Lexical Semantics and Distribution of Suffixes — A Visual Analysis

Christian Rohrdantz[1] **Andreas Niekler**[2] **Annette Hautli**[1] **Miriam Butt**[1] **Daniel A. Keim**[1]

[1] University of Konstanz
first.last@uni-konstanz.de

[2]Leipzig University of Applied Sciences
aniekler@fbm.htwk-leipzig.de

Abstract

We present a quantitative investigation of the cross-linguistic usage of some (relatively) newly minted derivational morphemes. In particular, we examine the lexical semantic content expressed by three suffixes originating in English: *-gate*, *-geddon* and *-athon*. Using data from newspapers, we look at the distribution and lexical semantic usage of these morphemes not only within English, but across several languages and also across time, with a time-depth of 20 years. The occurrence of these suffixes in available corpora are comparatively rare, however, by investigating huge amounts of data, we are able to arrive at interesting insights into the distribution, meaning and spread of the suffixes. Processing and understanding the huge amounts of data is accomplished via visualization methods that allow the presentation of an overall distributional picture, with further details and different types of perspectives available on demand.

1 Introduction

It is well-known that parts of a compound can begin to lead an additional life as derivational suffixes, or even as stand-alone items. A famous example is *burger*, which is now used to denote a food-item (e.g., *burger, cheese burger, veggie burger*) and is originally from the word *Hamburger*, which designates a person from the German city of Hamburg. These morphemes are generally known as *cranberry morphemes* (because of the prolific use of *cran*). Some other examples are *-(o)nomics, -(o)mat* or *(o)rama*.

While it is well-known that this morphological process exists, it is less clear what conditions trigger it and how the coinage "catches" on to become a regular part of a language. Given the current availability of huge amounts of digital data, we decided to investigate whether we could gain an insight into the use and spread of some of these morphemes via quantitative methods, thereby confirming our intuitions.

Furthermore, we decided to focus not just on the use of the cranberry morphemes in their language of origin, but also on their use and spread in other languages. In particular, we want to model the contexts in which these suffixes are used to coin new words and how these neologisms transport to other languages. We chose to look at the following three morphemes: *-gate*, *-geddon* and *-athon* because they tend to be used in "newsworthy" contexts and are therefore likely to appear in newswire and newspaper corpora, which are available to us in large amounts.

This paper describes work in progress, where we visually analyze the lexical semantics and use of the three suffixes *-gate*, *-geddon* and *-athon*. We were able to add some time-depth to our investigation via an analysis of the New York Times corpus from 1987–2007. This means that while we cannot pin-point the first occurrence and further spread of the morpheme uses, we can gain some idea as to their historical development.

Given that the amount of data we analyze is huge, we use methods from Visual Analytics in order to make the vast amount of information generated from the computational models easily accessible to the human eye and mind.

We proceed as follows: After a review of related work in Section 2, we describe our study in Section 3 and discuss the visual analysis in Section 4. In a case study we compare the meaning of words with the suffix *-gate* to other semantically related words (4.1) based on an optimized topic model. We also develop, customize and apply visualizations to investigate the productivity of new suffixes and their spread across news sources and languages (4.2). We conclude with Section 5.

2 Related Work

As already mentioned, the coinage and spread of new suffixes is well-known in theoretical linguistics. However, linguists are generally not sure what effects exactly are involved in the process (Baayen, 1992; Plag, 1999). We are not aware of any other computational work on cranberry morphemes. Work by Lüdeling and Evert (2005) on the German non-medical suffix *-itis* is closest to this paper; however, the type of the morpheme investigated is different and their focus is mainly on productivity. We concentrate more on the lexical semantic content of the suffixes, look at them across languages in bigger corpora to investigate their distribution and use and provide a layer of visual analysis.

One question we asked ourselves is whether we could predict from the context the likelihood of the suffixes *-gate*, *-geddon* and *-athon* and whether one can identify the lexical semantic content of the suffixes more precisely. This task can be formulated as a topic modeling problem for which we chose to employ Latent Dirichlet Allocation (LDA) (Blei et al., 2003). It has recently been used to perform word sense induction from small word contexts (e.g. Brody (2009)) and has also proven successful when detecting changes in word meanings over time on small word contexts in diachronic corpora (Rohrdantz et al., 2011).

