Finding diachronic sense changes by Unsupervised methods

Arun Kumar Jayapal

Thesis submitted for the Degree of Doctor of Philosophy

School of Computer Science & Statistics

Trinity College

University of Dublin

11 August 2017

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. Wherever there is published or unpublished work included, it is duly acknowledged in the text.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Arun Jayapal

Summary

An existing word in a language may acquire a new meaning as time passes, in addition to meanings it already possesses. Such a new word-meaning pairing is called a *semantic neologism*. An example of a semantic neologism in English is 'tweet' referring to 'a post on twitter', which is a newly acquired sense, in addition to its long established sense of referring to 'a bird song'. This thesis addresses the problem of the computational detection of such changes from time-stamped raw text i.e., text without sense annotations. For this, a generative model is proposed with variables for time Y, sense S, and contexts $\boldsymbol{w}$ around a given target word (a potential semantic neologism). A target word will be treated as having is exhibiting K senses over a given time period. The model has senses dependent on times, expressed by $P(S|Y)$, and context words dependent on senses, expressed by $P(\boldsymbol{w}|S)$. This reflects first a reasoning that sense ratios change over time. Secondly it reflects another reasoning that senses themselves, each seen as a probability distribution over words, can be regarded as more or less eternal, and subject to little change over time.

Two different estimation schemes Expectation Maximization (EM) [Dempster et al., 1977] and Gibbs sampling [Gelfand and Smith, 1990] are proposed and the parameter updates are also derived to get Maximum Likelihood and mean parameter estimates. For a genuine semantic neologism the expectation is that the estimated $P(S|Y)$ values will for some $S = k$ show an initial phase of being close to zero which it then departs from.

To evaluate the estimated parameters, a ground truth date of emergence is required: call this C_0 – the time at which the neologistic sense for the word departed from close to zero and continued to climb thereafter in a given corpus. It is hard to provide such C_0 dates because of the lack of diachronic sense-labeled corpora. In prior work there is a lack of consensus about, or reflection on, the appropriate way to deal with this difficult ground-truth problem. One of the contributions of the thesis is to reflect on the options, their strengths and weaknesses. The Oxford English Dictionary (OED) maintains the earliest citation date of a word-sense pair – call this D_0^e. This will be argued that this should not be taken to be C_0 (though it sometimes has been) but taken as a *lower-bound* for C_0. To establish a more exact C_0 a new so-called 'tracks-plot' method will be proposed. The idea behind this 'tracks-plot is that there may be words particularly associated with novel sense and not with other senses. If the per-year probabilities for such words, $P(w|Y)$, are plotted they are expected to be at close to 0 during an initial period and take off at C_0. For this work, Google 5-gram time-stamped corpus is used, which is based

on Google's digitized books holdings. Some arguments for choosing this this dataset over possible alternative sources of time-stamped data are further established in this thesis.

Before conducting actual experiments on the datasets, the model was first tested using a novel 'pseudo-neologisms' technique that was adopted from the 'pseudo-word' [Schütze, 1998] technique. Then a number of actual experiments were conducted on 'positive' targets – that is targets known to exhibit sense emergence over some particular time period. Parameter estimates were obtained from which it was mostly possible to identify a novel sense. Similarly experiments were conducted on 'negative' targets – targets known not to exhibit sense emergence. Mostly no novel sense was detected. There are a few semantic neologism targets for which the model did not discover an expected novel sense. These are apparent failures of the algorithm but some further analysis of the 5-gram data in these cases is at least suggestive of the possibility that the anticipated senses are objectively absent from the 5-gram data.

Comparisons with related work along a number of orthogonal dimensions are made, including model, datasets and evaluation approaches, though strictly quantitative comparison will not be undertaken because of differences of datasets and evaluation approaches. The most closely related work is that of Frermann and Lapata [2016], whose independently developed model involves components $P(S|Y)$ and $P(\boldsymbol{w}|S,Y)$, and so treats words as dependent on sense and time, unlike the proposal developed here where words depend only on sense. The model used in Frermann and Lapata [2016] will be argued to be conceptually a refinement of the model proposed here (though it was not developed as such). Given the success of the model proposed in the thesis in discovering the expected novel senses from the Google 5-gram data, it seems the greater sophistication of the proposal in Frermann and Lapata [2016] relative to the simpler model proposed here is not completely motivated.

Acknowledgement

This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at Trinity College Dublin.

Firstly, I would like to express my sincere gratitude to my supervisor Dr. Martin Emms for the continuous support throughout the course of my studies, for his hands-on involvement in this research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. In fact, some of the most important contributions included herein, such as the diachronic model in chapter 3 and evaluation by traks-plot in chapter 4 are based on original ideas by him.

Besides my advisor, I would like to thank the rest of my thesis advisors: Dr. Simon Wilson and Dr. Carl Vogel, for their insightful comments and encouragement. Credit and gratitude are also owed to the examiners of this thesis, Dr Brett Houlding and Dr Tony Veale, whose feedback and advice strengthened this work significantly. I would further like to thank Carmen and Erwan for taking their valuable time in reviewing the thesis. Many thanks also go to the other Ph.D. students and post-docs for the technical discussions: Akira, Alexander, Alfredo, Ashjan, Christy, Fasih, Grace, Justine, Kevin, Liliana and Shane.

A special thanks to my family. Words cannot express how grateful I am to my father, and mother for all of the sacrifices that you've made on my behalf. Your prayer for me was what sustained me thus far. I would also like to thank Deepak, his wife Nisha for their moral support and all other friends who supported me in writing, and incented me to strive towards my goal.

Above all, I praise God, the almighty for providing me this opportunity and granting me the capability to proceed successfully. This thesis appears in its current form due to the assistance and guidance of several people. I would therefore like to offer my sincere thanks to all of them.

Contents

Declaration ii

Summary iv

Acknowledgments v

Table of Contents x

List of Tables xiv

List of Figures xix

Glossary of notations xxi

1 Introduction 1

1.1 Formal and Semantic Neologisms 1

1.2 Other lexical changes 3

1.3 Research goal 3

1.4 Motivation 4

1.5 What affects word frequencies? 6

1.6 Granularity of senses 7

1.7 Thesis plan 7

2 Literature review 11

2.1 Problem evolution 11

2.2 Models and algorithms in the prior work 12

2.2.1 Approaches using probabilistic generative models 12

2.2.2 Approaches using clustering 15

2.2.3 Other approaches 16

2.3 Evaluation in the prior work 18

2.4 Discussion 21

3 Research theory **23**
3.1 Parameter estimation essentials . . . 23
3.1.1 Parameter estimation approaches . . . 23
3.1.2 Dirichlet priors . . . 25
3.1.2.1 Mean . . . 25
3.1.2.2 Mode . . . 27
3.1.2.3 Dirichlet as conjugate prior . . . 28
3.1.2.4 Impact of Dirichlet prior . . . 29
3.1.3 Expectation Maximization outline . . . 31
3.1.4 Gibbs sampling outline . . . 31
3.2 Proposed Diachronic model . . . 32
3.2.1 Static model alternative . . . 37
3.2.2 Alternative choices . . . 37
3.3 EM estimation for Diachronic model . . . 38
3.3.1 Deriving EM updates . . . 40
3.3.2 Proof of Monotone increase in likelihood . . . 43
3.3.3 MAP estimation . . . 44
3.3.4 Parameter initialization . . . 45
3.4 Gibbs sampling estimation for Diachronic model . . . 45
3.4.1 Deriving Gibbs sampling distributions . . . 46
3.4.2 Why 'burn-in'? . . . 49
3.4.3 Parameter initialization . . . 49
3.4.4 How is a Gibbs sampler different from the EM? . . . 49
3.4.5 Label switching . . . 50
3.4.6 Credible interval . . . 53
3.5 The model of Frermann and Lapata [2016] . . . 54

4 Evaluation & analysis options for neologism **57**
4.1 Ground truth for sense emergence . . . 57
4.1.1 Dictionary first citation . . . 58
4.1.2 Dictionary first inclusion . . . 59
4.1.3 Tracks-plots . . . 59
4.1.4 Emergence time detection . . . 60
4.2 Analyzing parameter outcomes . . . 64
4.2.1 *gist* words . . . 65
4.2.2 Sense examples . . . 66

5 Diachronic dataset & target possibilities **69**
5.1 Downloadable datasets . . . 69
5.2 Web-accessible datasets . . . 71
5.3 Google 5-gram . . . 73

5.4 Comparing eras . . . 74
5.5 Choosing targets . . . 74

6 Neologism Experiments - Google 5-grams 79
6.1 Generating sub-corpus . . . 80
6.2 pseudo-neologism model test . . . 80
6.2.1 byte-ostensible . . . 81
6.2.2 genocide-ostensible . . . 82
6.2.3 supermarket-ostensible . . . 83
6.3 Neologism targets . . . 84
6.3.1 mouse . . . 86
6.3.2 gay . . . 89
6.3.3 strike . . . 90
6.3.4 bit . . . 92
6.3.5 compile . . . 93
6.3.6 paste . . . 95
6.3.7 surf . . . 97
6.3.8 boot . . . 99
6.3.9 rock . . . 102
6.3.10 stoned . . . 103
6.4 Non-neologism targets . . . 106
6.5 Discriminating neologism vs non-neologism targets . . . 114
6.6 Neologisms which were undetected . . . 116
6.6.1 hip . . . 117
6.6.2 export . . . 118
6.6.3 mirror . . . 119
6.6.4 domain . . . 119
6.6.5 high . . . 121
6.7 Further model tests . . . 122
6.7.1 Ablation tests . . . 123
6.7.2 Merge tests . . . 123
6.8 Comparison with the prior work . . . 128
6.9 Discussion . . . 130

