# Investigation of Word Senses over Time Using Linguistic Corpora

Christian Pölitz[1](✉), Thomas Bartz[1], Katharina Morik[1], and Angelika Störrer[2]

[1] Artificial Intelligence Group, TU Dortmund University,
Otto Hahn Str. 12, 44227 Dortmund, Germany
christian.poelitz@tu-dortmund.de

[2] Germanistische Linguistik, Mannheim University,
Schloss, Ehrenhof West, 68131 Mannheim, Germany

**Abstract.** Word sense induction is an important method to identify possible meanings of words. Word co-occurrences can group word contexts into semantically related topics. Besides the pure words, temporal information provide another dimension to further investigate the development of the word meanings over time. Large digital corpora of written language, such as those that are held by the CLARIN-D centers, provide excellent possibilities for such kind of linguistic research on authentic language data. In this paper, we investigate the evolution of meanings of words with topic models over time using large digital text corpora.

**Keywords:** Word sense induction · Topic models · Time · Linguistic corpora

## 1 Introduction

Finding polysemy of words is an important linguistic analysis. For instance the word *bank* has multiple meanings in English. It can be used in the context of a credit institute or a river bank. Identifying these meanings not only helps understanding language better, it also can be used to filter out words in certain meanings from given contexts. The latter for instance is important to investigate usage of certain words in certain meanings.

Word sense induction (WSI) is a technique to find possible meanings of words based on automatic analysis methods. The most prominent methods in WSI are clustering of words, contexts of words or finding groups based on co-occurrence statistics. To automatically find possible meanings, large textual data set are usually used. In this paper, we investigate how good linguistic corpora are suited for WSI. We use Latent Dirichlet Allocation as introduced by Blei et al. [1] (LDA) for WSI on so called key words in context (KWIC) lists. A KWIC list contains snippets, usually some sentences, that contain a word for which we want to identify possible meanings. For convenience, throughout the paper we use the term document for the snippets in the KWIC lists. Linguistic infrastructure projects like Clarin-D provide excellent linguistic resources to retrieve such KWIC lists and to perform such linguistic research, see for instance McEnery et al. [8].

P. Král and V. Matoušek (Eds.): TSD 2015, LNAI 9302, pp. 191–198, 2015.
DOI: 10.1007/978-3-319-24033-6_22

Besides the pure identification of different meanings of words, the investigation of the development of these meanings over time is also an important linguistic task. For instance the English word *cloud* has recently got the new semantic context of cloud computing. Such emergences of new word meanings appear often over time. Interesting questions in this context are whether the meaning becomes the dominant meaning of the word or do several meanings coexist. Further in lexicography, the evolution of word meanings is important to construct descriptive examples to update existing dictionary entries as in Engelbert and Lemnitzer [4]. We use the continuous time topic model by Wang and McCallum [12] which is an extension of LDA to model also the temporal dimension of the word meanings.

In this paper, we investigate the development of meanings over time for the German language on the dictionary of the German language: "Wörterbuch der deutschen Sprache" (DWDS). The DWDS core corpus of the 20th century (DWDS-KK), constructed at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW), contains approximately 100 million running words, balanced chronologically (over the decades of the 20th century) and by text genre (belles-lettres, newspaper, scientific and functional texts). The newspaper corpus Die ZEIT (ZEIT) covers all the issues of the German weekly newspaper Die ZEIT from 1946 to 2009, approximately 460 million running words (see [5] by Geyken; [7] by Klein and Geyken).

## 2 Related Work

The induction of semantic meaning by usage patterns in the area of automatic analysis for linguistics is already well researched. An early statistical approach was completed by Brown et al. [3], Navigili [9] provides a comprehensive overview on the current research. Brody and Lapata [2] have shown, that they obtained the best results with the help of Latent Dirichlet Allocation [1]. In addition, they expanded their method to take into consideration various other context features besides the pure word occurrences (e.g. part of speech tags, syntax, etc.). Originally, LDA was used for thematic clustering of document collections. Navigli and Crisafulli [10] have already shown this to also be useful for the disambiguation of small text snippets, for example when clustering the search results from a web search engine. Rohrdantz et al. [11] showed the benefits of this method as a basis for the visualization of semantic change of example words from an English newspaper corpus, allowing them to observe the emergence of new meanings and reconstruct their development over time.