We applied an optimized topic model and combined the statistical results with methods from Visual Analytics. Visual Analytics is based on the tight coupling of algorithms for automatic data analysis and interactive visual components (Thomas and Cook, 2005; Keim et al., 2010). The idea is to exploit human perceptive abilities to support the detection of interesting patterns (see Card et al. (1999) for details). Examples for visualizations used previously to investigate linguistic questions are Mayer et al. (2010a) on vowel harmony, Mayer et al. (2010b) on consonant patterns, Honkela et al. (1995) on syntactic categories, Rohrdantz et al. (2011) on lexical semantics across time.

We also used visualizations to look at cross-linguistic use and productivity of the suffixes. Prominent theoretical work on the productivity of morphemes has been done by Baayen (1992) and Plag (1999), most computational approaches have worked on English due to the availability of large enough corpora (Nishimoto, 2004). To the best of our knowledge, no large-scale quantitative study has been performed which takes into account both the diachronic as well as the cross-linguistic dimension of the development.

3 Our Approach

3.1 Research Questions & Analysis Tasks

The object of research are three productive suffixes, namely *-gate*, *geddon* and *-athon*. What these suffixes have in common is that they trigger neologisms in various languages and all of them seem to carry some lexical semantic information. Whereas *-gate*, which was coined by the *Watergate* affair, is used for scandalous events or affairs, *-geddon* seems to denote a similar concept but more of a disastrous event, building on its original use in the bible. Usually, *-athon*, coming from *marathon*, denotes a long-lasting event. We assume that the lexical semantic content of these suffixes can be modeled with standard topic models.

3.2 Data & Statistics

Our investigations are based on two different data sets, one is a diachronic news corpus, the New York Times Annotated Corpus[1] containing 1.8 million newspaper articles from 1987 to 2007. To generate the second data set, we performed an online scan of the EMM news service,[2] which links to multilingual news articles from all over the world and enriches them with metada (Atkinson and der Goot, 2009; Krstajic et al., 2010). Between May 2009 and January 2012, we scanned about eleven million news articles in English, German and French.

For both data sources, we extract a context of 25 words before and after the word under investigation, together with its timestamp. In the case of the EMM data, we also save information on the news source, the source country and the language of the article. In a manual postprocessing step, we

[1] http://www.ldc.upenn.edu/

[2] http://emm.newsexplorer.eu/

clean the dataset from words ending in the suffixes by coincidence, many of which are proper names of persons and locations.

From the EMM metadata, we can attribute the employment of the suffixes to the countries they were used in. Table 1 shows the figures for the *-gate* suffix, what language it was used in, and its country of origin. We can see that the suffix was used in many countries and different world regions between May 2009 and January 2012.

Lang.	**Country**
English	GB (1142), USA (840), Ireland (364), Pakistan (275), South Africa (190), India (131), Australia (129), Canada (117), Zimbabwe (73)
French	France (2089), Switzerland (429), Belgium (108), Senegal (30)
German	Germany (493), Switzerland (151), Austria (151)

Table 1: Usage of the suffix *-gate* in different languages/countries. For each language only the countries with the most occurrences are listed.

Among the total 7,500 *-gate* appearances, *Rubygate* – the affair of Italian's ex prime minister Silvio Berlusconi with an under-aged girl from Morocco – was the most frequent word with 1558 matches, followed by *Angolagate* with 1025 matches and *Climategate* with 752 matches. The NYT corpus has 1,000 matches of *-gate* words, the top ones were *Iraqgate* with 148, *Travelgate* with 122, and *Irangate* with 105 matches. The frequency of *-geddon* and *-athon* was much lower.

3.3 Topic Modeling

The task of the topic modeling in this paper is to discover meaning relationships between our the suffixes and semantically related words, i.e. we want to determine from the word contexts whether *-gate* words share context features with words such as *scandal* or *affair*. For this task, we use LDA, which describes a generative hierarchical Bayesian model that relates the words and documents within a corpus through a latent variable. The interpretation of this latent variable could be seen as topics that are responsible for the usage of words within the documents. Within the LDA framework we can describe the generation of a document by the following process