7 Conclusions and future work 133
7.1 Summary of the contributions . . . 133
7.2 Future work . . . 135

A Appendix 139
A.1 Sense definitions . . . 139

Bibliography **143**

Index **150**

List of Tables

1.1 Sample English - German - Tamil translations via Google translate for sentences with words *gay* and *bricked*; S1, S2 are the original English sentences; G1, G2 are the German translations for S1, S2; T1, T2 are the Tamil translations for S1, S2 and; L1, L2 are the Transliterations for the Tamil text in T1, T2. 5

3.1 Example sentences with the target entities 33

3.2 Mock dataset used to test label switching for a target T with window 5 containing A and B as vocabulary items from a single year 1990. The $0's$ and $1's$ to the right are appended providing a sense annotation for each data item. . . . 50

3.3 Mock dataset used to test label switching for a target T with window 5 containing A and B as vocabulary items from two years 1990 and 2000. The $0's$ and $1's$ to the right are appended providing a sense annotation for each data item. . 52

4.1 Top 30 *gist* words for the target *mouse* given each sense listed here are ranked by computing the ratio of $P(W|S=k)$ to $P_{corp}(W)$. 66

4.2 Top 30 sense examples words for the target *mouse* given 'sense 1' from the year *1990* with their probabilities listed here are ranked by computing $P(S|Y,\boldsymbol{w})$. 67

5.1 A list of corpora explored for the Diachronic analysis with further details related to the corpora are provided here – see text for explanation on further details. 70

5.2 Sample 5-grams for target *mouse* from the year 1821 73

5.3 The table provides the information for targets that are neologisms 76

6.1 Words used to generate pseudo-neologisms 81

6.2 Top 30 *gist* words for *byte-ostensible* pseudo-word 82

6.3 Top 30 *gist* words for *genocide-ostensible* pseudo-word 83

6.4 Top 30 *gist* words for *supermarket-ostensible* pseudo-word 84

6.5 Google 5 gram dataset - the table provides the information for targets that are neologisms 85

6.6 D_0^c(OED) gives the date of first citation in the OED, under D_0^i LDCOE (resp. COE) gives the date of first dictionary inclusion in LDOCE (resp. COE), C_0(Tracks) gives a tracks-based corpus-emergence data in the relevant Google 5 gram sub-corpus, GS-Date and EM-Date give sense emergence dates derived from $\boldsymbol{\pi}_t$ parameters inferred by GS and EM, and $GS < 10\%$ and $EM < 10\%$ indicate whether these agree with 10% of the time-span with C_0(Tracks) 85
6.7 Top *gist* words for the neologism sense for target *mouse* ranked by comparing word distributions to corpus Probabilities 87
6.8 Top neologism sense examples for the target *mouse* extracted from inferred EM and Gibbs sampling estimates for *sense 1* 87
6.9 Top neologism sense examples for the target *mouse* – extracted from inferred EM estimates for *sense 1* 88
6.10 Top 30 *gist* words for the target *gay* ranked by comparing word distributions to corpus Probabilities 89
6.11 Top neologism sense examples for the target *gay* extracted from inferred EM and Gibbs sampling estimates for *sense 2* 90
6.12 Top 30 *gist* words for the target *strike* ranked by comparing word distributions to corpus Probabilities 91
6.13 Top neologism sense examples for the target *strike* extracted from inferred EM and Gibbs sampling estimates for *sense 1* and *sense 2* respectively 91
6.14 Top 30 *gist* words for the target *bit* ranked by comparing word distributions to corpus Probabilities 93
6.15 Top neologism sense examples for the target *bit* extracted from inferred EM and Gibbs sampling estimates for *sense 2* 93
6.16 Top 30 *gist* words for the target *compile* ranked by comparing word distributions to corpus Probabilities 94
6.17 Top neologism sense examples for the target *compile* extracted from inferred EM and Gibbs sampling estimates for *sense 2* 94
6.18 Top 30 *gist* words for the target *paste* ranked by comparing word distributions to corpus Probabilities 95
6.19 Top neologism sense examples for the target *paste* extracted from inferred EM and Gibbs sampling estimates for *sense 2* and *sense 1* 96
6.20 Top 30 *gist* words for the target *surf* ranked by comparing word distributions to corpus Probabilities 99
6.21 Top neologism sense examples for the target *surf* extracted from inferred EM and Gibbs sampling estimates for *sense 3* and *sense 4* respectively. 99
6.22 Top 30 *gist* words for the target *boot* ranked by comparing word distributions to corpus Probabilities 101
6.23 Top neologism sense examples for the target *boot* extracted from inferred EM and Gibbs sampling estimates for *sense 3* and *sense 2* respectively 101

6.24 Top 30 *gist* words for the target *rock* ranked by comparing word distributions to corpus Probabilities . . . 102
6.25 Top neologism sense examples for the target *rock* extracted from inferred EM and Gibbs sampling estimates for *sense 2* . . . 103
6.26 Top 30 *gist* words for the target *stoned* ranked by comparing word distributions to corpus Probabilities . . . 105
6.27 Top neologism sense examples for the target *stoned* extracted from inferred EM and Gibbs sampling estimates for *sense 0* . . . 105
6.28 Top neologism sense examples for the target *stoned* extracted from inferred EM and Gibbs sampling estimates for *sense 3* and *sense 2*. . . . 105
6.29 For (a-e) provides the top 30 *gist* words of non-neologism senses for the targets *mouse, surf, gay, bit, boot* – further continued in next page for other targets . . 107
6.29 continued from previous page – For EM and Gibbs sampling word distributions $\boldsymbol{\theta}_k$, the top 30 *gist* words of non-neologism senses for targets *mouse, surf, gay, bit, boot, compile, strike, rock, stoned, paste* (a-j) ranked by comparing word distributions to corpus Probabilities. . . . 108
6.30 Google 5 gram dataset - the table has the targets for non-neologisms . . . 109
6.31 For the EM outcomes, top 30 *gist* words for the target *promotion* ranked by comparing word distributions to corpus Probabilities are shown . . . 114
6.32 For the GS outcomes, top 30 *gist* words for the target *promotion* ranked by comparing word distributions to corpus Probabilities are shown . . . 114
6.33 Novelty scores for the neologism and non-neologism targets based on the EM inferred $\pi_t[k]$ sense parameter outcomes. . . . 115
6.34 Novelty scores for the neologism and non-neologism targets based on the Gibbs sampling inferred $\pi_t[k]$ sense parameter outcomes. . . . 115
6.35 Google 5 gram dataset - the table provides the information for targets that are neologisms . . . 116
6.36 Top 30 *gist* words for the target *hip* with emerging sense, ranked by comparing word distributions to corpus Probabilities for the EM-outcomes with 4 and 5 sense settings . . . 117
6.37 Top 30 *gist* words for the target *domain* with emerging sense, ranked by comparing word distributions to corpus Probabilities for the EM-outcome with 4 sense settings – for seemingly emerging senses (0 & 3) . . . 120
6.38 Top 30 *gist* words for the target *high* with emerging sense, ranked by comparing word distributions to corpus Probabilities for the EM-outcomes with 5 sense settings . . . 121
6.39 KL-divergence (symmetric) distance between inferred $\boldsymbol{\theta}_k$ distributions for *surf* . 125
6.40 Sense numbers allocated based on merging inferred $\boldsymbol{\theta}_k$ distributions for *surf* – see text for further details. . . . 126

6.41 Top 30 *gist* words for the target *surf* ranked by comparing word distributions to corpus Probabilities, where the word distributions are obtained from the new merged distributions. 127

List of Figures

1.1 Word frequencies for the words *supermarket, genocide* that are formal neologisms 2

1.2 Word frequency (solid line) and sense frequencies (dashed lines). 4

1.3 Word frequency plots to demonstrate how (a) People's opinion affect word frequencies (b) Changes in the world affect word frequencies. 6

2.1 Plate diagram for Topic model (Latent Dirichlet Allocation) 13

2.2 This is a screenshot from table 1 of Mitra et al. [2015]'s paper, showing clusters for the target 'compiler' 15

2.3 This is a sreenshot of the the schematic shown in figure 2 of Mitra et al. [2015] depicting the birth of a new sense. 16

2.4 This is a screen-shot of the plot for the word *tape* provided in figure 5 from Kulkarni et al. [2014]'a paper 17

2.5 This is a screen-shot of the plot for the word *tou ming* provided in figure 6 from Tang et al. [2015]'s paper 17

2.6 This is a screen-shot of the plot for a sense of the word *tou ming* provided in figure 7 from Tang et al. [2015]'a paper 17

3.1 2D plots of Dirichlet distribution, (from left to right), α at different settings: 1. $\alpha = 1$, 2. $\alpha = 0.1$, 3. $\alpha = 5$, 4. $\alpha = 0.01$. The figure provides 15 random draws at each of four different α settings. The graph has the number of topics plotted on x-axis and the probability of the topic on the y-axis. 30

3.2 3D plots of Dirichlet distribution. Each of these plots are produced at dimension $k = 3$ and the respective alpha settings are titled for each plot. It can observed that when $\alpha = 1$ (symmetric), the distribution is uniform, while $\alpha = 10$ (symmetric), the distribution has a mode at the maximum and the hump is dense. But when we have α's set asymmetric, the distributions are skewed. . . 30

