ISLE 2, Boston • Susanne Stegmayr & Daphné Kerremans (LMU Munich)

Neologisms on the Internet

The NeoCrawler
Overview

1. The Problem
2. The Web-as-Corpus
3. The NeoCrawler
   • architecture
   • manual search and search results
   • automatic searches and the Google Research Program
   • web interface and linguistic processing
4. Applied NeoCrawling
1. The Problem

**cherpumple**

- **n = 368**

**halfalogue**

- **n = 167**
2. The Web-as-Corpus

- The WaCky approach (Sharoff, Baroni, Bernardini, Kilgarriff et al.)

- Downloadable crawlers:
  - KWICFinder (Fletcher 2001, www.kwicfinder.com)

- On-demand crawlers:
  - Linguistic Search Engine (Kilgarriff 2003)
The WaCky approach (Sharoff, Baroni, Bernardini, Kilgarriff et al.)

Downloadable crawlers:

- **KWICFinder** (Fletcher 2001, [www.kwicfinder.com](http://www.kwicfinder.com))
- **GlossaNet 2** (Fairon, Macé and Naets 2008, [http://glossa.fltr.ucl.ac.be](http://glossa.fltr.ucl.ac.be))

On-demand crawlers:

- **Linguistic Search Engine** (Kilgarriff 2003)
- **WebCorp and WebCorp Linguistic Search Engine** (1998, 2005 RDUES Birmingham City University, [www.webcorp.ac.co.uk](http://www.webcorp.ac.co.uk))

2. The Web-as-Corpus

Susanne Stegmayr & Daphné Kerremans
Discoverer

Search the Web for new Words

Database of Neologisms

Observer

Search for Websites with the Neologism

Database of Websites
Susanne Stegmayr & Daphné Kerremans

6
http://www.google.com/advanced_search?q=halfalogue&hl=en&client=safari&rls=en&prmd=
http://www.google.com/search?q=halfalogue&hl=en&rls=en&num=10&lr=lang_en&tbs=qdr:m
Normally, halfalogues don’t bug me all that much…but there’s a doozy here at the gate. A guy behind me with a raspy voice is talking about people he’s considering hiring. While I’m sort of curious to know what his company does, and what positions he’s hiring for, right now I wish he’d just stop talking.

He stopped.

No, there’s more. He said goodbye, but then he started up again. Now he’s talking about clearances…

Now he’s done. Really done. Phew.
Halflologue at Gate C-6

Post by Joe Pioza

On a sunny day in the middle of the afternoon, I was driving through the city. Suddenly, a man in a white shirt and black pants came running out of a building right in front of me. He was shouting and waving his arms. I was shocked but managed to stop the car just in time.

I got out of the car and went over to him. He was breathing heavily and looked very agitated. He told me that he was a police officer and had just received a call about a crime that had been committed in a nearby house. He needed my help to get there as quickly as possible.

I immediately got back into my car and followed him to the scene of the crime. When we arrived, I could see that the house was damaged and there were signs of a struggle. The police officer explained that the suspects were still at large and that we needed to act quickly to catch them.

Together, we decided to search the house for any evidence that could help us identify the suspects. We checked every room, but unfortunately, we didn’t find anything that could help us.

After searching for a while, we decided to leave and report back to the police station. On the way, we passed by a diner and decided to have a quick meal before returning.

We ordered our orders and waited for our food. While we were eating, a group of people came into the diner and sat down at a table next to us. They started talking loudly and laughing. I couldn’t help but overhear their conversation.

It was clear that they were some sort of criminal gang. They were discussing plans and looking around nervously. I knew that we had to act fast.

After we finished our meal, we went back to the police station to report our findings. The police officer thanked us for our help and promised to follow up on the case.

As we were leaving, I noticed that one of the suspects was walking towards us. I quickly hid behind a column and watched as he walked away. I knew that I had to report this to the police officer immediately.

We went back to the police station as quickly as possible and reported our findings. The police officer was impressed with our help and thanked us again.

In the end, we were able to catch the suspects and bring them to justice. It was a difficult and dangerous situation, but I’m glad that we were able to help.