## 3 Topic Models

Topic models are statistical models that group documents and words from a document collection into so called topics. The words and documents that are associated with a topic are statistically related based on co-occurrences of words. Latent Dirichlet Allocation (LDA) as introduced by Blei et al. [1] has been successfully used for the estimation of such topics. In LDA, it is assumed that the words in a document are drawn from a Multinominal distribution that depends on latent factors, later interpreted as topics. We briefly summaries the generative process of document as the following:

1. For each topic z:
   (a) Draw $\theta_z \sim Dir(\beta)$
2. For each document d:
   (a) Draw $\phi_d \sim Dir(\alpha)$
   (b) For each word i:
      i. Draw $z_i \sim Mult(\phi_d)$
      ii. Draw $w_i \sim Mult(\theta_{z_i})$

Assuming a number of topics, we draw for each of them a Multinominal distribution of the words in this topic from a Dirichlet distribution $Dir(\beta)$ with metaparameter $\beta$. For each document we draw a Multinominal distribution of the topics in this document from a Dirichlet distribution $Dir(\alpha)$ with metaparameter $\alpha$. Finally, for each word in the document we draw a topic with respect to the topic distribution in the document and a word based on the word distribution for the drawn topic. The metaparameter $\alpha$ and $\beta$ are prior probabilities of the Multinominal distributions drawn from the Dirichlet distribution. These priors are the expected word probabilities in a topic before we have seen any data.

The generation of the LDA Topic Model is usually done by Variational Inference, as in the original work by Blei et al. [1], or via Gibbs samplers, as for proposed by Griffiths et al. [6]. We use Gibbs sampler to sample topics directly from the topic distribution. Integrating $\theta$ and $\phi$ out, we get for the probability of a topic $z_i$, given a word $w$ in a document $d$ and all other topic assignments:

$$\begin{aligned} &p(z_i|w, d, z_1, \cdots z_{i-1}, z_{i+1}, \cdots z_T) \\ &\propto \frac{N_{w,z_i} - 1 + \beta}{N_{z_i} - 1 + W \cdot \beta} \cdot (N_{d,z_i} + \alpha) \end{aligned} \tag{1}$$

We denote $N_{w,z}$ the number of times topic $z$ has been assigned to word $w$, $N_{d,z}$ the number of times topic $z$ has been assigned to any word in document $d$, $N_z$ the number of times topic $z$ has been assigned to any word, $W$ the number of words in the document collection and $T$ the number of topics.

After a sufficient number of samples from the Gibbs sampler we get estimates of the word distributions for the topics and the topic distributions for the documents:

$$\theta_{w|t} = \frac{N_{w,t} + \beta}{N_t + W \cdot \beta} \tag{2}$$

$$\phi_{d|t} = \frac{N_{d,t} + \alpha}{N_d + T \cdot \alpha} \tag{3}$$

## 4 Topic Models over Time

While the standard topic models group only words and documents in semantically related topics, we are further interested in the distribution of the topics over time. In order to extract the distribution of word senses over time, we use topic models that consider temporal information about the documents. Each document has a time stamp. These time stamps are assumed to be Beta distributed. This Beta distribution is integrated in an LDA topic model. Wang and McCallum [12] introduced this model to investigate topics over time. Throughout this paper we call this method topics over time LDA. The generative process given by the Enumeration 2 is extended such that for each word $w_i$ in each document, we also draw a time stamp $t_i \sim Beta(\psi_{z_i})$ with $\psi_{z_i} = (\alpha, \beta)$ the shape parameters of the Beta distribution. The shape parameters are estimated by the method of moments. After each Gibbs iterations the parameters are estimated in the following way: For each topic $z$ we estimate the mean $\hat{m}$ and sample variance $s^2$ of all time stamps from the documents that have been assigned this topic. By the method of moments, we set $\alpha = \hat{m} \cdot (\frac{\hat{m}\cdot(1-\hat{m})}{s^2} - 1)$ and $\beta = (1-\hat{m}) \cdot (\frac{\hat{m}\cdot(1-\hat{m})}{s^2} - 1)$ for each topic. Integrating the time stamp as Beta distributed random variable, we get for the probability of a topic $z_i$, given a word $w$ in a document $d$ with time stamp $t$ and all other topic assignments:

$$\begin{aligned} & p(z_i | w, d, t, z_1, \cdots z_{i-1}, z_{i+1}, \cdots z_T) \\ \propto & \frac{N_{w,z_i} - 1 + \beta}{N_{z_i} - 1 + W \cdot \beta} \cdot (N_{d,z_i} + \alpha) \cdot \frac{(1-t_d)^{\alpha-1} \cdot t_d^{\beta-1}}{Beta(\alpha, \beta)} \end{aligned} \tag{4}$$

where the last term comes from the density of the Beta distribution at time stamp $t_d$ and $Beta(\alpha, \beta)$ the Beta function.

## 5 Experiments

We perform experiments on the DWDS corpus for two German words with multiple meanings over time. The words “Platte” (with meanings board / disc / hard disc / plate / conductor), and “Ampel” (with meanings traffic light / a coalition of German parties (the social democrats (red), the liberals (yellow) and the green party).

In the first experiment, we investigate how good possible different meanings can be found with respect to time. We compare the standard LDA with topics over time LDA. We use 10 topics and performed 2000 Gibbs iterations. In Figure 1 we plot the distribution of the extracted topics over time for both methods. We use the DWDS core corpus in the experiment to retrieve KWIC lists from the documents from 1900 till 2000. Standard LDA does not consider time, but we can accumulate the probability proportions of the documents for the topics grouped by time periods and plot them.

From the two distributions we see that using topics over time LDA, we get a much clearer distinction of the topics over time. We can directly read off

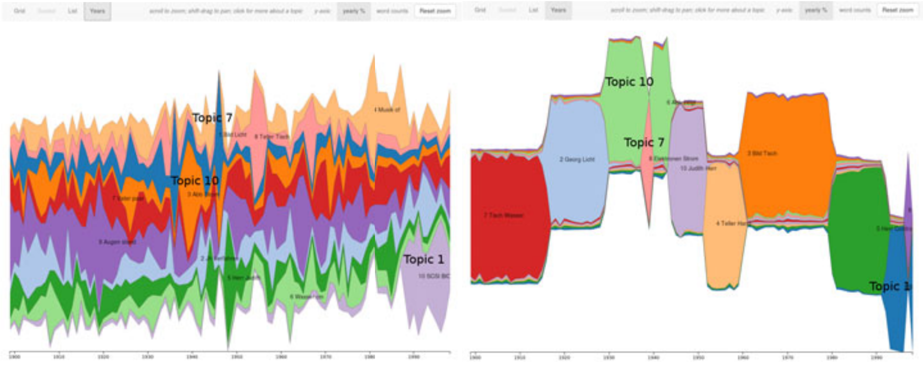
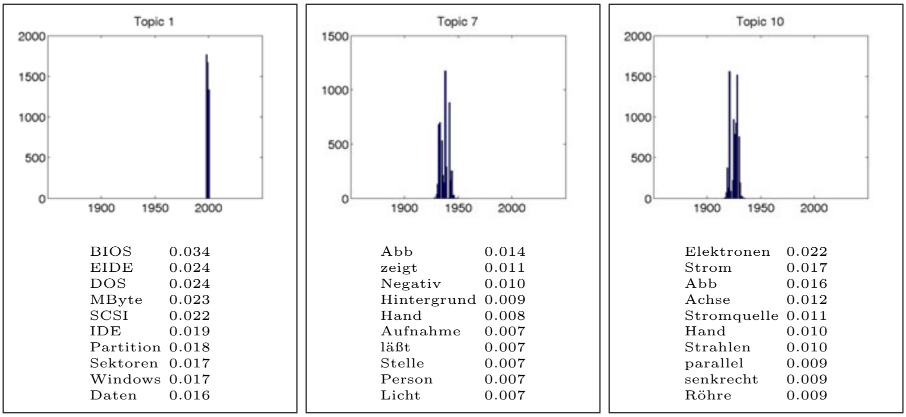