1. draw K multinomials $\phi_k \propto Dir(\beta_k)$, one for each topic k
2. for each document d, $d = 1, \dots, D$
   (a) draw multinomial $\theta_d \propto Dir(\alpha_d)$
   (b) for each word w_{dn} in document d, $n = 1, \dots, N_d$
      i. draw a topic $z_{dn} \propto Multinomial(\theta_d)$
      ii. draw a word w_{dn} from $p(w_{dn}|\phi_{z_{dn}})$, the multinomial probability conditioned on topic z_{dn}

Following this generative process we identify the hidden variables for every document in a corpus by computing the posterior distribution:

$$p(\theta, \phi, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \phi, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}. \quad (1)$$

Exact inference for this posterior distribution is not tractable and we use collapsed Gibbs sampling as in Griffiths and Steyver (2004). We compute the posterior distribution over all variables and model parameters instead of inferring θ and ϕ directly. The Gibbs sampling procedure samples a topic z_{dn} for each word in all documents of the corpus. This procedure is iterated until the approximated posterior distribution does not change the likelihood of the model with more iterations. As a result we get a sampled topic z_{dn} for each word in the corpus and can trace θ and ϕ. For our problem we can use the counts of z_{dn}, the count of words belonging to a topic, for each document in combination with the timestamps to see which word in question appears how often in a specific topic in which time slice. This allows us to observe the usage of a word within a certain timespan. The hidden variable ϕ can be interpreted as a matrix having the conditional probability $p(w_i|z_k)$ at the matrix position $\phi_{i,k}$. This means that every column vector in ϕ is a probability distribution over the whole vocabulary. These distributions can be seen as topics since they describe a mixture of words with exact probabilities. Having those distributions at hand we can analyze which words occur significantly often in the same topic or semantic context.

The purpose of the LDA model is to analyze the latent structure of the passages extracted from the NYT corpus. We decided to use the contexts of *Watergate*, *scandal*, *affair*, *crisis*, *controversy* in combination with the suffix *-gate*. We can then