I hope that this experience will help others to understand the importance of being vigilant and taking action when necessary.
TIME has come today

Halfologue at Gate C-6

Posted on: May 6, 2011
by: JoePaica

Normally, halfologues don’t bug me all that much, but there’s a doozy here at the gate. A guy behind me with a nasally voice is talking about people he’s considering hiring. While I’m sort of curious to know what his company does, and what positions he’s hiring for, right now I wish he’d just stop talking.

He stopped.

No, there’s more. He said goodbye, but then he started up again. Now he’s talking about clearances. Now he’s done. Really done. Phew.

This entry was posted in Uncategorized. Bookmark the permalink.
A
- avatar
- administrative
- affincion
- AffA
- agile business continuity
- alhambred
- anagramme
- aquapunk
- assault weapon
- avatar

(n.) a pale white person or object with a mental handicap (also adj.):

1. (n.) A person who plays a hero in the gaming world but is a real loser in the real world.
2. (n.) A person who is obsessed with the characters, worlds and animals of James Cameron's new science fiction film Avatar.

B
- basic scooping
- business
- backwater
- bearfish
- bearishness
- blog
- blogging
- blogs
- blogs
- blu-link
- blog
- boycott
-browser
- browser
- buzz
- buzz marketing

(n.) a posture around the hips and flanks to narrow the waist.

(n.) A person in the financial service industry who grows rich despite the continued impoverishment of those who depend on their services, and despite their apparent inability to succeed in business without constant government assistance.

1. (n.) A short blog entry
2. (n.) A form of address for fellow bloggers
3. (n.) E-mail update information system
   (n.) To update your blog
4. (n.) An advertising strategy that entices people to buy a product and get one for free.
5. (n.) A straight man who is very close to another man without being homosexual.

(n.) A Burke which is made into a bikini

(n.) Word-of-mouth marketing

C
- carbynitrile
- people
- Cy-Fa
- cheap and cheerful
- chopspumple
- coffee
- shopping
- check
- stuggle
- right computing
- cloud generator
- co-op

(n.) dirty, gray "slloa" of slushy snow and petrified ice that falls of oto automobile mud flaps all winter in Wisconsin (coined in the US)

(n.) An apple pie baked inside a spice cake, a pumpkin pie baked inside a yellow cake, and a cherry pie baked inside a white cake, all stacked on top of each other and covered in cream cheese icing.

(n.) Techn. women's soccer: blend of "jocks" and "chicks," girls who look pretty off the field but play just as hard as the boys.

(n.) Atmospheric global warming, history's biggest scam.

(n.) System of computing in which services and storage are provided over the Internet.
Progress of halfologue

<table>
<thead>
<tr>
<th>Process ID</th>
<th>Processed by</th>
<th>Time Restriction</th>
<th>Search String</th>
<th>No. of Pages</th>
<th>No. of Tokens</th>
<th>Date</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>0917</td>
<td>necrawler</td>
<td>w</td>
<td>halfologue</td>
<td>1</td>
<td>1</td>
<td>2010-08-09</td>
<td>summary details</td>
</tr>
<tr>
<td>9196</td>
<td>necrawler</td>
<td>w</td>
<td>halfologue</td>
<td>2</td>
<td>4</td>
<td>2010-08-16</td>
<td>summary details</td>
</tr>
<tr>
<td>9475</td>
<td>necrawler</td>
<td>w</td>
<td>halfologue</td>
<td>3</td>
<td>4</td>
<td>2010-08-23</td>
<td>summary details</td>
</tr>
<tr>
<td>11465</td>
<td>necrawler</td>
<td>m</td>
<td>halfologue</td>
<td>95</td>
<td>241</td>
<td>2010-10-05</td>
<td>summary details</td>
</tr>
<tr>
<td>11747</td>
<td>necrawler</td>
<td>m</td>
<td>halfologue</td>
<td>87</td>
<td>255</td>
<td>2010-10-12</td>
<td>summary details</td>
</tr>
<tr>
<td>12029</td>
<td>necrawler</td>
<td>m</td>
<td>halfologue</td>
<td>98</td>
<td>237</td>
<td>2010-10-19</td>
<td>summary details</td>
</tr>
<tr>
<td>12311</td>
<td>necrawler</td>
<td>m</td>
<td>halfologue</td>
<td>51</td>
<td>140</td>
<td>2010-10-26</td>
<td>summary details</td>
</tr>
<tr>
<td>12593</td>
<td>necrawler</td>
<td>m</td>
<td>halfologue</td>
<td>47</td>
<td>135</td>
<td>2010-11-02</td>
<td>summary details</td>
</tr>
<tr>
<td>12906</td>
<td>necrawler</td>
<td>m</td>
<td>halfologue</td>
<td>59</td>
<td>151</td>
<td>2010-11-09</td>
<td>summary details</td>
</tr>
</tbody>
</table>

