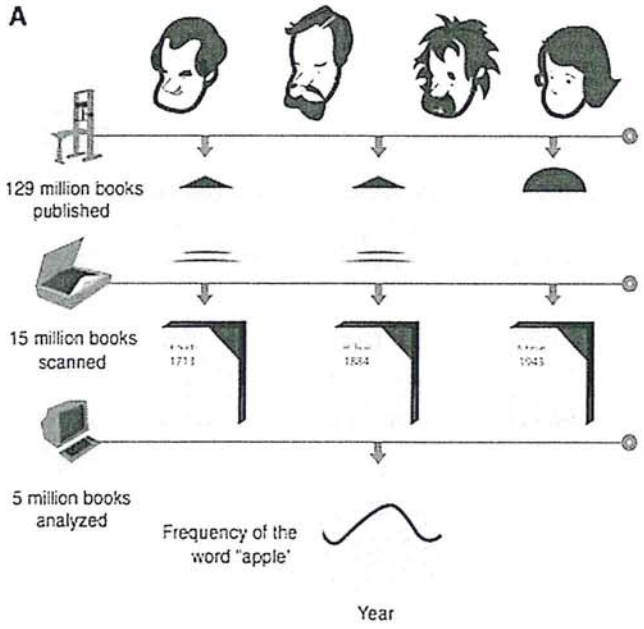
On a more serious note, I inputted *plague* to possibly track epidemics of plague in the past:



Using this device, Lieberman and Baptiste-Michel tracked the frequency of irregular verbs over time and found out when they regularized. They were able to quantify a part of the *evolution* of language. This visual and mathematical way of seeing the story of human culture and history is fascinating and offers a multitude of potential. The measurement of irregular verb's evolution is just one of them.

Lieberman and Baptiste Michel's Method and Findings

From Google Books's 15 million books -- about 12 percent of the approximately 129 million books published ever since the printing press (Michel 177) -- Lieberman and Michel created a body of 5,195,769 digitized books -- which is approximately 4% of the books ever published starting in the year 1500 (Michel 176). Although this is only a percentage of the books ever published, this sample size is enough follow trends in human culture. Out of these approximately 5 million books is more than 500 billion words in various languages; 361 billion of which are in English. The frequency of a specific word (an ngram) in a given year can then be found by dividing how much that word appears by the overall number of words used that year. Here is an image to summarize:



A. Culturomic analyses study millions of books □at once. (A) ***Top row****: Authors have been writing□ for millennia; ~129 mil-□lion book editions have□ been published since the□advent of the printing press□(upper left).* ***Second row****: □Libraries and publishing□ houses provide books to□Google for scanning (mid□dle left). Over 15 million□books have been digitized.□****Third row****: Each book is □associated with meta data.□Five million books are cho□sen for computational anal□ysis (bottom left).* ***Bottom□ row****: A culturomic time line□ shows the frequency of□ "apple" in English books□over time (1800–2000).*

Directly From "*Quantitative Analysis of Culture Using Millions of Digitized Books" page*
Jean Baptiste-Michel...

As one would expect with the increasing production, availability, and demand of books over time, more recent years have more words per year than past years. For instance, according to the body of data Lieberman and Baptiste-Michel compiled in the year 1800,

only 98 million words were used; in 1900, there were 1.8 billion; and in 2000, there were 11 billion (Michel 176).