3.3 34

3.4 Plate diagram - No prior 35

3.5 Plate diagram - with prior over parameters 36

3.6 The Gibbs samples of sense $\boldsymbol{\pi}_t[k]$ and word $\boldsymbol{\theta}_k[w]$ produced in with a 10k run are plotted in (a) and (b); the left hand plots in (a) and (b) shows the density of 10k Gibbs samples while the right hand plots show the Gibbs samples in 10k iterations. 51

3.7 The Gibbs samples of sense $\boldsymbol{\pi}_t[k]$ and word $\boldsymbol{\theta}_k[w]$ produced in with a 50k run are plotted in (a) and (b); the left hand plots in (a) and (b) shows the density of 50k Gibbs samples while the right hand plots show the Gibbs samples of 50k samples. 51

3.8 The Gibbs samples of sense and word distributions produced on a larger dataset with 24 items are plotted in (a) and (b); the left hand plots in (a) and (b) shows the density of 50k Gibbs samples from 50k iterations and the right hand plots show the Gibbs samples of 50k samples. 52

3.9 A Density plot showing 90% HPD interval with area under the curve shaded. . 53

3.10 Plate diagram for the model used in Frermann and Lapata [2016] 54

4.1 A hypothetical sense emergence plot - Dictionary and corpus emergence dates (D_0^c & C_0) annotated 58

4.2 Tracks plot for *mouse* - True C_0 annotated 60

4.3 example taken from Granjon [2013]. (a) data, which is samples from successive Gaussians with means $\mu_0 = 0$, $\mu_1 = 1$, changing at sample 1000. (b) generalised likelihodd G, which reaches a threshold value of 100 at sample 1201 (c) cumulative sum S_n giving estimated change-point at sample 1001 62

5.1 Plots of words adjacent to *mouse* and *surf* are based on values from ratio queries given to the Google n-gram viewer API – values are normalized by their means over time. The black line in each case concerns adjacent words intuitively particularly associated with an emerging sense, the red lines concerns words associated with a long established sense. 75

6.1 The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *byte-ostensible* pseudo-neologism, and the third plot shows the known *byte* and *ostensible* proportions 82

6.2 The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *genocide-ostensible* pseudo-neologism, and the third plot shows the known *genocide* and *ostensible* proportions 83

6.3 The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *supermarket-ostensible* pseudo-neologism, and the third plot shows the known *supermarket* and *ostensible* proportions . . 83

6.4 The first and second plots show the EM and Gibbs sampling algorithm's inferred $\boldsymbol{\pi}[k]$ sense parameter outcomes for *mouse* experiment, and the third plot shows the probability 'tracks' for some words that are intuitively associated with the 'computer peripheral' sense of *mouse*. Dating information – EM:1982, GS:1982, OED:1965, Tracks: 1982) 87

6.5 The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *gay* experiment, and the third plot shows the probability 'tracks' for some words that are intuitively associated with the 'homosexual person' sense of *gay*. Dating information – EM:1970, GS:1969, OED:1941, Tracks:1969 89

6.6 The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *strike* experiment, and the third plot shows the probability 'tracks' for some words that are intuitively associated with the 'industrial action' sense of *strike*. Dating information: EM:1904, GS:1901, OED:1822, Tracks: 1899 90

6.7 Tracks plot for words *slip, emptive, capability* 92

6.8 The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *bit* experiment, and the third plot shows the probability 'tracks' for some words that are intuitively associated with the 'basic unit of information' sense of *bit*. Dating information – EM:1958, GS:1958, OED:1948, Tracks: 1966 93

6.9 The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *compile/compiling* experiment, and the third plot shows the probability 'tracks' for some words that are intuitively associated with the 'transform to machine code' sense of *compile/compiling*. Dating information – EM:1965, GS:1965, OED:1952, Tracks: 1972 94

6.10 The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *paste* experiment, and the third plot shows the probability 'tracks' for some words that are intuitively associated with the 'homosexual person' sense of *paste*. Dating information – EM:1981, GS:1981, OED:1975, Tracks:1982 96

6.11 Tracks plot for words *scissors, eugenol, oxide* in comparison with the words *cut, copy, text, image, clipboard* 96

6.12 EM inferred plot – *surf/surfing/surfed* – 3 and 4 sense settings 98

6.13 The first and second plots show the EM and Gibbs sampling algorithm's inferred $\boldsymbol{\pi}[k]$ sense parameter outcomes for *surf/surfing/surfed* experiment, and the third plot shows the probability 'tracks' for some words that are intuitively associated with the 'exploring internet' sense of *surf/surfing/surfed*. Dating information – EM:1992, GS:1992, OED:1992, Tracks: 1993 98

6.14 'Tracks' plot showing tracks for words *turf, n* 100

6.15 The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *boot/boots/booted/booting* experiment, and the third plot shows the probability 'tracks' for some words that are intuitively associated with the 'computer start up' sense of *boot/boots/booted/-booting*. Dating information – EM:1981, GS:1980, OED:1980, Tracks: 1982 . 100

6.16 Tracks plot for words *car, camp* 101

6.17 The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *rock* experiment, and the third plot shows the probability 'tracks' for some words that are intuitively associated with the 'genre of music' sense of *rock*. Dating information – EM:1960, GS:1960, OED:1956, Tracks: 1967 103

6.18 EM inferred plot – *stoned* – 5 sense setting on Google 5-gram books dataset . . 104

6.19 The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *stoned* experiment, and the third plot shows the probability 'tracks' for some words that are intuitively associated with the 'under the influence of drug' sense of *stoned*. Dating information – EM:1965, GS:1960, OED:1952, Tracks: 1959 104

6.20 For plots (a-d), the first and second columns show the outcomes of EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *ostensible, cinema, present, promotion* non-neologism targets – (e-h) continued in the next page. 110

6.20 For plots (e-h), the first and second columns show the outcomes of EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *theatre, play, plant, spirit* non-neologism targets. 111

6.21 For plots (a-d), the first and second columns show the outcomes of EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *ostensible, cinema, present, promotion* non-neologism targets – (e-h) continued in the next page. 112

6.21 For plots (e-h), the first and second columns show the outcomes of EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *theatre, play, plant, spirit* non-neologism targets. 113

6.22 The first, second and third plots show the EM inferred $\pi_t[k]$ sense parameter outcomes for *hip* experiment with 3, 4 and 5 sense settings. 117

6.23 The plot (a) shows the n-gram viewer 'tracks' for 2-grams and plot (b) shows probabilities normalized 'tracks' for some words that we expect to be associated with the 'trendy' sense of *hip* (c) shows the probability (un-normalized) 'tracks' for words that we expect to be associated with the 'trendy' sense of *hip* 118

6.24 Plots (a) shows the EM inferred $\pi_t[k]$ sense parameter outcomes for *export* experiment with 5 sense settings (b) shows probability 'tracks' for some words that we expect to be associated with 'convert file format' usage of the word *export* that are normalized between 0 and 1 and plot (c) shows the 'non-normalized' version of (b). 119
6.25 Plots (a) shows the EM inferred $\pi_t[k]$ sense parameter outcomes for *mirror* experiment with 5 sense settings (b) shows probability 'tracks' for some words that we expect to be associated with 'store copies of data' usage of the word *mirror* that are normalized between 0 and 1 and plot (c) shows the 'non-normalized' version of (b). 120
6.26 Plots (a) shows the EM inferred $\pi_t[k]$ sense parameter outcomes for *domain* experiment with 5 sense settings (b) shows probability 'tracks' for some words that we expect to be associated with 'suffix of internet address' usage of the word *domain* that are normalized between 0 and 1 and plot (c) shows the 'non-normalized' version of (b). 120
6.27 Plots (a) shows probability 'tracks' for some words from 'gist' that are not relevant to 'convert file format' usage of the word *domain* that are normalized between 0 and 1 and plot (c) shows the 'non-normalized' version of (a). 121
6.28 The first and second plots show the EM inferred $\pi_t[k]$ sense parameter outcomes for *high* experiment with 3 and 5 sense settings. 122
6.29 Ablation test outcomes on *mouse* dataset – plots from (a-h) shows the EM outcomes from $75\%, 60\%, 45\%, 30\%, 15\%, 5\%, 2\%$ and 0.5% of the complete dataset between the years 1970 and 2008. 124
6.30 Hierarchical cluster plots based on KL-divergence distances between $\boldsymbol{\theta}_k$ distributions for *surf*. The sense numbers are represented as $s0, \ldots, s4$ and the height provided in the plot gives the distance between the clusters. 126
6.31 The plots show the initialized and inferred sense distributions $\pi_t[k]$ for the target *surf* – shows first level merge 127
6.32 The plots show the initialized and inferred sense distributions $\pi_t[k]$ for the target *surf* – shows second level merge 127
6.33 The plots show the initialized and inferred sense distributions $\pi_t[k]$ for the target *surf* – shows second level merge 128
6.34 Screen-shot of Table 2 from Lau et al. [2012] showing top-10 terms for each of the senses induced by HDP for the word 'cheat'. 129