x axis: calendar week (2011)
y axis: number of tokens per week
x axis: calendar week (2011)

y axis: number of tokens per week

<table>
<thead>
<tr>
<th>Process ID</th>
<th>Processed by</th>
<th>Time Restriction</th>
<th>Search String</th>
<th>No. of Pages</th>
<th>No. of Tokens</th>
<th>Date</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>6917</td>
<td>neocrawler</td>
<td>w</td>
<td>halfalogue</td>
<td>1</td>
<td>1</td>
<td>2010-06-09</td>
<td>summary details</td>
</tr>
<tr>
<td>9196</td>
<td>neocrawler</td>
<td>w</td>
<td>halfalogue</td>
<td>2</td>
<td>4</td>
<td>2010-08-16</td>
<td>summary details</td>
</tr>
<tr>
<td>9475</td>
<td>neocrawler</td>
<td>w</td>
<td>halfalogue</td>
<td>3</td>
<td>4</td>
<td>2010-06-23</td>
<td>summary details</td>
</tr>
<tr>
<td>11465</td>
<td>neocrawler</td>
<td>m</td>
<td>halfalogue</td>
<td>56</td>
<td>237</td>
<td>2010-10-06</td>
<td>summary details</td>
</tr>
<tr>
<td>11747</td>
<td>neocrawler</td>
<td>m</td>
<td>halfalogue</td>
<td>50</td>
<td>151</td>
<td>2010-10-12</td>
<td>summary details</td>
</tr>
<tr>
<td>12020</td>
<td>neocrawler</td>
<td>m</td>
<td>halfalogue</td>
<td>56</td>
<td>140</td>
<td>2010-10-26</td>
<td>summary details</td>
</tr>
<tr>
<td>12311</td>
<td>neocrawler</td>
<td>m</td>
<td>halfalogue</td>
<td>47</td>
<td>135</td>
<td>2010-11-02</td>
<td>summary details</td>
</tr>
<tr>
<td>12593</td>
<td>neocrawler</td>
<td>m</td>
<td>halfalogue</td>
<td>50</td>
<td>151</td>
<td>2010-11-09</td>
<td>summary details</td>
</tr>
</tbody>
</table>
### Summary of Process No. 12906