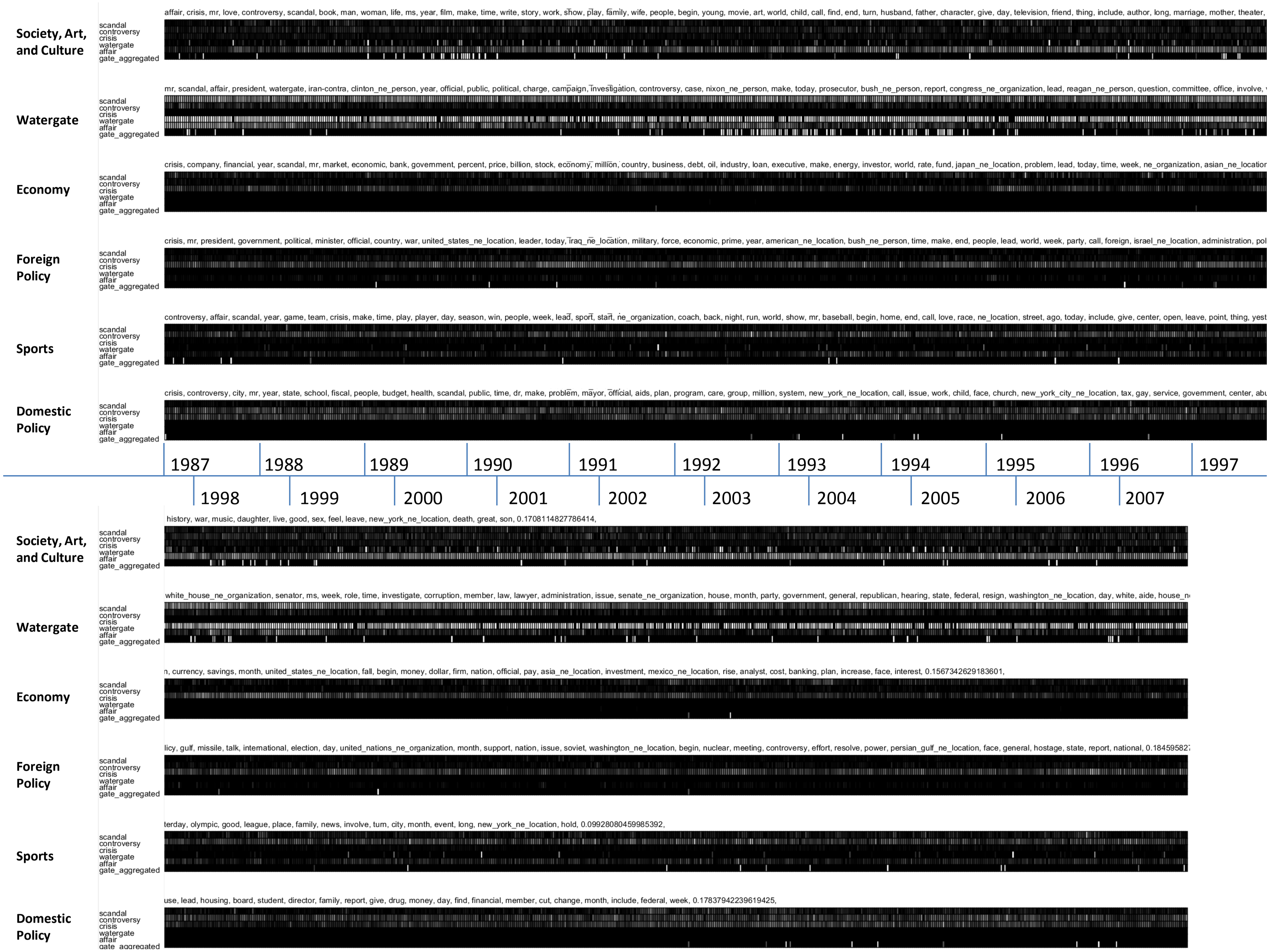

Figure 1: The diachronic distribution of the words under investigation over the 6 topics learned from the New York Times Corpus.

see where these terms co-occur and hence what the semantic context is. We infer a model which consists of six topics under the assumption that if the word senses of the six words given above do not overlap at all, there should not be more than six senses to analyze. The fixed parameter K in the model leads us to an optimization problem of the hyper-parameter β. The hyper-parameter α is not as important as β since it scales the topic per document mixture. For that reason we do not optimize α explicitly. We rather estimate the optimal value after optimizing the value for β. Since the β parameter is of crucial impact to the generation of the hidden variable ϕ and thus the topics, we need to find the optimal hyper-parameter that generalizes the model to the given data. Most approaches show that one can optimize the model for fixed parameters α and β when testing models with different values for K as in (Griffiths and Steyver, 2004). Since we are fixing K we must test the dataset for an optimal model given different values for β. This can be done by utilizing the model perplexity (Blei et al., 2003) and thus maximizing the likelihood of a test dataset from the same corpus.

In our experiment we used a relatively small number of topics and we expected a large number of words aligned to a topic.

4 Visual Analytics

4.1 Topic Modeling

The topics extracted from the NYT corpus by the model described in Section 3.3 was further investigated with respect to the correlation between the lexical semantic content of the suffixed words and a development over time. For this purpose we designed a pixel visualization (see Figure 1), mapping the data facets to the visual variables as follows: The data is divided according to the topics mapping each topic to one horizontal band. The descriptive words of a topic as found by LDA are listed above its band. In addition, each topic is manually assigned an interpretive label. These labels are at the far left of a topic band.

Each topic band is further subdivided according to the words under investigation. Under the label "gate-aggregated", all words with *-gate* suffixes (except *Watergate*) are summarized. The bands are aligned with a time axis and vertically divided into cells, each cell representing one week of data. The cell color indicates whether the corresponding word under investigation occurred within the corresponding topic in the corresponding week. The black color means that there was no such occurrence, whereas the brightest white is assigned to the cell of the week where most occurrences (*max*) of a word under investigation are found, independent from the topic. Other occurrence counts are colored in grey tones according to a linear mapping into the normalized color range from black=0 to white=*max*. Note that the normalization depends on the word under investigation, i.e. is relative to its maximal occurrence.

In Figure 1, the data has to be split into two chunks to fit the page. The upper part shows the years from 1987 to 1997 and the lower part from 1997 to 2007. There are several possibilities for user interaction: A semantic zoom allows the data to be displayed in different levels of time granularity, e.g. day, week, month, year. By mousing over a cell, the underlying text passages are displayed in a tooltip.

Findings Figure 1 shows that the topics are dominated by different words under investigation, i.e. the words under investigation cannot be clearly separated into self-contained meanings. This mixture indicates that the words under investigation have similar meanings, but that in different contexts they are used in different combinations:

1. Society, Art, and Culture: This seems to be the most general topic with the broadest usage of the words under investigation. The descriptive terms show that it is a lot about interpersonal relations and dominated by "affair". In 1989/1990 the play *Mastergate* becomes visible in the "gate-aggregated" band.