**Fig. 1.** The distribution of 10 topics extracted from the KWIC lists of the word "Platte" (board/disc/hard disc/plate/conductor). Left: Standard LDA; Right: LDA with topics over time.

the topics and the temporal period when this topic was prominent. From the standard LDA, we get a much more diffuse distribution of the topics over the time. The results indicate three possible main meanings that clearly separate over time. These topics are summarize in Figure 5. There we show the most likeliest words for the topic and the distributions of the time stamps as histogram. First, in topic 1 we find computer related words as most likely. The distribution of the time stamps shows a peak between 1990 and 2000. Before this period, this topic has not appeared. For topic 7, the most probable words indicate the meaning of a photographic plate for the word "Platte". The two most likeliest words are "Abb" which is short for "Abbildung" (Engl. picture) and "zeigt" (Engl. to show). The distribution of the time stamps shows a major usage of this meaning



| | |
|---|---|
| BIOS | 0.034 |
| EIDE | 0.024 |
| DOS | 0.024 |
| MByte | 0.023 |
| SCSI | 0.022 |
| IDE | 0.019 |
| Partition | 0.018 |
| Sektoren | 0.017 |
| Windows | 0.017 |
| Daten | 0.016 |

| | |
|---|---|
| Abb | 0.014 |
| zeigt | 0.011 |
| Negativ | 0.010 |
| Hintergrund | 0.009 |
| Hand | 0.008 |
| Aufnahme | 0.007 |
| läßt | 0.007 |
| Stelle | 0.007 |
| Person | 0.007 |
| Licht | 0.007 |

| | |
|---|---|
| Elektronen | 0.022 |
| Strom | 0.017 |
| Abb | 0.016 |
| Achse | 0.012 |
| Stromquelle | 0.011 |
| Hand | 0.010 |
| Strahlen | 0.010 |
| parallel | 0.009 |
| senkrecht | 0.009 |
| Röhre | 0.009 |

**Fig. 2.** Three topics extracted from the KWIC lists for the word "Platte" using topics over time LDA. Top: histograms of the time stamps in the topics. Bottom: most likely words in the topics.

till the 50. Topic 10 is associated with the meaning conductor that has most of its usage in 1920 and 1930. The two most likeliest words are "Elektronen" (Engl. electrons) and "Strom" (Engl. current). From the time stamps in the topic it seems that the "Platte" is no longer used with this meaning. On the other hand, from standard LDA we seem that this "Platte" is still used with this meaning. Here we see the clear limitation of the topics over time LDA: the separation into the time spans absorbs all probability mass. This mean, the density of the Beta distribution overwhelms the remaining parts of the topic probabilities.

The results from the first experiment show that the density of the Beta distribution of the time stamps tends to put to much of weight on the single topics. This gets worse the more topics we have since than we have less different time stamps per topic and hence the density of the corresponding Beta distribution gets very large at these time stamps.

To investigate this further, we perform another experiment with only 2 topics. On the Zeit corpus, we investigate the development of possible meanings from the word "Ampel". Figure 3 shows the distribution over time for the two topics. Compared to the previous experiment, we still get a very strict separation between the topics with respect to time. In Figure 4 we further investigate the distribution of the time stamps in the two topics. Additional to the histograms of the time stamps, we also show the fitted Beta distributions from each of the topics and the histogram of the time stamps over all topics.