Using this data, they tracked the evolution of 177 Old-English irregular verbs, finding that 145 stayed as irregular verbs in Middle English, and only 98 still remained irregular verbs today (Lieberman 1). They found that the rate an irregular verb regularizes is based on how much that word is used and the more an irregular verb is used in common everyday language, the longer it will take to regularize and eventually have an *–ed* past tense (Lieberman 1). They found the rate $R(\omega), or\ words\ per\ year,$ in which irregular verbs regularize to be "inversely proportional to the square root of their usage frequency ω" (Lieberman 714):

$$R(\omega) \propto \frac{1}{\sqrt{\omega}}$$

For instance, an irregular word used once in a given year will regularized at the rate of 1 word per year while an irregular word used 100 times in a given year will regularized at a rate of $\frac{1}{10}$ words per year.

$$R(100) \propto \frac{1}{\sqrt{100}} = \frac{1}{10} \qquad R(1) \propto \frac{1}{\sqrt{1}} = 1$$

This also means that "an irregular verb that is 100 times less frequent is regularizes 10 times as fast" (Lieberman 714) or an irregular verb used 100 times more frequent regularizes 10 times slower.

To be able to determine the regularization of irregular verbs in terms of years, I derived an alternative function, a function of half- life, $T_{\frac{1}{2}}(\omega)$, meaning half the time that an irregular verb will regularize:

The regularization of irregular verbs also means the decay of irregular verbs into regular verbs. Thus, I can use the exponential decay to find the half-life.

Exponential Decay Function $y = a\,(1-r)^t$ $a = initial\ amount$

$r = rate\ of\ decay$

$t = time$

To find when half of the initial amount will be present after the decay:

$$\frac{1}{2}a = a\,(1-r)^t$$

$$\frac{1}{2} = (1-r)^t$$

$$\ln\left(\frac{1}{2}\right) = t * ln(1-r)$$

$$T_{\frac{1}{2}} = \frac{\ln\left(\frac{1}{2}\right)}{\ln(1-r)}$$

Since I have a function for the rate $R(\omega) \propto \frac{1}{\sqrt{\omega}}$, this can be substituted for r.

$$T_{\frac{1}{2}} = \frac{\ln\left(\frac{1}{2}\right)}{\ln\left(1-\frac{1}{\sqrt{w}}\right)}$$

This function is what I will use to find when know and swim will regularize.

*When will **know** and **swim** regularize?*

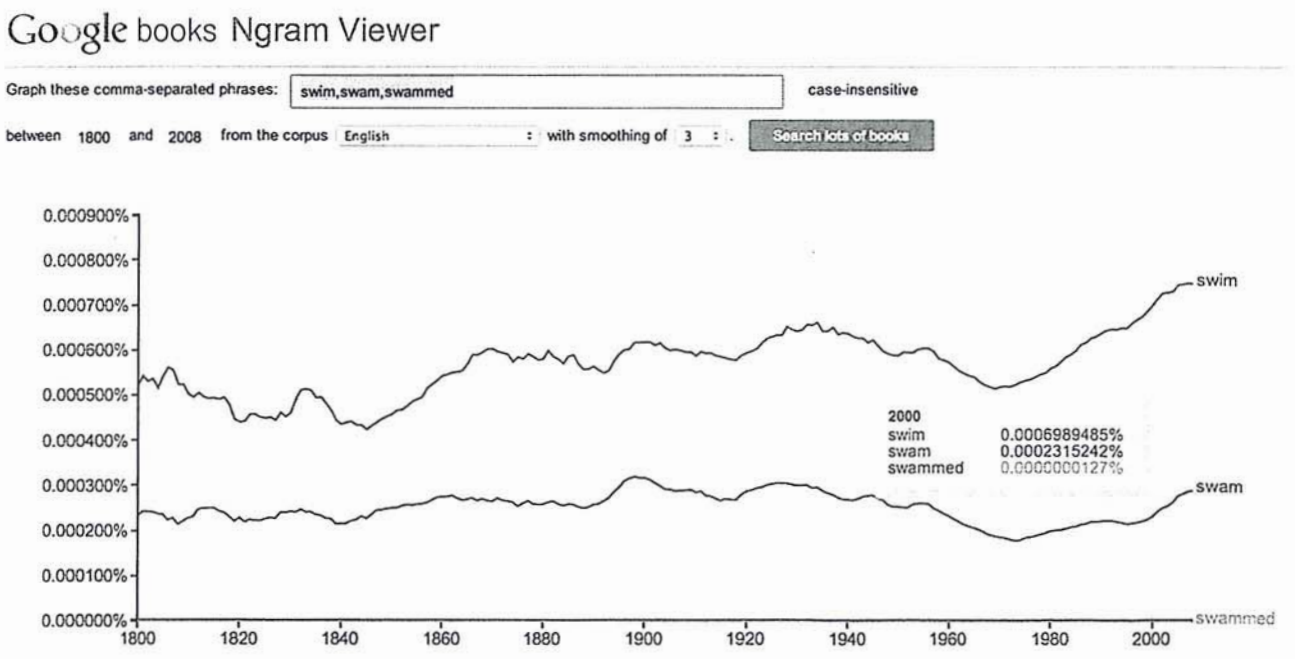
Known Information (all of which are estimates):