Glossary of notations

Notation	Description
T	target word, the word in question
$\boldsymbol{D}$	dataset with the target word T that has $1 \dots D$ data items
d	is a particular data item (or index of one in $\boldsymbol{D}$)
Y	variable for time stamp (year) of a data item (when it was authored)
N	the number of possible values for Y (the number of unique time stamps in the dataset)
S	variable for the sense of the target in a data item
K	the number of possible values for S
$\boldsymbol{w}$	variable for the vector of context words around the target word T in a data item
W^d	the length of $\boldsymbol{w}$ for data item d
V	the number of unique words in $\boldsymbol{D}$ – the size of the vocabulary
t	variable to index through time stamps
k	variable to index through senses
v	variable to index through vocabulary
$\boldsymbol{\tau}$	a length N vector of time probabilities
τ_t	probability of time stamp t
$\boldsymbol{\pi}_t$	a length K vector of sense probabilities at time t
$\pi_{t,k}$	probability of sense k at time t
$\boldsymbol{\theta}_k$	a length V vector of context word probabilities for sense k
$\theta_{k,v}$	the probability of vocabulary item v in context of sense k
$\boldsymbol{t}^{1:D}$	sequence of all the time stamps from the data items in $\boldsymbol{D}$
$\boldsymbol{s}^{1:D}$	sequence of all the senses from the data items in $\boldsymbol{D}$
$\boldsymbol{w}^{1:D}$	sequence of all the word vectors from the data items in $\boldsymbol{D}$
$\boldsymbol{\gamma}_\pi$	Dirichlet hyperparameter for prior on $\boldsymbol{\pi}_t$
$\boldsymbol{\gamma}_\theta$	Dirichlet hyperparameter for prior on $\boldsymbol{\theta}_k$
$\boldsymbol{\Theta}$	notation to group all parameters together

Chapter 1

Introduction

Languages have continuously evolved over time and such changes are called 'diachronic' in nature. The different forms of language changes include change in word pronunciation, grammatical change, change in word spelling, syntactical change and lexical changes. There could be many aspects such as social changes and technological advancements that influence such changes. For most of this, the evidence comes from *written* time-stamped text inspected by humans. In Natural language processing (NLP), there are a number of works that deal with different aspects of these language changes, but there are not many works that deal with semantic changes (a form of lexical change) in languages. In section 1.1, formal and semantic neologisms – forms of lexical change that are of interest for this thesis are introduced. Also, some other lexical changes are also discussed in section 1.2. Then in sections 1.3 and 1.4, the research goal and motivation for this research is established. There could be a number of factors that affects word and sense frequencies, only one of which is genuine language change. Such factors are discussed in section 1.5. Then in section 1.6, there is a short discussion on word sense granularity and the thesis-plan is provided in section 1.7.

1.1 Formal and Semantic Neologisms

There are two forms of lexical change that are of interest to understand the research problem. They are,

1. When a new word is coined it is called a 'formal neologism'. In this case a letter/-phoneme sequence comes to be acceptable as a word where it was not before.
2. When an existing word acquires a novel sense, it is called a 'semantic neologism' [Tournier, 1985]. In this case the letter/phoneme sequence already exists, but takes a meaning it did not have before.

Some examples of recent *formal neologisms* are; *crowdsourcing*: which refers to 'getting work done by a large community through a website' introduced in the year 2006; *selfie*: which refers to 'a self-photograph taken using a smart-phone' introduced in 2002; *bromance*: which refers to 'an intimate relationship between men' introduced in the year 2001. These are first

citation dates from online Oxford English Dictionary (OED).

To consider some less recent examples, the words *supermarket* and *genocide* date from the years 1931 and 1944 respectively, according to the OED. Their emergence can actually be verified by using n-gram frequency data provided by Google. Figure 1.1 is based on the *relative* frequency of these words in successive years; the raw relative frequencies for a particular word are normalized by their mean over the times[1]. For the words *supermarket* and *genocide*, it can be seen from the figure they have close to zero frequency until they emerged sometime around 1935 and 1940.

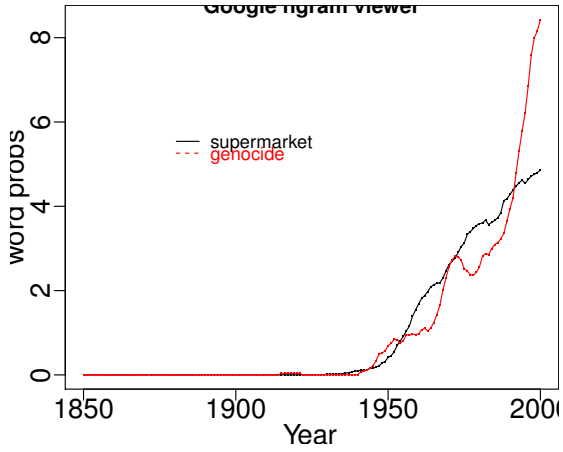


Figure 1.1 – Word frequencies for the words *supermarket, genocide* that are formal neologisms

Turning to *semantic neologism* some recent examples are; *tweet*: which has an older sense referring to bird noise and a newer sense referring to post a message on Twitter; *bricked*: which has an older sense concerning construction with bricks and a newer sense which concerns making some computing device unresponsive, probably by a software update. Some examples of these usages are given below, where: (a) and (a′) correspond to old and new senses for *tweet* and (b) and (b′) refer to the old and new senses for *bricked*. These examples and their year indications are obtained by searching for these words using Google timeline search[2].

(a) *When a bird* **tweets***, it's telling you what it is and where it is.* (1995)

(a′) *An Embedded* **Tweet** *brings the best content created on Twitter into your article or website.* (2009)

(b) **bricked** *up the windows of the old house* (1990)

(b′) *He managed to get his new iMac* **bricked** *while trying to boot WinXP on it.* (2013)

Given a plain text time-stamped corpus it is easy to identify the date of emergence of a formal neologism (ie., a newly coined term) just by using the frequency count of the word itself (seen in the figure 1.1). It is not so straightforward to identify the emergence date of a semantic neologism (ie., 'new' sense of an existing word).

[1]This plot is based on data downloaded via Google-Ngrams API `https://github.com/econpy/google-ngrams`. This is further discussed in section 5.5. Fairly similar plots can be obtained online using the Google n-gram viewer `https://books.google.com/ngrams`.

[2]Google provides a search feature which allows its users to search for words for a particular time-period. We call this a Google timeline search

The above examples portray existing words *tweet, bricked* acquiring a novel sense. It is important to note there are other forms of lexical changes such as pejoration, amelioration, broadening and narrowing senses but this research is concerned only with *semantic* neologisms. In the next section 1.2, there is a further discussion about these lexical changes to establish how the sense emergence problem is different from these other language changes.

1.2 Other lexical changes

In addition to *semantic* neologisms, the other forms of language changes are discussed in this section.

'Pejoration' refers to when a word's sense becomes more negative over time, while 'Amelioration' refers to a when word's sense becomes more positive over time. As an example for 'Pejoration' consider the word *awful* which used to refer to something 'worthy of respect', but recently it has taken a negative usage to mean 'not worthy' or 'bad'. For 'Amelioration', consider the word *geek* used to refer to 'a fool' early 20^{th} century, while in the recent past it has evolved to have a more positive sense meaning 'a person who is extremely knowledgeable'. Cook and Stevenson [2010] use corpora from different time periods to study the change in the semantic orientation of words with respect to 'Pejoration' and 'Amelioration'.

Broadening is a type of language change where the word meaning gets *more inclusive* than its earlier meaning. An example of *broadening of word sense*, *Guy Fawkes* infamous first name lost its specificity with the proliferation of November 5th effigies of the criminal; then *guys* began to be used of males of strange appearance, then it was broadened to refer to any males, and now it is generalized (especially in the plural) to any group of people, including groups of females.

Narrowing is another type of language change where the word meaning gets *less inclusive* than its earlier meaning. Following is an example to demonstrate the narrowing of word sense: *Ammunition* originally referred to military supplies of all kinds, military supplies that explode, such as bullets and rockets.

1.3 Research goal

Consider a semantic neologism word w and suppose it had K different senses over some time period. Very compactly stated the research goal is to find how $P(w|t)$ is divided up in $P(w \text{ in sense } k|t)$ among different senses from K by unsupervised means.

Consider Fig 1.2 where the solid black-line is a hypothetical plot of $P(w|t)$, showing how a given word's probability might vary over time – the kind of plot easily producible using the Google n-gram viewer. If the word had K senses over its history, this plot is a superposition of K word-*sense* plots and Fig 1.2 shows a hypothetical decomposition into two senses (the dashed lines). In the hypothetical case shown, one of the senses emerged around 1970. Essentially the aim of this research is to propose a method which can carry out the kind of decomposition

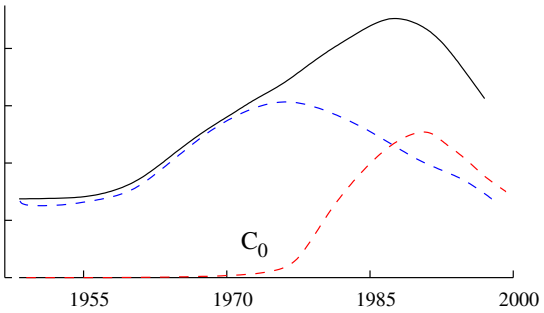


Figure 1.2 – Word frequency (solid line) and sense frequencies (dashed lines).

illustrated in this picture. Once we have $P(w \text{ in sense } k|t)$, one can detect whether it is a neologistic sense. For a semantic neologism you expect the neologistic sense to have probabilities close to 0 in the early years and then to go up in the later years. In the plot, the point at which this happens is marked as C_0.

For a formal neologism detecting its emergence is straightforward – it just requires the time-series of $P(w|t)$. For a semantic neologism you need the time-series $P(w \text{ in sense } k|t)$, for each k, which is not visibly directly in the data. To deal with this, the intention is to use an unsupervised model, that exploits our reasoning that the context words change when the sense emerges (the actual model is given in section 3.2). In a way, by looking at the context words the unsupervised model tries to imitate what an expert would do. This requires a dataset containing occurrences of a target word T each of which has a time-stamp. This task is closely related to word sense induction (WSI), but is different from WSI as it needs time-stamps.