**Lemma:** hallogue

**Date:** 2010-11-09, 08:11
**Restriction:** m
**Pages:** 59
**Tokens:** 151
**By:** neocrawler

**Source:** Association for Psychological Science - hallogue

<table>
<thead>
<tr>
<th>#</th>
<th>Tokens</th>
<th>Title</th>
<th>Co-text</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>Association for Psychological Science - hallogue</td>
<td>Association for Psychological Science - hallogue Home About APS Contact -</td>
<td><img src="https://example.com" alt="field of discourse" /> <img src="https://example.com" alt="type of source" /> <img src="https://example.com" alt="authorship" /> <img src="https://example.com" alt="data entry" /> <img src="https://example.com" alt="blacklist site" /></td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>Psychological Science - Coming Backwards and Setting Trends - Association for Psychological Science - hallogue</td>
<td>source for the latest psychological research. Previous Observation &gt;&gt; &lt;&lt; Next Observation Psychological Science: Coming Backwards and Setting Trends Posted by &gt;&gt; By &gt;&gt; When we hear a hallogue, our brain tries to fill in the other half of the conversation. Today's Express article &quot;Stuck In Your Ear: Meta-Hallogues Aren't Just Rude -- They Are Disturbingly Irresistible!&quot; (page 10) highlights not only a recent psychological science article, but the term &quot;hallogue&quot; used by the author, Lauren Embree, PhD candidate in psychology at Cornell University, coined the catchy term to describe the scenario in which only one side of a cell phone conversation.</td>
<td><img src="https://example.com" alt="field of discourse" /> <img src="https://example.com" alt="type of source" /> <img src="https://example.com" alt="authorship" /> <img src="https://example.com" alt="data entry" /> <img src="https://example.com" alt="blacklist site" /></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>Small the Coffee: Hallogue is not better than no log at all - Life - The Charleston Gazette - West Virginia News and Sports -</td>
<td>Small the Coffee: Hallogue is not better than no log at all - Life - The Charleston Gazette - West Virginia News and Sports -</td>
<td><img src="https://example.com" alt="field of discourse" /> <img src="https://example.com" alt="type of source" /> <img src="https://example.com" alt="authorship" /> <img src="https://example.com" alt="data entry" /> <img src="https://example.com" alt="blacklist site" /></td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>Small the Coffee: Hallogue is not better than no log at all - Time -</td>
<td>Small the Coffee: Hallogue is not better than no log at all - Time -</td>
<td><img src="https://example.com" alt="field of discourse" /> <img src="https://example.com" alt="type of source" /> <img src="https://example.com" alt="authorship" /> <img src="https://example.com" alt="data entry" /> <img src="https://example.com" alt="blacklist site" /></td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>gulfnews: Sound affects</td>
<td>gulfnews: Sound affects</td>
<td><img src="https://example.com" alt="field of discourse" /> <img src="https://example.com" alt="type of source" /> <img src="https://example.com" alt="authorship" /> <img src="https://example.com" alt="data entry" /> <img src="https://example.com" alt="blacklist site" /></td>
</tr>
</tbody>
</table>
Starter of the Discoverer-Module

Please select the action you would like the script to perform.

**Crawling:**

- [ ] Crawl one or more URLs for neologisms
- [x] Use Google Blog Search to retrieve a list of blog URLs and process the pages
- [ ] Upload file from on local disk
- [ ] Select multiple files from server HDD

**Resource Managing:**

- [ ] Build trigram statistics for phonotactic check
- [ ] Build automata from dictionary resource(s)
- [ ] Purge all temporary lists (warning: all previously found types will be deleted from the variable LEX lists and MDICs. You can revert backups if this was done unintentionally)
- [ ] Undo last changes to temporary lists (revert to last backup)
- [ ] Re-compile variable .LEX files to temporary MDICs

**Settings:**

- Display stripped text and frequency list:
- Find hyphenated compounds:
- Only display candidates that have quality...
Starter of the Discoverer-Module

Please select the action you would like the script to perform.

Crawling:

- All
- Politics
- US
- World
- Business
- Technology
- Video Games
- Science
- Entertainment
- Movies
- Television
- Sports

Go!

Settings:
- Display stripped text and frequency list: 
- Find hyphenated compounds: 

http://neocrawler.gwi.uni-muenchen.de/index.php?abfrage=discoverer
Opened 10 LookupMD instances successfully
Crawling Google Blog Search to retrieve 10 URLs...
retrieved 10 Blog URLs from Google Blog Search.
List of URLs:
3.: "http://dailycapitalist.com/2011/05/31/case-shiller-down-5-1-q1-2011-from-q1-2010/"
4.: "http://www.balloon-juice.com/2011/05/31/base-how-low-can-you-go/"
5.: "http://www.cultofmac.com/ios5-eliminate-the-need-for-itunes-sync-could-icloud-include-facebook-style-apps-speculation/98669"
6.: "http://www.cultofmac.com/"
7.: "http://politicalbetting.com/index.php/archives/2011/06/05/is-inverclyde-a-certain-labour-hold/"
8.: "http://politicalbetting.com/"
10.: "http://www.ritholtz.com/blog/"
Starting filtering process of retrieved URLs...

Preprocessing Text #1
Encoding is UTF-8, attempting full punctuation stripping method
Filtering out stopwords and strings with length < 3... Done.
Filtering out known names... Done.
Filtering out known tokens from Dictionary... Done.


☐ Show Frequency list of URL "http://la.curbed.com/archives/2011/05/shiller_time_6.php", text number 1:

No tokens remained after filtering.