2. Economy: This topic is strongly related to "crisis" and apart from the moderate frequency of "scandal", other words are rarely used in this context. Apparently, financial scandals were usually not described attaching the suffix "-gate" in the years between 1987 and 2007.

3. Foreign Policy: This is another topic dominated by "crisis", with moderate occurrences of "controversy". Some "gate-words" also appear.

4. Sports: Here, "controversy" is the dominating element, with a raised frequency of "affair" and small frequency of "scandal". Again, "gate-words" appear from time to time, with a slightly

increased frequency towards the end.
5. Domestic Politics: The dominant words are “controversy” and “crisis”. It’s noteworthy that “controversy” is a lot more frequent here than for Foreign Policy. Especially in the last years “gate-words” appeared from time to time.

In sum, we find that there are preferred contexts in which *-gate* is used, namely mainly in topics to do with society, art and culture and that topics to do with the economy, *-gate* is hardly used. The lexical semantic content of *-gate* seems to be most closely linked to the word *affair*.

4.2 Productivity

The cases of suffixation presented above should also be considered from the standpoint of morphological productivity. For Baayen (1992), morphological productivity is a complex phenomenon in which factors like the structure of the language, its processing complexities and social conventions mingle. Whereas he focuses on the the correlation between productivity and frequency, we can take into account another variable for productivity. In particular, we can consider the number of newspapers that use a certain term. This will normalize the measures usually taken in that a term like “Watergate”, which is highly frequent and mentioned in a variety of sources is more productive than a term that occurs frequently, but only in one source. Using this methodology we can at least partly circumvent the problem of productivity effects that are merely based on the specific style of one particular newspaper.

First, we visually evaluate the productivity of the different suffixes plotting the sum of different coinages against time, see Figure 2. As can be expected, in all three cases there is a steeper slope in the beginning of the monitored period. This is an artifact because all older coinages that had been around before the monitoring started will be observed for the first time. As more time passes all plots show a linear overall trend, indicating that the rate with which new coinages appear remains somewhat constant. Yet, there are some local oscillations in the rate that become more visible in the plots of *-geddon*- and *-athon*-coinages, which are in general much more infrequent than *-gate*-coinages. It can be concluded that over the last two and a half years the suffixes kept their rate of productivity in English, German, and French newswire texts fairly constant.

To investigate the cross-linguistic productivity of the new coinages we customized a visualization with the Tableau software.[3] Figure 3 shows the appearances of the 15 most frequent *-gate*-coinages across the three languages over time. Along the y-axis the data is divided according to *-gate*-coinages and languages, whereas the x-axis encodes the time. Whenever a certain coinage appears in a certain language at a certain point in time, a colored triangle is plotted to the corresponding position. The color redundantly encodes the language for easier interpretation.

Figure 3 shows many interesting patterns. The most salient patterns can be summarized as:
1. No language barrier: The top *-gate*-coinages belong to scandals that are of international interest and once they are coined in English they immediately spread to the other languages, see *Rubygate*, *Climategate*, *Cablegate*, *Antennagate*, and *Crashgate*. Only in the case of *Angolagate* and *Karachigate* there is a certain delay in the spread, possibly due to the fact that it was coined in French first and initially did not achieve the same attention as coinages in English.
2. Pertinacity partly depends on language: Some *-gate*-coinages re-appear over and over again only in individual languages. This especially holds for words that were coined before the monitoring started, e.g. *Sachsgate*, *Oilgate*, *Troopergate*, and *Travelgate* which all persist in English. Examples can be found for other languages, e.g. *Angolagate* for French. Interestingly, in German *Nipplegate* persists over the whole monitored period, but only in German, and even outperforms its German spelling *Nippelgate*.
3. Some coinages are special: Some of the recent coinages such as *Memogate*, *Asiagate*, and *Weinergate* reach an extremely high frequency within very short time ranges, but can be found almost exclusively in English. These will be subject of further investigation in Section 4.2.1. It has to be noted that many of the infrequent coinages appear only once and are never adopted.