1.4 Motivation

As lexical information is central to so many NLP tasks, means to automatically identify changes to the required information could be useful. For a semantic neologism, the emergence date of the neologism sense can be obtained based on the outcomes of the unsupervised algorithm. This emergence date information would presumably render word-meaning representations more accurate. Consider one of the language processing tasks, Statistical Machine Translation (SMT). In such systems, this might be helpful in predicting the mistranslations of such words. If an SMT system trained from aligned corpora from particular times is to be applied to text from different times it could be of use to know whether there have been sense changes, perhaps identifying which occurrences can be anticipated to be poorly translated. Table 1.1 provides examples of mis-translation via Google translate that relate to this.

In table 1.1, the English original[3], S1 comes from 1931, at which time *gay* simply meant 'happy'. In recent times, it has acquired a frequently used 'homosexual' sense. G1 is the German translation of S1 via Google translate (last executed on Apr 14, 2016), while T1 is

[3]This sentence comes from 'sons and lovers' by D.H. Lawrence

S1. With Clara, however, his brow cleared, and he was **gay** again (1931)

G1. Mit Clara, aber seine Stirn gelöscht, und er war wieder **Homosexuell**

T1. கிளாரா, ஆயினும், அவரது புருவம் அகற்றப்படும், மற்றும் அவர் மீண்டும் **ஓரினச்சேர்க்கையாளர்**

L1. clara, aayinum, avarathu puruvam akatrappadum, matrum avar meendum **Orinaccerkkaiyalar**

S2. Rooting my Android phone went well, but I've tried to flash a custom ROM and now I think I've **bricked** my phone (2011)

G2. Rooting mein Android-Handy ist gut gelaufen, aber ich habe versucht, eine benutzerdefinierte Flash-ROM und jetzt denke ich, dass ich mein Handy **gemauert** habe

T2. **என்** அண்ட்ராய்டு தொலைபேசியில் நன்றாக சென்றது வேர்விடும், ஆனால் நான் ஒரு தனிபயன் ரோம் ப்ளாஷ் முயற்சி மற்றும் இப்போது நான் **என்** தொலைபேசி **bricked** என்று நான் நினைக்கிறேன்

L2. en android tolaipesiyil nandraga sendrathu vaervidum, anaal naan oru tanippayan rome plash muyarchi marrum ippodhu naan en tholaippesi **bricked** endru naan ninaikkiraen

Table 1.1 – Sample English - German - Tamil translations via Google translate for sentences with words *gay* and *bricked*; S1, S2 are the original English sentences; G1, G2 are the German translations for S1, S2; T1, T2 are the Tamil translations for S1, S2 and; L1, L2 are the Transliterations for the Tamil text in T1, T2.

the Tamil translation of S1 and L1 is T1's transliteration[4] via Google translate. The word is translated as *Homosexuell* and *Orinaccerkkaiyalar* in German and Tamil respectively, in the 'homosexual' sense. The translations have probably gone wrong because the training data is *recent* and in it the 'homosexual' sense predominated. Following is another example translation demonstrating the newer usage of word with older training data. The English original, S2, comes from 2011, and uses the word *bricked* in its more recent 'render-inert' sense, but the German translation uses *gemauert* in the older sense of 'enclosing with bricks', while the less resourced language Tamil leaves the word *bricked* non-translated in T2 and L2. In this case the translations have probably gone wrong for the opposite reason, namely that the training data is not recent enough, and does not contain examples of the newer sense sufficiently often.

In addition to machine translation systems, the emergence date information may also be helpful in updating dictionaries. When the word 'stoned' is searched on wordnet[5], it produces an entry for the 'under drug influence' sense, but when the word 'brick' is when searched on

[4]This transliteration is a representation of the pronunciation of the original Tamil text using English.

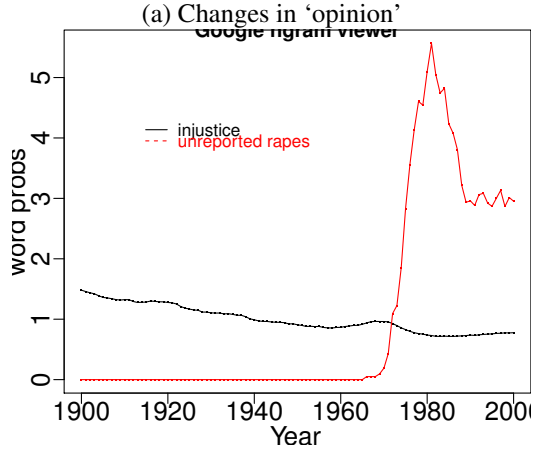
[5]The search outcome can be accessed from `http://wordnetweb.princeton.edu/perl/webwn?s=stoned&sub=Search+WordNet&o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&h=`

the online Wordnet[6] database, it does not produce any entry relating to the 'render inert' sense (a recent innovation). To deal with such dictionary incompletions, the diachronic analysis to discover the novel sense may be useful. As further examples, consider the information retrieval and question answering tasks, where the emergence date information could increase the precision of query disambiguation and document retrieval.

1.5 What affects word frequencies?

Both formal and semantic neologisms are examples of language change which takes the form of some kind of frequency change. It is worth noting that other things beside genuine language change will lead to frequency changes, roughly speaking changes 'in the world' and changes in people's attitudes or pre-occupations. To demonstrate this, let us first consider unambiguous words and phrases. Figure 1.3(a) shows frequencies[7] of the 2-gram *unreported rapes* and the word *injustice*. There is a dramatic change for *unreported rapes* and little change for *injustice*. Based on our knowledge, this does not reflect a language change or changes 'in the world', but more probably a change in people's opinion towards *unreported rape*. To illustrate changes 'in the world', figure 1.3(b) shows the word frequency plots for the 2-grams *America attacked, British attacked, Germans attacked, Japan attacked*. The first three 2-grams have an increase in frequencies around 1910 and a decrease followed by another increase around 1935. The fourth one has an increase just around 1935. Again this does not seem to reflect language changes or just changes in people's opinions but does reflect changes 'in the world' namely the world wars 1914 - 1918 and 1939 - 1945.

(a) Changes in 'opinion'



(b) Changes in 'World'

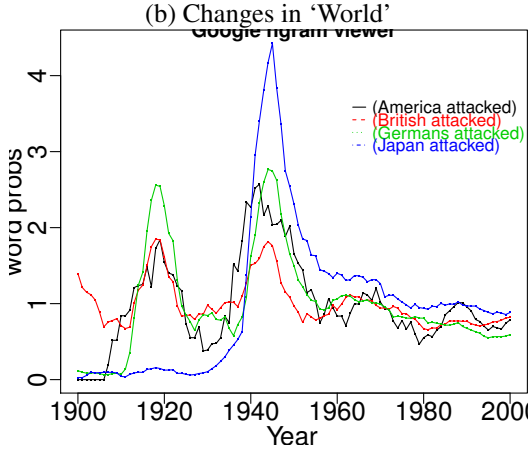


Figure 1.3 – Word frequency plots to demonstrate how (a) People's opinion affect word frequencies (b) Changes in the world affect word frequencies.

[6]The word 'brick' search on wordnet online can be accessed from `http://wordnetweb.princeton.edu/perl/webwn?s=brick&sub=Search+WordNet&o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&h=`

[7]This plot is based on data downloaded via Google-Ngrams API `https://github.com/econpy/google-ngrams`. This is further discussed in section 5.5.

Whilst the above looked at unambiguous words, presumably also for ambiguous words the frequency of use of a particular sense is not only a reflection of language change but also world change and opinion change. It's an interesting question as to whether these different influences can be distinguished automatically, but it is a question far beyond the scope of this research to do this. For the case of semantic neologisms, we will assume that an initial period of frequency close to zero and then a climbing frequency *is* a reasonably reliable indicator of *language* change.

1.6 Granularity of senses

Traditionally people distinguish polysemy from homonymy[Véronis, 2002]. When two or more words either sound the same (homophones), have the same spelling (homographs), or both, but have unrelated meanings, they are called *homonyms*. An example is the word *bank* which refers to 'a financial institution' and 'a river shore'. A word is *polysemous* if it can be used to express different meanings that are semantically related (may be obvious or subtle). An example for this is again the word *bank* referred in two different senses 'a financial institution' and 'as a place of safekeeping or storage' (as in a computer's memory bank).

It is not the case that we will be seeking to distinguish between homonymy and polysemy. One reason is that the distinction is not always straightforward. Relatedly it is notoriously difficult to say how many distinct senses a word has. Later on when experiments are performed on an ambiguous word, there will be a flexibility about the number of senses it might have. In particular no reference will be made to any particular dictionary to decide this. This pragmatic approach to homonymy, polysemy and number of senses seems to be a fairly widespread assumption made in this area of work.

1.7 Thesis plan

In this chapter, the research goal and motivation for this research was established. In the next chapter 2, first the problem evolution from static word sense induction to novel sense detection is introduced and then prior work related to this research is discussed. The alternative models and also alternative approaches to evaluation in such prior works are discussed. Some of the main points of contrast between this prior work and the approaches to be taken in the current work are noted and motivated.