Preprocessing Text #2
Warning, this source is not encoded in UTF-8, errors might occur...
Encoding is "utf-8", using compatibility punctuation stripping method
Filtering out stopwords and strings with length < 3... Done.
Filtering out known names... Done.
Filtering out known tokens from Dictionary... Done.
Filtering out known tokens from previous program runs... Done.
Constructing Candidate objects... Done.
Preprocessing Text #10
Encoding is UTF-8, attempting full punctuation stripping method
Filtering out stopwords and strings with length < 3... Done.
Filtering out known names... Done.
Filtering out known tokens from Dictionary... Done.
Filtering out known tokens from previous program runs... Done.
Constructing Candidate objects... Done.

☐ Show stripped Text of URL "http://www.ritholtz.com/blog/", text number 10:

☐ Show Frequency list of URL "http://www.ritholtz.com/blog/", text number 10:

Adding unknown types to List...
took 167 seconds
26101 Tokens processed, 115 unknown types found.

List of unknown types
Display Explanation: ☐

List of all encountered unknown types:
☐ Show unknown types in Text #2:
☐ Show unknown types in Text #3:
☐ Show unknown types in Text #4:
☐ Show unknown types in Text #5:
Adding unknown types to List...

Preprocessing Text #3
Encoding is UTF-8, attempting full punctuation stripping method
Filtering out stopwords and strings with length < 3... Done.
Filtering out known names... Done.
Filtering out known tokens from Dictionary... Done.
Filtering out known tokens from previous program runs... Done.
Constructing Candidate objects... Done.

Show stripped Text of URL: "http://dailycapitalist.com/2011/05/31/case-shiller-down-5-1-q1-2011-from-q1-2010/", text number 3:

Show Frequency list of URL
"http://dailycapitalist.com/2011/05/31/case-shiller-down-5-1-q1-2011-from-q1-2010/", text number 3:

the --> 81
of --> 47
and --> 38
to --> 36
is --> 35
The --> 29
that --> 23
in --> 23
at --> 19
will --> 16
be --> 15
housing --> 14
<table>
<thead>
<tr>
<th>ID</th>
<th>Term</th>
<th>Type Quality</th>
<th>Avg. Trigrams</th>
<th>Extended Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>house-to-key</td>
<td>Ignore</td>
<td>10</td>
<td>A house that had no equivalent house-to-key. The high-end rental market [...].</td>
</tr>
<tr>
<td>2.</td>
<td>urbanseaman</td>
<td>Ignore</td>
<td>10</td>
<td>It did not come true urbanseaman on Toldeo El Gato [...].</td>
</tr>
<tr>
<td>6.</td>
<td>well-picked</td>
<td>Ignore</td>
<td>10</td>
<td>A limited resource. And a well-picked location will appreciate [...].</td>
</tr>
<tr>
<td>7.</td>
<td>TBOGG</td>
<td>Ignore</td>
<td>10</td>
<td>Talk Left. Talking Points. Menu TBOGG. The Beacon Tribune. The [...].</td>
</tr>
<tr>
<td>8.</td>
<td>nonbublikation</td>
<td>Ignore</td>
<td>10</td>
<td>Residential construction boom, based on nonbublikation historical rates of return [...].</td>
</tr>
<tr>
<td>10.</td>
<td>hostcessaroood</td>
<td>Ignore</td>
<td>10</td>
<td>She's a lawyer. He's a lawyer. Cassa Business. Mondello. Neither is going to [...].</td>
</tr>
<tr>
<td>14.</td>
<td>white-nose</td>
<td>Ignore</td>
<td>10</td>
<td>of the UK housing market. What’s his name in the quote. snake-oil [...].</td>
</tr>
<tr>
<td>15.</td>
<td>lawness</td>
<td>Ignore</td>
<td>10</td>
<td>It’s a garage. It’s a garage. Lower-middle-class subdivision. It’s a college [...].</td>
</tr>
<tr>
<td>17.</td>
<td>Wingularity</td>
<td>Ignore</td>
<td>10</td>
<td>The War on Drugs. The Wingularity. This is not where [...].</td>
</tr>
<tr>
<td>18.</td>
<td>koolaid</td>
<td>Ignore</td>
<td>10</td>
<td>Much of the “American Dream” koolaid. There is more than [...].</td>
</tr>
<tr>
<td>19.</td>
<td>Irick Lifter</td>
<td>Ignore</td>
<td>10</td>
<td>Late Night Weekend Thread. Competitive Exhibit. Ane. All. All. Stumped [...].</td>
</tr>
</tbody>
</table>
the past six weeks, his *tactically_allocated* accounts have moved from...
<table>
<thead>
<tr>
<th>Rank</th>
<th>Word</th>
<th>Candidate</th>
<th>Garbage</th>
<th>Ignore</th>
<th>Type</th>
<th>Quality</th>
<th>Freq</th>
<th>Trigram</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>streets-you</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>Type</td>
<td>Quality</td>
<td>Freq</td>
<td>Trigram 91.39%</td>
<td>[...] there is blood in the streets-you can not time the [...]</td>
</tr>
<tr>
<td>2.</td>
<td>Stagflationary</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>Type</td>
<td>Quality</td>
<td>Freq</td>
<td>Trigram 91.39%</td>
<td>Sinking Manufacturing Is A Stagflationary Trend</td>
</tr>
<tr>
<td>3.</td>
<td>monetary-induced</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>Type</td>
<td>Quality</td>
<td>Freq</td>
<td>Trigram 91.39%</td>
<td>[...] are related to positive economic growth, rising wages, and unfortunately monetary-induced price inflation. I don’t [...]</td>
</tr>
<tr>
<td>4.</td>
<td>Econophile</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>Type</td>
<td>Quality</td>
<td>Freq</td>
<td>Trigram 91.39%</td>
<td>[...] Capitalist on Fox Business News Econophile On Fox Business News [...]</td>
</tr>
<tr>
<td>5.</td>
<td>neg-equity</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>Type</td>
<td>Quality</td>
<td>Freq</td>
<td>Trigram 91.39%</td>
<td>[...] methodologies used in the Zillow/Core-Logic neg-equity estimates only use [...]</td>
</tr>
<tr>
<td>6.</td>
<td>Destruction-Parts</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>Type</td>
<td>Quality</td>
<td>Freq</td>
<td>Trigram 91.39%</td>
<td>[...] Come The Economics of Mass Destruction-Parts I and II Archived [...]</td>
</tr>
<tr>
<td>7.</td>
<td>median-sized</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>Type</td>
<td>Quality</td>
<td>Freq</td>
<td>Trigram 91.39%</td>
<td>[...] of gold can buy a median-sized home in the US. [...]</td>
</tr>
<tr>
<td>8.</td>
<td>CoreLogi</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>Type</td>
<td>Quality</td>
<td>Freq</td>
<td>Trigram 91.39%</td>
<td>[...] in their homes. At this point, CoreLogic would put it at [...]</td>
</tr>
<tr>
<td>9.</td>
<td>20-City</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>Type</td>
<td>Quality</td>
<td>Freq</td>
<td>Trigram 91.39%</td>
<td>[...] the nation. The National Index, the 20-City Composite and 12 MSAs [...]</td>
</tr>
<tr>
<td>Neologism-Suggester Tool</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24. : transit-poor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Candidate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type Quality: 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg. Trigram Freq: 3.21754%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[...] limits their interest in the transit-poor areas where single family</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Garbage Ignore</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| 25. : snake-oiling      |
| Candidate               |
| Type Quality: 10        |
| Avg. Trigram Freq: 3.20926% |
| [...] market. What’s-his-nose in the quote snake-oiling, “BUY! BUY |
| Garbage Ignore          |

| 26. : never-lived-in    |
| Candidate               |
| Type Quality: 10        |
| Avg. Trigram Freq: 1.78209% |
| [...] close friends bought a million never-lived-in foreclosure |
| Garbage Ignore          |

| 27. : Galtian           |
| Candidate               |
| Type Quality: 10        |
| Avg. Trigram Freq: 2.94458% |
| [...] situation will improve until our Galtian overlords are made to |
| Garbage Ignore          |

| 28. : hoocoodanode      |
| Candidate               |
| Type Quality: 10        |
| Avg. Trigram Freq: 2.82051% |
| [...] For Conservatives Green Balloons Hoekstroika hoocoodanode Can |
| Garbage Ignore          |

