Lexical Semantics and Distribution of Suffixes — A Visual Analysis

Christian Rohrdantz\textsuperscript{1} Andreas Niekler\textsuperscript{2} Annette Hautli\textsuperscript{1} Miriam Butt\textsuperscript{1} Daniel A. Keim\textsuperscript{1}

\textsuperscript{1}University of Konstanz
\textsuperscript{2}Leipzig University of Applied Sciences

EACL 2012 Joint Workshop of LINGVIS & UNCLH
Motivation

1. increasing amount of diachronic data electronically available
2. demand of linguists to process these corpora and uncover patterns of language use and language change
Motivation

1. increasing amount of diachronic data electronically available
2. demand of linguists to process these corpora and uncover patterns of language use and language change

Challenge

Make the data accessible for exploration and provide insight.
Motivation

1. increasing amount of diachronic data electronically available
2. demand of linguists to process these corpora and uncover patterns of language use and language change

Challenge
Make the data accessible for exploration and provide insight.

Research question
How far do we get exploring massive diachronic language data combining surface statistical methods with visualization? Can we test existing hypotheses of change and can they even generate new ones?
Lexical semantics

Productivity

Research object

The object under investigation is the **lexical semantics** and **productivity** of three derivational morphemes: -gate, -geddon, -athon

- part of a word can begin to lead an extra life as a derivational suffix → cranberry morpheme
  - e.g. burger from Hamburger (citizens from the German city Hamburg) to a food item
- these morphemes carry semantic content that carries over to new expressions (also in other languages)
The object under investigation is the lexical semantics and productivity of three derivational morphemes: *-gate*, *-geddon*, *-athon*

- part of a word can begin to lead an extra life as a derivational suffix → cranberry morpheme
  - e.g. *burger* from *Hamburger* (citizens from the German city Hamburg) to a food item
- these morphemes carry semantic content that carries over to new expressions (also in other languages)

To examine

What conditions trigger the spread of these morphemes? Are there any observable diachronic developments in their lexical semantics or productivity?
Our Investigation

Research object and methodology

- *-gate, -geddon, -athon* are relatively new
Our Investigation

Research object and methodology

- -gate, -geddon, -athon are relatively new
- It has been shown that diachronic shifts in word meaning/use can be detected and described by topic modeling (Rohrdantz et al. 2011)
Our Investigation

Research object and methodology

- *-gate, -geddon, -athon* are relatively new
- It has been shown that diachronic shifts in word meaning/use can be detected and described by topic modeling (Rohrdantz et al. 2011)

Research hypotheses

- The meaning and use of the suffixes is becoming broader
- The suffixes are about to spread
Our Investigation

Research object and methodology

- *-gate, -geddon, -athon* are relatively new
- It has been shown that diachronic shifts in word meaning/use can be detected and described by topic modeling (Rohrdantz et al. 2011)

Research hypotheses

- The meaning and use of the suffixes is becoming broader
- The suffixes are about to spread

Limitations

- While the diachronic data snapshots we base the analysis on are quite large, they only have a limited time-depth
- The statistics work on the surface, no deep linguistic analysis
Data

- **New York Times (NYT) corpus**
  - 1.8 million newspaper articles from 1987 to 2007
  - each article has a specific time stamp

- **European Media Monitor (EMM) news service data**
  - 11 million news articles from all over the world in English, French and German, from May 2009 to January 2012
  - enriched with metadata (Atkinson and der Goot 2009, Krstajic et al 2010)
Data: NYT
Data: EMM
Data: EMM

Statistics for -gate:

- 7500 -gate matches (700 distinct)
- *Rubygate* the most frequent with 1558 matches, followed by *Angolagate* (1025) and *Climategate* (752)

<table>
<thead>
<tr>
<th><strong>Lang.</strong></th>
<th><strong>Country</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>GB (1142), USA (840), Ireland (364), Pakistan (275), South Africa (190), India (131), Australia (129), Canada (117), Zimbabwe (73)</td>
</tr>
<tr>
<td>French</td>
<td>France (2089), Switzerland (429), Belgium (108), Senegal (30)</td>
</tr>
<tr>
<td>German</td>
<td>Germany (493), Switzerland (151), Austria (151)</td>
</tr>
</tbody>
</table>
Outline