In chapter 3 the theoretical proposal is set forth. As preliminaries to that there is a brief discussion on the (unsupervised) parameter estimation approaches Maximum Likelihood Estimation (MLE), Maximum A Posteriori (MAP) and Mean estimates for the unknown parameters in any given model. An outline to Expectation Maximization (EM) and Gibbs sampling techniques involved in estimating the parameters are presented. Then, a probabilistic model – call this a 'diachronic' model – is proposed in section 3.2 which conditions words in a target's context on that target's sense and conditions senses on times, making a simplifying assumption that

context words are independent of time given the target's sense. For the said model, the EM and Gibbs sampling algorithms are presented in further sections that will be used for experiments later. The relationships of this model to the model recently proposed in Frermann and Lapata [2016] are also noted in section 3.5.

This being an unsupervised task and no standard annotated datasets available, it is challenging to evaluate a sense emergence (neologistic sense) when it is identified by the model. From the literature review in chapter 2 this issue emerges as one where there is little consensus or reflection on the possibilities. Chapter 4 is concerned specifically with the issue of establishing the ground truth for sense emergence claims. Besides identifying strengths and weakness of approaches that have been previously attempted a proposed so-called 'tracks-based' method is put forward. In the experimental work the data-set used represents an extended sequence of *year-by-year* data, which is in contrast to much of the prior work. This presents some new issues in the identification of a sense emergence date from a time series and this is also addressed in section 4.1.4. Further, the possibilities to analyze parameter outcomes to verify the identified neologistic sense are also provided.

For this novel sense detection task, we require time-stamped raw text from a long timespan. The dataset possibilities for this task are analyzed in chapter 5, where informed decisions are also made in choosing the appropriate corpus for this work. There is also some discussion of the steps taken to arrive at a set of target items for the later experiments.

Chapter 6 reports and analyses all the experiments conducted on the Google 5-gram dataset. In chapter 6, a *pseudo-neologisms* technique is introduced (adapted from *pseudo-word* technique introduced by Schütze [1998]) to test the model's ability in identifying a neologistic sense. Following the success of the diachronic model in identifying a neologistic sense, a real set of experiments using EM and Gibbs sampling are conducted that include positive (semantic neologism) and negative (non-semantic neologism) targets. The said experiments are evaluated based on the evaluation possibilities introduced in chapter 4. There were also some semantic neologism targets such that the experiments on them did not produce an expected neologistic sense. An analysis on such targets are also reported and discussed in chapter 6.

Finally, chapter 7 presents the main concluding points of the thesis and discusses areas for future work.

Some of this research has been previously been published in:

Emms, Martin and Jayapal, Arun. Detecting change and emergence for multiword expressions. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 89–93, Gothenburg, Sweden, 2014. Association for Computational Linguistics

Emms, Martin and Jayapal, Arun. An unsupervised em method to infer time variation in sense probabilities. In *ICON 2015: 12th International Conference on Natural Language Processing*, pages 266–271, Trivandrum, India, December 2015

Emms, Martin and Jayapal, Arun. Dynamic generative model for diachronic sense emergence detection. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, page to be published, Osaka, Japan, December 2016

Chapter 2

Literature review

Languages change with time and lexical changes form an important part of that. In section 1.1, two forms of lexical changes were introduced, namely *formal* and *semantic* neologisms. Further in section 1.2 examples for how other lexical changes influence language were also provided. Then in section 1.3, the aim of this research in identifying the neologism sense for a semantic neologism word was established with a hypothetical plot of an ambiguous word decomposed into individual senses, one of which emerged at a particular time. The motivation (section 1.4) for this work was further established with real world examples that are affected by semantic neologisms.

Section 2.1 of this chapter introduces the evolution of research problem followed by a detailed discussion on the models and algorithms used in the prior work in section 2.2. Further in section 2.3, various evaluation schemes used in the prior work are discussed. Having said these, it is necessary to emphasize that in this chapter all the relevant works are discussed and the relevant comparisons of the current work with respect to the prior work are postponed to chapter 6, after we introduce our model and data, and experiments are published.

2.1 Problem evolution

With the research goal introduced in section 2.1, one may rightly perceive that the current work is related to word sense disambiguation (WSD) and word sense induction (WSI) tasks otherwise called as word sense discrimination tasks.

Word sense disambiguation (WSD) [Agirre and Edmonds, 2007, Ide and Véronis, 1998] is a classical natural language processing (NLP) task to determine the sense of a given ambiguous word from its context (phrase, sentence, paragraph, text). WSD can be considered as a classification problem – for a target word T (an ambiguous word that is polysemous or homonym[1]) in a sentence be disambiguated, WSD system requires an inventory of different sense usages and

[1]Words in sentences are associated to a particular sense where they appear. These words are ambiguous in nature because of the multiple meanings associated with them.

with this T can be assigned a sense usage. The target T is disambiguated based on its context[2], which determines the usage of T in a given sentence.

Word sense discrimination (or) induction (WSI) [Schütze, 1998] is closely related to WSD, but is different from WSD as this is independent of any annotated datasets or a word-sense inventory. Initially WSD was applied to the English language due to a vast amount of resources available for the language, but it was not possible to extend this to less-resourced languages, which led to the usage of WSI. Instead of relying on a dictionary of words or a sense annotated corpora, WSI deals with sense discrimination of T from a large data collection with T, by unsupervised means ie., WSI concerns automatic identification of senses of a given word. WSI is challenging as there is no dictionary of word senses available nor there is any supervision available (in terms of sense annotations) to discriminate data items with T.

As introduction in section 1.3, one may perceive that the proposed research is closely related to WSI as the neologism sense detection task is unsupervised. But this is different from WSI as we also consider the time of origin (publishing date) of the occurrence of T while discriminating the word's sense in an attempt to determine when a sense of T first comes into use in a time series

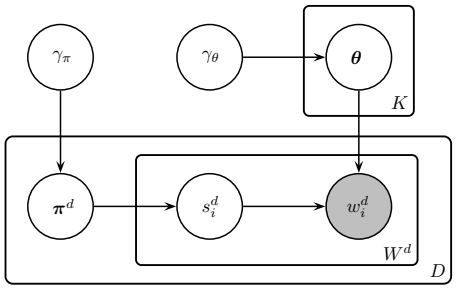
2.2 Models and algorithms in the prior work

For novel sense detection tasks there seem to be two widely used approaches in the literature, one involving generative models and the other distance based clustering, and these are discussed in this section. Some further approaches are then also discussed in section 2.2.3. In discussing these alternative approaches in this section, we mostly describe the concepts and algorithms involved. Some of the further details (such as data-sets used) will be taken up in later sections. For this work, each document d from a dataset $\boldsymbol{D}$ is considered to be a window of words W^d (contexts) around the target word T.

2.2.1 Approaches using probabilistic generative models

Topic models are probabilistic models used to automatically discover topics such as *medical science, computer, statistics, government* and *tax* from a collection of documents by unsupervised means. An LDA topic model (Blei et al. [2003], Griffiths and Steyvers [2004]) can be seen as a generative model of word sequences $\boldsymbol{w}$. In the context where these models were first developed, these sequences are the text content of documents. Where there are W^d words in a document d the model has W^d hidden so called *topic* variables $\{s_1, \ldots, s_{W^d}\}$. In its generative story, the values of these variables are first chosen, in each case choosing amongst K values, and then each word w_i^d is generated depending on s_i^d. The plate diagram for a topic model is given in figure 2.1.

[2]According to Firth [1957] – a famous phrase "a word is characterized by the company it keeps", one can understand the role of words associated with the ambiguous word, also called as *context words*.



γ_π : hyperparam of dirichlet distribution to sample for π
γ_θ : hyperparam of dirichlet distribution to sample for θ
$\boldsymbol{\pi}^d$: parameter for topics (topic probs in d)
s_i^d : topic for w_i^d
w_i^d : word i in document d
W^d : words in document d
D : number of documents
K : total number of topics (classes)
$\boldsymbol{\theta}_k$: parameter for words (word probs in topic k)
$\boldsymbol{w}^{1:D}$: set of all documents
V : all words in the vocabulary

Figure 2.1 – Plate diagram for Topic model (Latent Dirichlet Allocation)

As the plate diagram indicates, there is a per-topic word distribution $\boldsymbol{\theta}_{1:K}$, thought of as drawn from a Dirichlet prior[3]. For each document d there is a topic distribution $\boldsymbol{\pi}^d_{1:K}$, again thought of as drawn from a Dirichlet prior. Blei et al. [2003] uses variational EM for parameter estimation, and many papers since have used various Gibbs sampling methods.

This model has been adopted for WSI starting with Brody and Lapata [2009] and Yao and Durme [2011] more or less by treating 'topics' as senses. In the place of conventional documents, they have tiny 'documents' consisting of the context words $\boldsymbol{w}$ with a given target word T. Rather than 50 or more 'topics', K is instead in the region of around 10. Finally in order to have the model give a single sense to a target T the following final step is adopted. After estimation is complete, from each d the most-likely 'topic' sequence $\{s_1, \ldots, s_{W^d}\}$ is inferred. Then the 'sense' for item d is the value of k that occurs most frequently in this final topic sequence i.e., sense $=$ argmax(k) [freq(k in $s_1, \ldots, s_{W^d}$)]. So this approach equates the sense of target T with the most frequent topics in its context.

Whilst Brody and Lapata [2009] and Yao and Durme [2011] used topic models for WSI without any concern for change over time, the relevance to this thesis is that several people have used such topic models in connection with sense emergence. Rohrdantz et al. [2011] seems to be the earliest attempt of this kind, and uses the most basic LDA topic model ie., the one in plate diagram 2.1. This has been followed by Cook et al. [2013, 2014], Lau et al. [2012]. They use the Hierarchical Dirichlet Process (HDP) (Teh et al. [2004]) variant of LDA, which was first applied to WSI in Yao and Durme [2011]. Without attempting to describe HDP's details it has the feature that the number of topics can be inferred.