| 29. : end-ups           |
| Candidate               |
| Type Quality: 10        |
| Avg. Trigram Freq: 2.53102% |
| [...] a home always and necessarily end-ups with “nothing” (quote |
| Garbage Ignore          |

| 30. : Crazification     |
| Candidate               |
| Type Quality: 10        |
| Avg. Trigram Freq: 2.37612% |
| [...] Cooking Corrections Crab Bucket Politics Crazification Factor |
| Garbage Ignore          |

| 31. : Spectering        |
| Candidate               |
| Type Quality: 10        |
| Avg. Trigram Freq: 2.30362% |
| [...] Village is Missing its Idiot Spectering Sports Starbursts Sturve |
| Garbage Ignore          |

| 32. : dollar-valued     |
| Candidate               |
| Type Quality: 10        |
| Avg. Trigram Freq: 1.78209% |
| [...] close friends bought a million dollar-valued never-lived-in foreclosure |
| Garbage Ignore          |
| 24. : transit-poor  | Candidate | Garbage | Ignore | Type Quality  | Avg. Trigram Freq | context... limits their interest in the transit-poor areas where single family... |
| 25. : snake-oiling  | Candidate | Garbage | Ignore | Type Quality  | Avg. Trigram Freq | context... market. What’s-his-nose in the quote snake-oiling, “BUY! BUY... |
| 26. : never-lived-in | Candidate | Garbage | Ignore | Type Quality  | Avg. Trigram Freq | context... friends bought a million dollar-valued never-lived-in foreclosure... |
| 27. : Galtian       | Candidate | Garbage | Ignore | Type Quality  | Avg. Trigram Freq | context... situation will improve until our Galtian overlords are made to... |
| 28. : hoocooodano0e | Candidate | Garbage | Ignore | Type Quality  | Avg. Trigram Freq | context... For Conservatives Green Balloons Hoekstrolka hoocooodano0e Can [...]
<p>| 29. : end-ups       | Candidate | Garbage | Ignore | Type Quality  | Avg. Trigram Freq | context... a home always and necessarily end-ups with “nothing” (quote q... |
| 30. : Crazification | Candidate | Garbage | Ignore | Type Quality  | Avg. Trigram Freq | context... Cooking Corrections Crab Bucket Politics Crazification Factor G... |
| 31. : Spectering    | Candidate | Garbage | Ignore | Type Quality  | Avg. Trigram Freq | context... Village is Missing its Idiot Spectering Sports Starbursts Sturve... |
| 32. : dollar-valued | Candidate | Garbage | Ignore | Type Quality  | Avg. Trigram Freq | context... close friends bought a million dollar-valued never-lived-in foreclo... |</p>
<table>
<thead>
<tr>
<th>No.</th>
<th>Word</th>
<th>Type</th>
<th>Candidate</th>
<th>Garbage</th>
<th>Ignore</th>
<th>Avg. Trigram Freq</th>
<th>Trigram Quality</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>56</td>
<td>GSAVE®</td>
<td>0</td>
<td>Candidate</td>
<td>Ignore</td>
<td></td>
<td>3.94955%</td>
<td>5</td>
<td>War War on Terror aka GSAVE® WATB We Are All</td>
</tr>
<tr>
<td>57</td>
<td>insane®</td>
<td>0</td>
<td>Candidate</td>
<td>Ignore</td>
<td></td>
<td>1.75352%</td>
<td>5</td>
<td>quit because i was going insane® j Reply Matt R - May</td>
</tr>
<tr>
<td>58</td>
<td>opie_jeanne</td>
<td>2</td>
<td>Candidate</td>
<td>Ignore</td>
<td></td>
<td>10.66998%</td>
<td>2</td>
<td>of this... Horrible—just horrible... Reply opie_jeanne - June 1,</td>
</tr>
<tr>
<td>59</td>
<td>goblue72</td>
<td>2</td>
<td>Candidate</td>
<td>Ignore</td>
<td></td>
<td>2.30218%</td>
<td>2</td>
<td>and would likely default. Reply goblue72 - June 1, 2011</td>
</tr>
<tr>
<td>60</td>
<td>mid-$500s</td>
<td>0</td>
<td>Candidate</td>
<td>Ignore</td>
<td></td>
<td>0.01182%</td>
<td>0</td>
<td>for sale now, one for mid-$500s and one for over</td>
</tr>
<tr>
<td>61</td>
<td>PB&amp;J</td>
<td>0</td>
<td>Candidate</td>
<td>Ignore</td>
<td></td>
<td>0%</td>
<td>0</td>
<td>lunch that week, or make PB&amp;J, rather then order sushi</td>
</tr>
<tr>
<td>62</td>
<td>WWJBD</td>
<td>0</td>
<td>Candidate</td>
<td>Ignore</td>
<td></td>
<td>0%</td>
<td>0</td>
<td>You Can Afford It) WTF? WWJBD- What Would Jenna Bush (</td>
</tr>
<tr>
<td>63</td>
<td>sh*t</td>
<td>0</td>
<td>Candidate</td>
<td>Ignore</td>
<td></td>
<td>0%</td>
<td>0</td>
<td>If you get in the sh*t for a year, you</td>
</tr>
<tr>
<td>64</td>
<td>$4mm</td>
<td>0</td>
<td>Candidate</td>
<td>Ignore</td>
<td></td>
<td>0%</td>
<td>0</td>
<td>Friends own what was a $4mm home when they bought</td>
</tr>
</tbody>
</table>
60. : mid-$500s