- Lieberman's and Baptiste-Michel's findings are from about 4% of books ever published.
- In this 4% of books, the resulting body of data has 500 billion words.
- In this 500 billion words, only 361 billion is in English. (72.2 % of 500 billion)
- In the indivudual year of 2000, the body of data yields 11 billion words.

From these known information, I assume/reason (all estimates):

- In 2000, the words out of 11 billion that is in English is 7.942 billion.
  - Calculation: (11 billion x 0.722)
- **7.942 billion** is the total number of English words used in 2000
  - Disclaimer: 7.942 billion words that exist in the year 2000 only for a fraction of the books published that year.

Swim



Swim occurs 0.0006989485% in 2000.

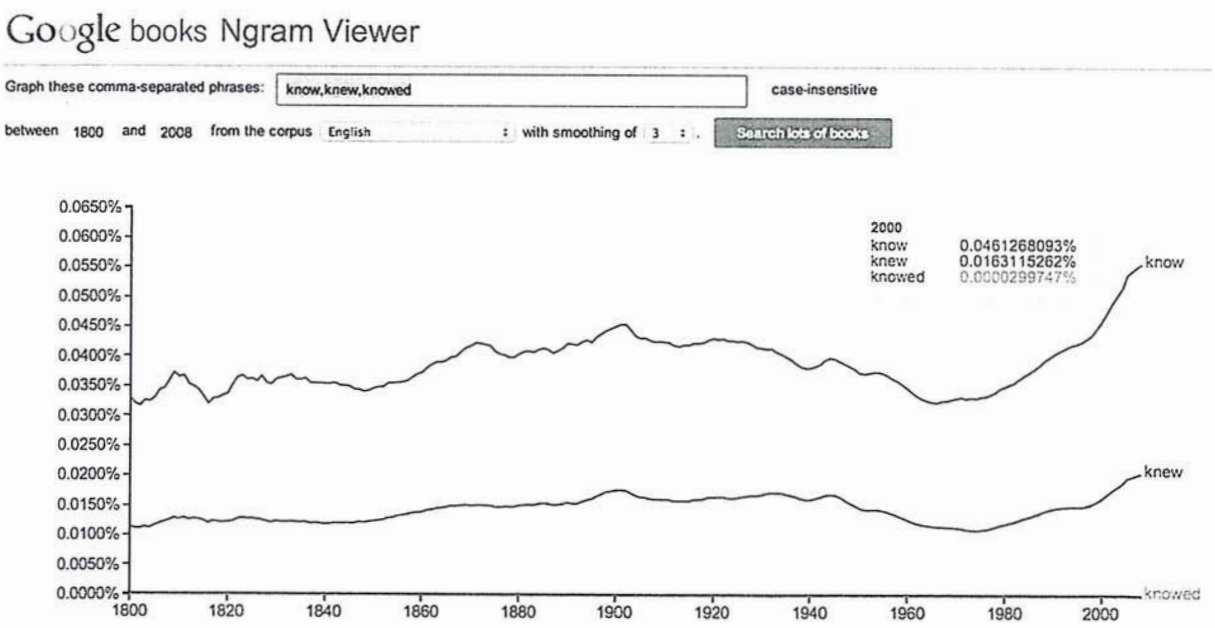
$$0.000006989485 * 7\,942\,000\,000 \text{ words} = 55\,510.48987 \text{ words}$$