4.2.1 Spread across News Sources and Countries

Figure 3 clearly shows that *Memogate* is heavily mentioned within English speaking news

[3] http://www.tableausoftware.com/

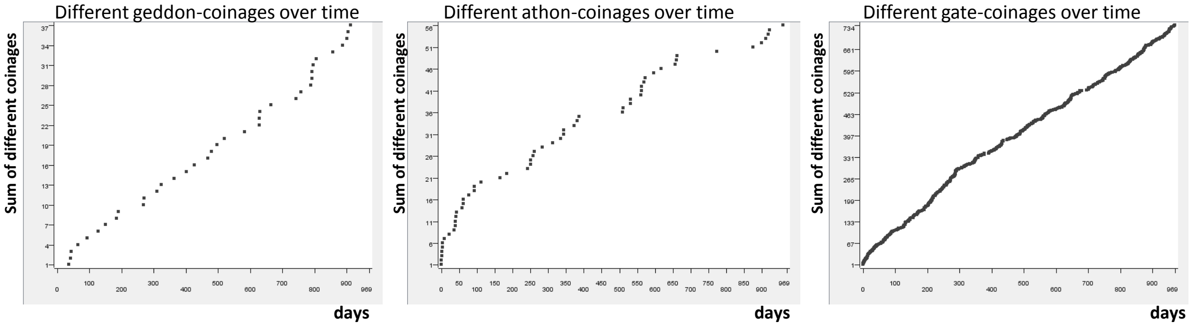

Figure 2: The number of different coinages containing the suffixes under investigation (on the y-axis) plotted against the number of days passed during the monitoring process (on the x-axis)