1. Lexical semantics
2. Productivity
Lexical semantics

Task

- discover meaning relationships between words with suffixes -gate, -geddon and -athon and semantically related words
  - e.g. between the suffix -gate and words like scandal, affair
- determine from word contexts whether suffixed words share context features with other words
- use statistics to model word senses on the basis of word contexts
Lexical semantics

Modelling

- Latent Dirichlet Allocation (LDA) (Blei et al., 2003)
  - not applied to documents but on contexts
- we predefine the number of generated senses, each word (both suffixed and semantically related word) is assigned to one sense
- Words under investigation: affair, scandal, crisis, controversy, Watergate, ...-gate
- Visual Analysis of diachronic behaviour
Lexical semantics: Topics for *-gate*

- **Society & Art**: affair, crisis, love, controversy, scandal, book, man, woman, life, year, film, time, write, story, work, show, play, family, wife, people, begin, young, movie, art,...
- **Watergate**: scandal, affair, president, watergate, iran-contra, clinton, year, official, public, political, charge, campaign, investigation, controversy, case, nixon, today, prosecutor, bush, report, congress,...
- **Economy**: crisis, company, financial, year, scandal, market, economic, bank, government, percent, price, billion, stock, economy, million, country, business, debt, oil, industry, loan, executive, energy, investor,...
- **Foreign Policy**: crisis, president, government, political, minister, official, country, war, united states, leader, today, iraq, military, force, economic, prime, year, american, bush, time, end, people, lead, world,...
- **Sports**: controversy, affair, scandal, year, game, team, crisis, time, play, player, day, season, win, people, week, lead, sport, start, coach,...
- **Domestic Policy**: crisis, controversy, city, year, state, school, fiscal, people, budget, heath, scandal, public, time, problem, mayor, official,...
Lexical semantics: Diachronic view

**Foreign Policy**
- scandal
- controversy
- crisis
- watergate
- affair
- gate_aggregated

Relative number of "crisis"-contexts of the first week that belong to the topic Foreign Policy

**Sports**
- scandal
- controversy
- crisis
- watergate
- affair
- gate_aggregated

**Domestic Policy**
- scandal
- controversy
- crisis
- watergate
- affair
- gate_aggregated

- crisis, mr, president, government, political, minister, official, country, war, unite
Outline

1. Lexical semantics

2. Productivity
Productivity

- investigate the cases of suffixation from the standpoint of morphological productivity
- productivity for Baayen (1992) is correlated with frequency
  - complex phenomenon where factors like language structure, processing complexity and social convention contribute
- here: productivity in terms of suffix frequency, the number of news sources and languages that the suffix carries over to
Productivity: New \textit{-geddon} coinages
Productivity: New -athon coinages
Productivity: New \textit{-gate} coinages
Productivity: New coinages

- Different geddon-coinages over time
- Different athon-coinages over time
- Different gate-coinages over time
# Productivity: Different Writings for -geddon

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Carmadgeddon&quot;</td>
<td>en</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Carmageddon&quot;</td>
<td>de</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>en</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>fr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Carmageddon&quot;</td>
<td>en</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Carmageddon&quot;</td>
<td>fr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Schmarmageddon&quot;</td>
<td>en</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Shmarmageddon&quot;</td>
<td>en</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Snomageddon&quot;</td>
<td>de</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>en</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Snowmageddon&quot;</td>
<td>de</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>en</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>fr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Snowmageddon&quot;</td>
<td>fr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Snowmageddon&quot;</td>
<td>en</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Snowmageddon&quot;</td>
<td>de</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>en</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>fr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Sowmageddon&quot;</td>
<td>de</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure created with Tableau Software
### Productivity: Different Languages for -gate