It may seem paradoxical that these kinds of models which make no mention of time could be applied to the problem of sense emergence. The way this is done is that first in the parameter estimation phase all time information in the data is ignored i.e., all data is *pooled*. Then second only after parameter estimation is completed, there is a phase of assigning a sense label to each data item. Finally one can look to see if the assigned senses are related to times, in particular that some particular sense is only getting assigned to *later* data items. This explains a general approach to applying topic-modeling to sense emergence, one that has been followed by Cook et al. [2013, 2014], Lau et al. [2012], Rohrdantz et al. [2011].

In the arena of topic modeling in documents, people have also sought to address the issue

[3] Background on Dirichlet prior's is given in chapter 3

that a 'topic' might have an interesting history. In work of that kind, Hall et al. [2008] point out the possibility of using exactly the same strategy of time-unaware topic-modeling that Cook et al. [2013, 2014], Lau et al. [2012] use: for training ignore time, and afterwards look at distributions of assigned labels in particular time-blocks. Hall et al. [2008] also point out the alternative, that of involving time in the modeling. In the setting of studying documents such approaches seem to be generally referred to as 'Dynamic Topic Models'. Wijaya and Yeniterzi [2011] is an example of work which seeks to exploit an existing dynamic topic model [Wang and McCallum, 2006] for the sense emergence task. Perhaps surprisingly this seems to be the only such piece of work, though in principle any dynamic topic model could be exploited. The dynamic topic model used by [Wang and McCallum, 2006] is called 'Topics over time' (TOT) [Wang and McCallum, 2006], a variant of LDA with a time parameter in the model to determine Topics over a certain time period. Somewhat differently to the other approaches they collapse all the contexts for a target T which share a time into single 'document' for that time. They did not really propose a sense emergence algorithm, but rather make some observations. For example in a 2-topic model they find a plausible increase of one particular sense of 'gay' at a particular time.

In the approach put forward in this thesis (section 3.2), time is involved in the modeling, with senses seen as having time-dependent probabilities. There is a recent article from Frermann and Lapata [2016], also concerning a model with time-dependent sense probabilities. It uses a model which is not a dynamic topic model as such but which takes a lot of inspiration from a proposal in that area Mimno et al. [2008]. Given a target word, their model has parameters[4] $\boldsymbol{\pi}_t$ for $P(S|Y)$ and $\boldsymbol{\theta}_{t,k}$ for $P(\boldsymbol{w}|S,Y)$ chosen from some discrete distribution, where $\boldsymbol{\pi}_t$ is the sense parameter that can capture senses that change over time. This model is in some ways like the one proposed in this thesis, and very briefly stated the contrast is that the model which is put forward in section 3.2 has $\boldsymbol{\pi}_t$ for $P(S|Y)$ and $\boldsymbol{\theta}_k$ for $P(\boldsymbol{w}|S)$, so words are treated as conditionally independent of time given a sense. These models were developed independently, with the model described in section 3.2 actually proposed earlier (see Emms and Jayapal [2014], Emms and Jayapal [2015]) than that in Frermann and Lapata [2016] ie. strictly speaking it is not *prior* work. The model used in Frermann and Lapata [2016] is conceptually a refinement of the model proposed here (though it was not developed as such). For that reason we will postpone making a clear exposition of the model's details to section 3.5, after the presentation of this thesis' model in section 3.2. It will be much clearer to see at that point the conceptual relationships between the two. The choices made concerning testing and evaluation in Frermann and Lapata [2016] are largely independent of the technicalities of their model proposal, and they will be turned to sooner in section 2.3.

[4] They have used different notations for the parameters, but are adapted here for the convenience of comparisons that will be made with the current work.

2.2.2 Approaches using clustering

Another possible approach involves distance-based clustering. This has been used in the work of [Mitra et al., 2014, 2015]. They divide the data into different eras and on each era they run a particular clustering algorithm. The outcome of this for each era is a set of clusters where each cluster is a set of words similar to a Wordnet 'synset' (Figure 2.2).

Time-period	Cluster ID	Words
1909–1953	C_{11}	*publishing, collection, editions, text, compilers, reprint, revision, author, copies, edition, authenticity* ...
	C_{12}	*novelist, poet, illustrator, proprietor, moralist, auditor, correspondent, reporter, editor, dramatist* ...
2002–2005	C_{21}	*administrator, clinician, listener, viewer, observer, statesman, teacher, analyst, planner, technician* ...
	C_{22}	*implementations, controller, program, preprocessor, api, application, specification, architecture* ...

Figure 2.2 – This is a screenshot from table 1 of Mitra et al. [2015]'s paper, showing clusters for the target 'compiler'

They use two different dependency parsed data sources (i) Google syntactic n-grams between 1520 and 2008 [Goldberg and Orwant, 2013] (ii) random tweets from twitter between 2012 and 2013 and construct a so called 'Distributional Thesaurus'(DT) – follows a method as provided in [Rychlý and Kilgarriff, 2007], from the word co-occurrence graph[5] based on syntactic bi-gram distribution across all times t and for each target T to track its sense change. Although they have considered a longer time span, they divide the time-line into time-groups containing equal amounts of data. To construct such a DT, a feature selection is done for each word using a so called 'Lexicographer Mutual Information' (LMI) [Kilgarriff et al., 2004] metric and the top 1000 features are retained. The clustering is done from these feature sets for each time-group by first constructing a neighborhood graph and the graph is clustered using the 'Chinese Whispers' algorithm (as provided in [Biemann et al., 2013]).

They then have to relate the sense-representing clusters from different eras to each other. One aspect of this is a criteria they propose which identifies a cluster as a 'sense birth'. Consider sense clusters $s_1^{N_1}, \ldots, s_n^{N_1}$ from an earlier era N_1 and $s_1^{N_2}, \ldots, s_n^{N_2}$ from a later era N_2. When a sense cluster $s_i^{N_2}$ appearing in N_2 has sufficiently few of its member words belonging to any of the clusters for the earlier era N_1, they count as $s_i^{N_2}$ as sense birth. Figure 2.3 shows a screenshot of the schematic shown in figure 2 of Mitra et al. [2015] explaining the birth of a new sense, where S_n provided with a green circle to show a new cluster that was did not exist in an earlier time.

Referring to figure 2.2, by their criteria the cluster C_{22} is a sense birth for the target *compiler*. This approach does not attempt to annotate the target's occurrences with a sense label,

[5] In a word co-occurrence graph, words are denoted by nodes, and there exists an edge between two nodes, if the corresponding words co-occur in a sentence.

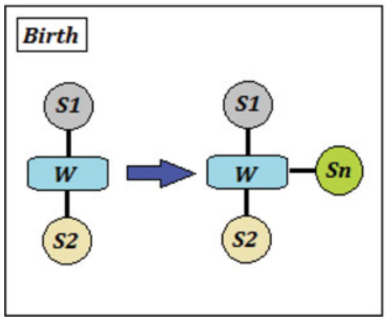


Figure 2.3 – This is a sreenshot of the the schematic shown in figure 2 of Mitra et al. [2015] depicting the birth of a new sense.

and it does not seem easy to make it do so.

This is a point of contrast to the current thesis proposal and the approaches based on topic modeling, which both being generative models, inevitably do have this instance-labeling feature, which is arguably a desirable one.

2.2.3 Other approaches

Kulkarni et al. [2014]'s work primarily concerns the detection of broadening and narrowing of word senses, however they are also concerned with sense emergence. To model the temporal evolution of language, they construct a time series per T using three different methods to define a time-dependent statistic for the target. One method simply uses the frequency of the target word T. In a second method, the time-specific statistic is the distance between the target word's part of speech (POS) distributions in successive time periods. In the third method, they make time-specific vectors – call this $\boldsymbol{v}_t$ – based on the time-specific co-occurring words. Then in this case the time-specific statistic they use[6] is $T_t = 1 - \text{cosine}(\boldsymbol{v}_t, \boldsymbol{v}_0)$, that is a comparison to the vector for the first time point.

Without getting into all the details of their work, their vectors $\boldsymbol{v}_t$ are not just the raw co-occurrence statistics, but are derived using the 'word-embeddings' [Mikolov et al., 2013] technique. Figure 2.4 shows a screen-shot of the time-series via this third method for the word *tape*. They suggest the pattern in this plot reflects the introduction of magnetic tape in the 1950's and its prevalence by the 1970's.

The most striking difference to the other techniques already discussed is that those had some form of sense representations, whilst none of the alternatives in Kulkarni et al. [2014] seek to make any kind of sense representations. In some sense it is concerned with what is detectable once senses have been aggregated over. Correspondingly the techniques used arguably are not *unsupervised* machine learning techniques.

Kim et al. [2014] is another approach using almost identical techniques.

Tang et al. [2015] has an aim to detect sense emergence as well as other kinds of semantic change. Their work is applied to the Chinese language. Where $\boldsymbol{w}$ is the context words around

[6] An alternative would again be to make a time series of difference between consecutive vectors.