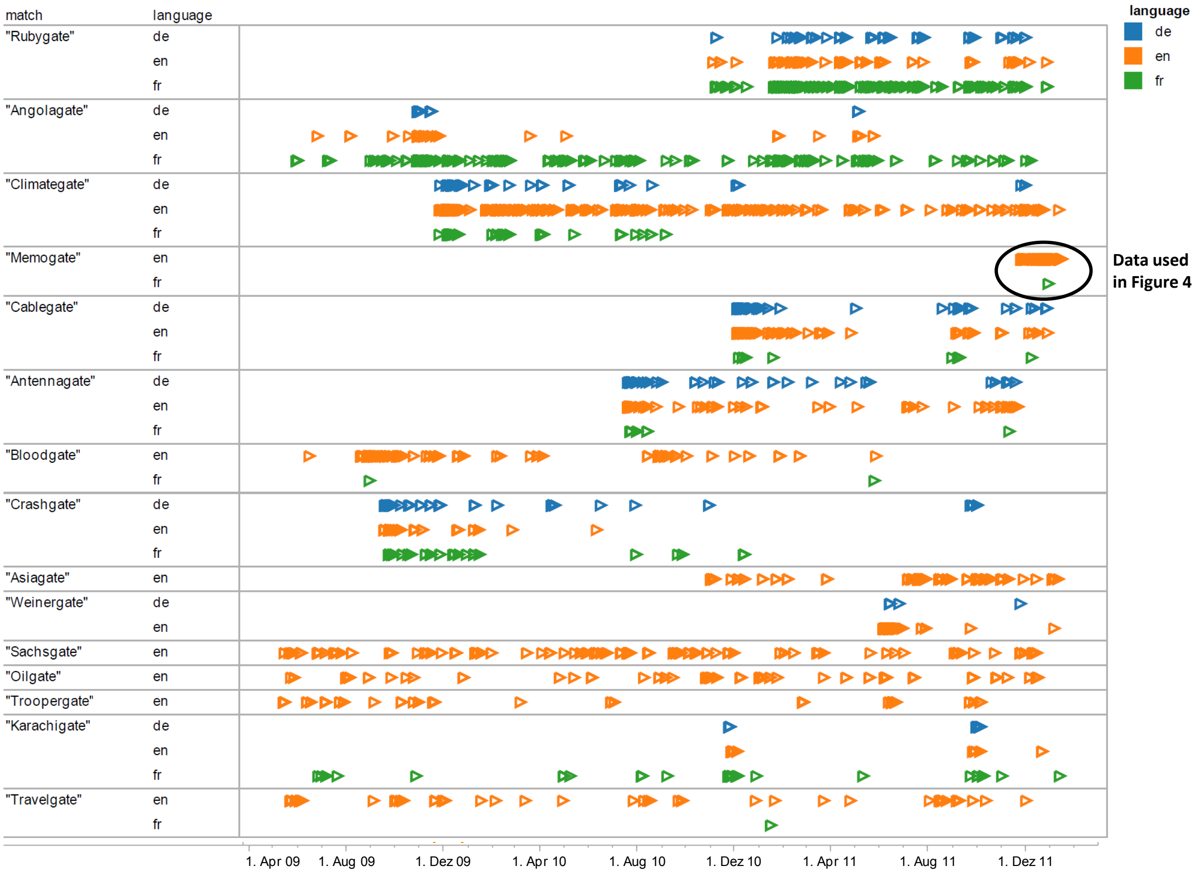

Figure 3: The appearances of the 15 most frequent *-gate* coinages over time and across the different languages

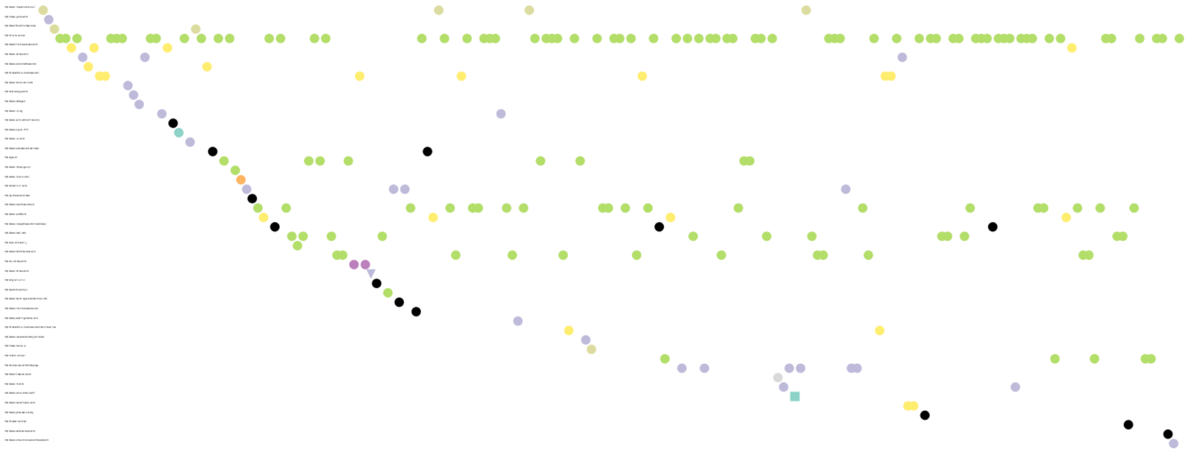

Figure 4: Detailed analysis of the *Memogate* cluster highlighted in Figure 3 using alternative visual mappings: Sequence of spread over different countries and news sources.

sources within a short time range. We developed a further visualization that shows how these mentions sequentially distribute over different news sources and countries. In Figure 4 each article mentioning *Memogate* is represented by a colored icon. The y-axis position encodes the news source, the x-axis position encodes the temporal order of the occurrences. Note that exact time differences are omitted to make the display more compact. The shape of an icon indicates the language of the article; Circles (English) heavily dominate. The color encodes the country of origin of the news source, here green (Pakistan), yellow (India), and purple (USA) dominate.

Findings: While the first three mentions of *Memogate* could be found in British and American Newspapers, early on it was adopted by *http://tribune.com.pk/* in Pakistan (fourth line from the top) and used so heavily that it kept being adopted and became constantly used by further sources from Pakistan and also India. Apparently, individual sources may have a huge influence on the spread of a new coinage.

5 Future work and conclusion

We have presented initial experiments with respect to the application of topic modeling and visualization to gain a better understanding of developments in morphological coinage and lexical semantics. We investigated three relatively new productive suffixes, namely *-gate*, *-geddon*, and *-athon* based on their occurrences in newswire data. Even though our data set was huge, the occurrences of the suffixes are comparatively rare and so we only had enough data for *-gate* to investigate the contexts it occurs in with an optimized topic modeling. The results indicate that it is used in broader contexts than *affair*, with which it is most related. Different domains of usage could be distinguished, even though a clear development over time could not be detected based the NYT corpus. Investigating the multilingual newswire data it became evident that all three suffixes under investigation have a relatively stable rate of appearance. Many more different *-gate*-coinages could be found, though. We could observe that *-gate* was usually attached to one specific single event, and especially in many of the less frequent coinages the suffix was combined with proper names of persons, institutions, or locations. In contrast, *-athon* and *-mageddon* coinages seem to be easier to generalize. For example, the two most widely spread coinages *Snowmageddon* and *Carmageddon*, while initially referring to a certain snow storm and a certain traffic jam, have been applied to further such events and can be found listed in resources such as the Urban Dictionary.[4]