<table>
<thead>
<tr>
<th>match</th>
<th>language</th>
<th>language</th>
<th>language</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Rubygate&quot;</td>
<td>de</td>
<td>en</td>
<td>fr</td>
</tr>
<tr>
<td>&quot;Angolagate&quot;</td>
<td>de</td>
<td>en</td>
<td>fr</td>
</tr>
<tr>
<td>&quot;Climategate&quot;</td>
<td>de</td>
<td>en</td>
<td>fr</td>
</tr>
<tr>
<td>&quot;Memogate&quot;</td>
<td>en</td>
<td>fr</td>
<td></td>
</tr>
<tr>
<td>&quot;Cablegate&quot;</td>
<td>de</td>
<td>en</td>
<td>fr</td>
</tr>
<tr>
<td>&quot;Antennagate&quot;</td>
<td>de</td>
<td>en</td>
<td>fr</td>
</tr>
<tr>
<td>&quot;Bloodgate&quot;</td>
<td>en</td>
<td>fr</td>
<td></td>
</tr>
<tr>
<td>&quot;Crashgate&quot;</td>
<td>de</td>
<td>en</td>
<td>fr</td>
</tr>
<tr>
<td>&quot;Asiagate&quot;</td>
<td>en</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Weinergate&quot;</td>
<td>de</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Sachsgate&quot;</td>
<td>en</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Oilgate&quot;</td>
<td>en</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Troopergate&quot;</td>
<td>en</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Karachigate&quot;</td>
<td>de</td>
<td>en</td>
<td>fr</td>
</tr>
<tr>
<td>&quot;Travelgate&quot;</td>
<td>en</td>
<td>fr</td>
<td></td>
</tr>
</tbody>
</table>

*Figure created with Tableau Software*
Observations

- *-gate* has rather vague semantics
- The underspecified meaning of *-gate* seems to contribute to the fact that many new coinages appear over time
- No clear recent development in the semantics of *-gate* observable in our data snapshot
- New coinages come up at a constant rate, the spread does not stop
- No language barrier
Future work

- Try to fill the gap in the data back to the first appearance of Watergate. Any semantic developments in the initial phase?
- Try to get more multi-lingual data from the past. Since when has -gate spread internationally?
- Which role does phonology play in the creation of new coinages?
- Are other aspects more relevant than linguistic aspects, when it comes to the spread of a new coinage? E.g. the influence of certain sources?
- Can we identify candidates for new derivational suffixes exploring massive data?
Thank you for your attention! Any questions or comments?
Abbas-gate, Adamugate, Afghan-gate, Africagate, Agliottigate, Aid-gate, Airportgate, Alicante-gate, Alinghigate, Altai-gate, Altargate, Alugate, Alu-gate, Amazonasgate, Amazongate, Amosgate, Anelka-gate, Angolagate, Angola-gate, Angologate, Antennagate, Antenna-gate, Antennegate, Apple-gate, Apprentice-Gate, Apuestagate, Arrivalsgate, Arsmgate, Asiagate, Asia-gate, Assange-Gate, Atomgate, Babygate, Baligate, Ballgate, Ballsgate, Bananagate, Bandargate, Bannergate, Bari-gate, Batterygate, Battery-gate, Beckgate, Bee-gate, Bees-gate, Belenegate, Bench-gate, Berlingate, Bertiegate, Betsygate, Bettencourtgate, Bettencourt-Gate, Biffogate, Bigotgate, Bigot-gate, Billinsgate, Biscuitgate, Biscuit-gate, Bittergate, Blackberry-Gate, Blackjack-gate, Blackoutgate, Blackwatergate, Blattergate, Bloggergate, Bloodgate, Blue-gate, Bondage-gate, Bonusgate, Boobgate, Boob-Gate, Boo-Gate, Boozegate, Bostitch-Gate, Bottomgate, Boubagate, Boulogne-gate, Bourgigate, Bra-Gate, Breadgate, Breakfast-gate, Bribery-gate, Bridgegate, Broad-gate, Brook-gate, Browgate, Bruneigate, Buggygate, Bullygate, Bulloggate, Bulog-gate, Bumpgate, Bunkergate, Butlergate, Buttongate, Buwog-Gate, Cablegate, Cable-gate, Cablegate-Gate, Caddie-gate, Caddygate, Cadmangate, Caldergate, Callistagate, Camerongate, Camillagate, Camilla-gate, Cannonsgate, Cargate, Carpetgate, Cashgate, Casinogate, Casino-gate, Casoria-Gate, Castle-gate, Catgate, Cat-gate, Cattlegate, Cementgate, Census-gate, Centralgate, Centurygate, Chaingate, Champagnegate, Cheriegate, Cherie-gate, Cherylgate, Chickengate, Chinagate, Chogm-gate, Choppergate, Christalmightygate, Cimategate, Cingapuragate, Cleavagate, Clementgate, Climategate, Climategate, Clintongate, Cllonugate, Cmpagane, Cnagapate, Cngapague, Clnetgate, Clnetgate, Clnategat,