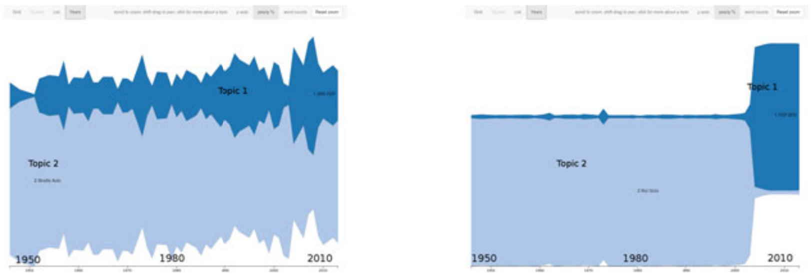
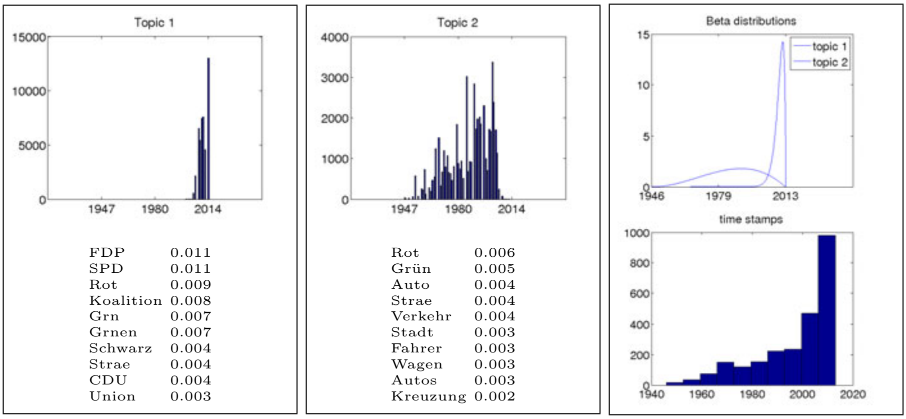


**Fig. 3.** The distribution of 2 topics extracted from the KWIC lists of the word "Ampel" (traffic light/a coalition of German parties (the social democrats (red), the liberals (yellow) and the green party)). Left: Standard LDA; Right: LDA with topics over time.

The histograms show also a strict separation of topic 1 and topic 2. Only a look on the curve of the fitted Beta distribution indicates that topic 2 is still present today. Investigating the number of documents respectively time stamps per year, we see that from 2000 on we have much more documents. This means, for topic 2, we have a much larger variance in the time stamp that makes the density of the Beta distribution smaller for individual time stamps for this topic. In topic 1 on the other hand, there are many times stamps from a small time period. This makes the Beta density much larger for those time stamps compared to the time stamps from topic 1.



| | |
|---|---|
| FDP | 0.011 |
| SPD | 0.011 |
| Rot | 0.009 |
| Koalition | 0.008 |
| Grn | 0.007 |
| Grnen | 0.007 |
| Schwarz | 0.004 |
| Strae | 0.004 |
| CDU | 0.004 |
| Union | 0.003 |

| | |
|---|---|
| Rot | 0.006 |
| Grün | 0.005 |
| Auto | 0.004 |
| Strae | 0.004 |
| Verkehr | 0.004 |
| Stadt | 0.003 |
| Fahrer | 0.003 |
| Wagen | 0.003 |
| Autos | 0.003 |
| Kreuzung | 0.002 |

**Fig. 4.** Two topics extracted from the KWIC lists for the word "Ampel" using topics over time LDA. Top: histograms of the time stamps in the topics. Bottom: most likely words in the topics. Right top: fitted Beta distributions in the topics. Right bottom: histogram of time stamps over all topics.