Assuming the word *swim* has a frequency of 55,510 in 2000:

$$T_{\frac{1}{2}} = \frac{\ln\left(\frac{1}{2}\right)}{\ln\left(1 - \frac{1}{\sqrt{55,510}}\right)} = 162.962$$

The word *swim* has a half-life of about 163 years, meaning that in about 326 years, *swimmed* would be used instead of *swam*.

Know



Know occurs 0.0461268093% in 2000.

$$0.000461268093 * 7\,942\,000\,000 \text{ words} = 3\,663\,391.19461 \text{ words}$$

Assuming the word *know* has a frequency of 3,663,391 in 2000.

$$T_{\frac{1}{2}} = \frac{\ln\left(\frac{1}{2}\right)}{\ln\left(1 - \frac{1}{\sqrt{3,663,391}}\right)} = 1,326.336$$

The word *know* has a half-life of about 1, 326 years, meaning that in about 2, 653 years, *knowed* would be used instead of *knew*.

Conclusion

Since *know* is one of the most common irregular verbs, it will be more difficult to eradicated its irregular past tense from the English lexicon than *swim*. It will take about 400 years for me to be able to have *swimmed* and about 3,000 years to be able to have *knowed*. I guess I'm stuck to just using *swam* and *knew*!

This investigation was fascinating, not due to the aim of the exploration (finding the life-expectancy of *knew* and *swam),* but because of the potential of knowledge from a digitizing massive scales of data to analyze human culture and the fact that cooperation between language with mathematics was able to make this happen. It also shows two dichotomies about the of human culture and communication. First is the dynamic mutability of language and culture to make communication more efficient and the second is language and culture's predictability and scientific nature.

In doing this Internal Assessment, I was able to see for myself how mathematics is used beyond doing problems in textbooks. Mathematicians and scientists all over the world have applied math to facets of nature and living, and not just in the physical world, in order to more fully explain the world and the people and creatures within it.

Works Cited

Aiden, Erez, and Jean-Baptiste Michel. *Uncharted*. New York: Penguin, 2013. Print.

The Google Ngram Viewer Team. "Google Ngram Viewer." *Google Ngram Viewer*. Google Research, n.d. Web. 21 Mar. 2014.

Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin A. Nowak. *Letters: Quantifying the Evolutionary Dynamics of Language*. N.p.: Nature: International Weekly Journal of Science, 11 Oct. 2007. PDF.

Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin A. Nowak. *Supplementary Information: Quantifying the Evolutionary Dynamics of Language*. N.p.: Nature: International Weekly Journal of Science, 11 Oct. 2007. PDF.

The Mathematics of History. Perf. Jean-Baptiste Michel. *TED*. TED2012, Feb. 2012. Web. 10 Mar. 2014.

Michel, Jean-Baptiste, Martin A. Nowak, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pinkett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, and Erez Liberman Aiden. *Supporting Online Material for Quantitative Analysis of Culture Using Millions of Digitized Books*. N.p.: SCIENCE, 11 Mar. 2011. PDF.

Michel, Jean-Baptiste, Martin A. Nowak, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pinkett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, and Erez Liberman Aiden. *Quantitative Analysis of Culture Using Millions of Digitized Books*. N.p.: SCIENCE, 14 Jan. 2011. PDF.

"Predicting the Future of the past Tense Mathematicians Apply Evolutionary Models to Language." *MIT News Office*. N.p., 15 Oct. 2007. Web. 01 Apr. 2014.

What We Learned from 5 Million Books. Perf. Jean-Baptiste Michel and Erez Lieberman Aiden. *TED*. TEDxBoston 2011, July 2011. Web. 10 Mar. 2014.