In conclusion, we demonstrated that visual analyses can help to gain insight and generate new hypotheses about the behavior of the distribution and use of new morphemes. In our future research we aim to investigate how much the success of a certain coinage depends on the event as such and its news dynamics, and what role linguistic features like e.g. phonology (two vs. three syllables, etc.) might play.

[4] http://www.urbandictionary.com/define.php?term=Carmageddon

Acknowledgments

This work has partly been funded by the Research Initiative "Computational Analysis of Linguistic Development" at the University of Konstanz and by the German Research Society (DFG) under the grant GK-1042, Explorative Analysis and Visualization of Large Information Spaces, Konstanz. The authors would like to thank Volker Rehberg for his programming support and Thomas Mayer for comments on previous versions of the paper.

References

Martin Atkinson and Erik Van der Goot. 2009. Near real time information mining in multilingual news. In Juan Quemada, Gonzalo León, Yoëlle S. Maarek, and Wolfgang Nejdl, editors, *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 1153–1154.

R. Harald Baayen. 1992. On frequency, transparency, and productivity. *Yearbook of Morphology*, pages 181–208.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Samuel Brody and Mirella Lapata. 2009. Bayesian word sense induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 103–111, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman, editors. 1999. *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Thomas L. Griffiths and Mark Steyver. 2004. Finding scientific topics. In *Proceedings of the National Academy of Sciences 101*, pages 5228–5235.

Timo Honkela, Ville Pulkki, and Teuvo Kohonen. 1995. Contextual relations of words in grimm tales, analyzed by self-organizing map. In *Proceedings of International Conference on Artificial Neural Networks (ICANN-95)*, pages 3–7.

Daniel A. Keim, Joern Kohlhammer, Geoffrey Ellis, and Florian Mansmann, editors. 2010. *Mastering The Information Age - Solving Problems with Visual Analytics*. Goslar: Eurographics.

Milos Krstajic, Florian Mansmann, Andreas Stoffel, Martin Atkinson, and Daniel A. Keim. 2010. Processing Online News Streams for Large-Scale Semantic Analysis. In *Proceedings of the 1st International Workshop on Data Engineering meets the Semantic Web (DESWeb 2010)*.

Anke Lüdeling and Stefan Evert. 2005. The emergence of productive non-medical *-itis*. corpus evidence and qualitative analysis. In S. Kepser and M. Reis, editors, *Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives*, pages 351–370. Berlin: Mouton de Gruyter.

Thomas Mayer, Christian Rohrdantz, Miriam Butt, Frans Plank, and Daniel Keim. 2010a. Visualizing vowel harmony. *Journal of Linguistic Issues in Language Technology (LiLT)*, 4(2).

Thomas Mayer, Christian Rohrdantz, Frans Plank, Peter Bak, Miriam Butt, and Daniel A. Keim. 2010b. Consonant co-occurrence in stems across languages: Automatic analysis and visualization of a phonotactic constraint. In *Proceedings of the ACL 2010 Workshop on NLP and Linguistics: Finding the Common Ground (NLPLING 2010)*, pages 67–75.

Eiji Nishimoto. 2004. Defining new words in corpus data: Productivity of english suffixes in the british national corpus. In *26th Annual Meeting of the Cognitive Science Society*.

Ingo Plag. 1999. *Morphological productivity. Structural constraints in English derivation*. Berlin/New York: Mouton de Gruyter.

Christian Rohrdantz, Annette Hautli, Thomas Mayer, Miriam Butt, Daniel A. Keim, and Frans Plank. 2011. Towards tracking semantic change by visual analytics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Langauge Technologies (ACL-HLT '11): shortpapers*, pages 305–310, Portland, Oregon. Association for Computational Linguistics.

James J. Thomas and Kristin A. Cook. 2005. *Illuminating the Path The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center.