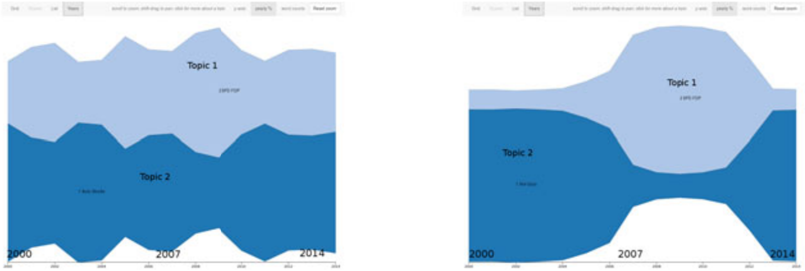


**Fig. 5.** The distribution of 2 topics extracted from the KWIC lists from 2000 till 2014 of the word "Ampel" (traffic light/a coalition of German parties (the social democrats (red), the liberals (yellow) and the green party)). Left: Standard LDA; Right: LDA with topics over time.

In Figure 5, we show the results of LDA and topic over time LDA on the same date as before but we filter out all KWIC list entries that have a time stamp from before 1999. Both figures show that the topic 2 still is present after year 2000. Topic 1 is very prominent from 2008 to 2010. Due to election periods and coalition talks between parties, these times make sense. Further, it also is clear that "Ampel" in the meaning of traffic light will no get lost. Topics over time LDA separates the topics still better and now also keeps topic 2 over the time.

To conclude, standard LDA can be used to identify the emergence of new topics to a certain degree. Topics over time LDA on the other hand separates topics over time quite quite, but sometimes to strict.

## 6 Conclusion and Future Work

In this paper, we proposed to use topic models that model also time for the investigation of evolution of meanings of words. In such a setting, time is modelled as addition random variable that is Beta distributed. We performed two extensive experiments on large linguistic corpora to test standard LDA and topics over time LDA for word sense induction over time. The results are promising but leave also some questions. In the future, we want to further investigate how to handle imbalances in the number of documents and words over the time.

## References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. Brody, S., Lapata, M.: Bayesian word sense induction. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2009, pp. 103–111. Association for Computational Linguistics, Stroudsburg (2009)
3. Brown, P.F., Pietra, S.A.D., Pietra, V.J.D., Mercer, R.L.: Word-sense disambiguation using statistical methods. In: Proceedings of the 29th Annual Meeting on Association for Computational Linguistics, ACL 1991, pp. 264–270. Association for Computational Linguistics, Stroudsburg (1991)
4. Engelberg, S., Lemnitzer, L.: Lexikographie und Wörterbuchbenutzung. Number 14 in Stauffenburg-Einführungen; 14; Stauffenburg-Einführungen. Stauffenburg-Verl., Tübingen, 2. aufl. edition (2004)
5. Geyken, A.: The DWDS corpus. A reference corpus for the german language of the twentieth century. In: Fellbaum, C. (ed.) Idioms and Collocations. Corpus-Based Linguistic and Lexicographic Studies, pp. 23–40. Continuum, London (2007)
6. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences **101**(Suppl. 1), 5228–5235 (2004)
7. Klein, W., Geyken, A.: Das Digitale Wörterbuch der Deutschen Sprache (DWDS) **26**, 79–96 (2010)
8. Mautner, G.: Tony mcenery, richard xiao and yukio tono, corpus-based language studies: an advanced resource book. Routledge, London (2006). pp. xix, 386. pb. Language in Society, 37:455–458, 7 2008
9. Navigli, R.: Word sense disambiguation: A survey. ACM Comput. Surv. **41**(2), 10:1–10:69 (2009)
10. Navigli, R., Crisafulli, G.: Inducing word senses to improve web search result clustering. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, pp. 116–126. Association for Computational Linguistics, Stroudsburg (2010)
11. Rohrdantz, C., Hautli, A., Mayer, T., Butt, M., Keim, D.A., Plank, F.: Towards tracking semantic change by visual analytics. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers, HLT 2011, vol. 2, pp. 305–310. Association for Computational Linguistics, Stroudsburg (2011)
12. Wang, X., McCallum, A.: Topics over time: a non-markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2006, pp. 424–433. ACM, New York (2006)