http://neocrawler.gwi.uni-muenchen.de

Average trigram frequency < .1%: -1
Average trigram frequency < .05%: -1
Token contains non-latin chars or non-letter-chars other than hyphen or single-quote: -5
Token contains numbers: -1
Token contains numbers other than 1, 2, 4, 8: -1
Token contains consecutive numbers: -3

OK
Discoverer: World Wide Web
Observer: energ. author
id_author: int(11)
age: varchar(256)
gender: varchar(1)
role: varchar(256)
nation: varchar(2)
nickname: varchar(256)
first_name: varchar(256)
last_name: varchar(256)
energ. blacklist
id_blacklist: int(11)
id_lemma: int(11)
blacklist_item: varchar(256)
energ. lemma
id_lemma: int(11)
lemma: varchar(256)
word_formation: varchar(256)
semantic_transfer: varchar(256)
morphological_relatives: varchar(256)
semantic_competitors: varchar(256)
transparency: varchar(256)
energ. process_info
id_process_info: int(11)
processed_by: varchar(256)
processed_date: int(11)
process_comments: varchar(4096)
search_string: varchar(256)
time_restriction: varchar(256)
no_pages: int(11)
no_tokens: int(11)
id_lemma: int(11)
access_date: varchar(25)
energ. source
id_source: int(11)
url: varchar(256)
field: varchar(256)
source_type: varchar(256)
energ. token
id_token: int(11)
id_source: int(11)
id_version: int(11)
id_author: int(11)
pos: varchar(256)
form: varchar(256)
count: int(11)
component: varchar(256)
cotext_1: varchar(4096)
cotext_2: varchar(4096)
collocation: varchar(256)
mode_use: varchar(256)
style_use: varchar(256)
intertextuality: varchar(256)
novelty_markers: varchar(1)
explanation: varchar(256)
paraphrase: varchar(2048)
ref_ontology: varchar(256)
new_ref: varchar(1)
connotation: varchar(256)
energ. version
id_version: int(11)
creation_date: varchar(256)
id_process_info: int(11)
no_tokens: int(11)
title: varchar(256)
html: mediumtext

1. periodical monitoring
2. web search
3. find links to relevant pages
4. return links to NeoCrawler
5. download web pages
6. automatic post-processing
7. save downloaded web pages
8. show results in web interface

c. download web pages
b. get links to blog entries
d. compare web sites to word list
e. list of rated suggestions
f. manually select neologisms
g. save to database for future crawling
halfalogue

Cumulated pages per month

n=167
halfalogue - Mode of Use

- Metalinguistic: 97%
- Objectlinguistic: 3%

Susanne Stegmayr & Daphné Kerremans
cherpumple

Cumulated pages per month

n=368
cherpumple - Mode of Use

- 99% Objectlinguistic
- 1% Metalinguistic
ISLE 2, Boston • Susanne Stegmayr & Daphné Kerremans (LMU Munich)

Neologisms on the Internet

The NeoCrawler