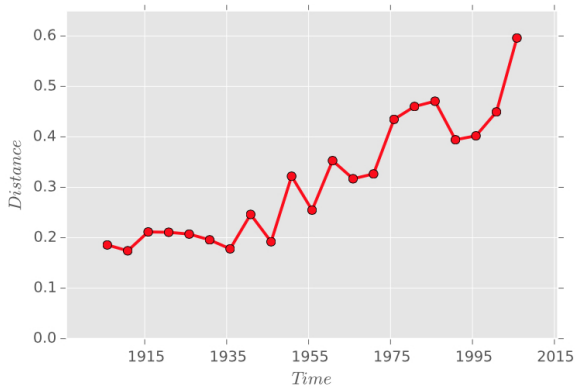


Figure 2.4 – This is a screen-shot of the plot for the word *tape* provided in figure 5 from Kulkarni et al. [2014]'a paper

a particular occurrence of a target T, they adopt the unusual starting position of identifying the 'sense' of this occurrence with that word $w \in \boldsymbol{w}$ which maximizes a particular 'association score' $assoc(T,w)$. Without going into the details of this score, this means there are as many potential 'senses' for the target T as there are words in the vocabulary V. Arguably this conflates a feature selection technique with a sense labeling technique. Also noticeably this sense concept does not involve any unsupervised machine learning.

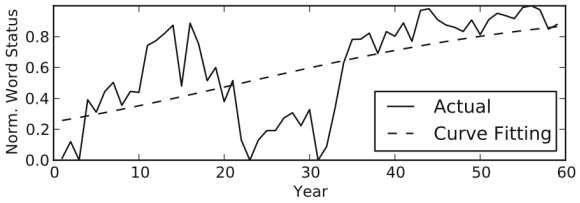


Figure 2.5 – This is a screen-shot of the plot for the word *tou ming* provided in figure 6 from Tang et al. [2015]'s paper

They define what they call a 'aggregative' approach and a 'segregative' approach. In the aggregative approach, they compute the entropy of the sense distribution in a given time and plot this in a time-series, analogous to Kulkarni et al. [2014]. Figure 2.5 shows the time-series of this kind for the word *tou ming*. The time series plot is rather jagged and a further part of their proposal involves fitting a smoother line to it.

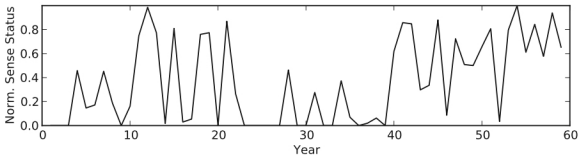


Figure 2.6 – This is a screen-shot of the plot for a sense of the word *tou ming* provided in figure 7 from Tang et al. [2015]'a paper

In the segregative approach, they are interested in the time-line of individual senses. Before doing this they attempt to merge the "senses" as defined above. They use a particular clustering

algorithm for this, though not an entirely unsupervised one as it refers to a Chinese dictionary of synonyms. Even after the clustering they can still have as many as 490 senses of a single target. Figure 2.6 shows the time-series plot for a sense obtained in this way for the word *tou ming*. Again part of their proposal is to fit a smoother line to it.

They seek to use parameters of their fitted lines to make observations about sense dynamics. Its hard to assess the success of this on their Chinese examples. However, it seems fair to say that their approach like Kulkarni et al. [2014]'s approach does not seem to be intrinsically able to label instances of the target with senses.

2.3 Evaluation in the prior work

In this section, the evaluation approaches used in the relevant prior work are discussed: the different algorithmic approaches used in prior work were already discussed in the previous section 2.2.

Recall [Rohrdantz et al., 2011] used an LDA model for their work on novel sense detection. They seem to have considered 7 topics to find the sense emergence from a range of years considered for their work. For all their topics, their occurrence over time is visualized over time (years) as a form of density plot. This way they predict a novel sense when a particular topic has dense areas identified in the later years and not during the initial period. Their approach to assessing the correctness of the prediction is using a 'dictionary first inclusion' approach so they compare the date of an apparent sense emergent with a date based on when that sense was first included in a dictionary. In particular they referred to multiple dictionaries from different periods (Longman Dictionary from 1987, the English WordNet4 [Fellbaum, 1998] and 2007 Collins dictionary) to verify the emerging sense information of the identified novel sense. There is a detailed discussion about this kind of approach to ground truth in section 4.1.2.

[Cook et al., 2013, 2014, Lau et al., 2012] also had an approach based on topic modeling (section 2.2). For their work, they did not consider a long time span but rather considered two datasets, a focus corpus f_c and a reference corpus r_c, from later and earlier times respectively. The BNC [Burnard, 2007] corpus that has data representing late 20th century has been used as r_c and the ukWac [Ferraresi et al., 2008] from 2008 as f_c. [Cook et al., 2013, 2014, Lau et al., 2012] have chosen their neologism targets by comparing successive editions of Concise Oxford English dictionary (1995, 2008). So they too are adopting the 'dictionary first inclusion' approach to ground truth. They evaluated success in the following way. They have both known neologisms and known non-neologisms and for each target they define a novelty score where they compute a ratio of the inferred sense frequencies $p_2 : p_1$ belonging to $r_c : f_c$ times. The evaluation then takes the form of seeing to what extent neologism targets rank higher than non-neologisms by this score.

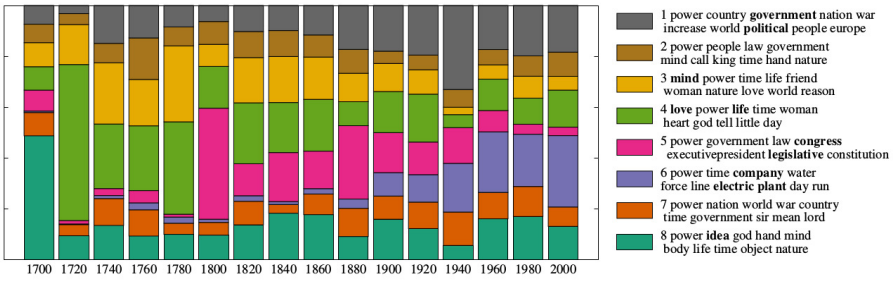
Recall that Mitra et al. [2014, 2015] adopted a clustering approach to identify novel sense – when a cluster $s_k^{N_2}$ is generated for a particular time period N_2 and is counted as not the same as any cluster $s_{k'}^{N_1}$ from an earlier time period N_1, they count this a sense 'birth'. They

divide the available data into 8 eras of equal data size and ever diminishing time-span: 1520-1908, 1909-1953, 1954-1972, 1973-1986, 1987-1995, 1996-2001, 2002-2005, 2006-2008. So 'sense birth', as they term it, can be assessed relative to any pair of eras drawn from this set. For evaluation purposes they considered the inferred 'sense births between the eras N_1 = 1909-1953 and N_2 = 2002-2005.

In [Mitra et al., 2014] to verify an inferred sense-birth they adopted the very simplest of all ground truth approaches: reference to the intuitions of one of the authors. In the later version of the paper Mitra et al. [2015] they indicate a switch to a 'dictionary first citation' approach to ground truth, so the date of the earliest citation for the sense in a dictionary (This general approach to ground truth be discussed further in section 4.1.2). Rather than using the OED they get their dating information from the 'Online Etymology Dictionary' `http://www.etymonline.com`, an online source created by Douglas Harper. They themselves describe the source as the online `dictionary.reference.com` but this itself is a secondary source that pulls information from other sources and the dating information in fact comes from the above-mentioned `http://www.etymonline.com`. Surprisingly the reported verification of sense emergences in the two papers, one due to author intuition and one due to use of dictionary first citations, are identical.

In addition to the dictionary based evaluation conducted, they propose an evaluation of the dating using WordNet. To do this they first define a mapping (based on amount of shared words) from any of their clusters to a Wordnet sense-id. Then they check whether when a sense cluster $s_k^{N_2}$ at N_2 has sufficiently little overlap with the clusters at N_1 – ie. that it is a 'sense birth' in their terms – that it is also the case that the mapped image of $s_k^{N_2}$ is also distinct from all the mapped images of the clusters at N_1. If this is so they count this as verifying that the $s_k^{N_2}$ sense truly only emerged in the later time period N_2. It is not actually clear it is reasonable to argue that this tells you something about the correct *dates*. It seems to beg the question that the algorithm was definitely correct to not have found a cluster in the earlier N_1 data that is sufficiently similar to the cluster $s_k^{N_2}$ found in the later N_2 data. Suppose one took a word definitely known to not have acquired a new sense between times N_1 and N_2, and suppose the system generates a N_2 cluster $s_k^{N_2}$ sufficiently different to all N_1 clusters. If the mapped WordNet image of $s_k^{N_2}$ is different to the mapped images for the all the N_1 clusters you would be forced to conclude the system is correct to predict a new sense for the word at N_2. So this approach at least arguably is not giving an evaluation of dating information but rather it seems to verify to some extent that discovered sense distinctions are real in the sense of being correlated with different WordNet ids.

In Frermann and Lapata [2016] one of the the model outcomes is a vector, $\boldsymbol{\pi}_t$ of sense probabilities given times, $P(S|Y)$. A possibility is to inspect this for evidence of sense emergence. In Fig 4 of their paper, for 4 target words there is a display of this parameter. The picture below reproduces this for their target *power*



About the 8 senses they say

> *three senses emerge: the institutional power (colors gray/1, brown/2, pink/5, orange/7 in the figure),mental power (yellow/3, lightgreen/4, darkgreen/8), and power as supply of energy (violet/6)*

They go on to suggest that the observed pattern for sense 6 (violet) is indicative of a "sense birth", and use a dictionary first citation date of 1889 from the OED to corroborate this. This indicates that they see one possibility for evaluating the outcomes is to look at the trajectory of the per-year sense probabilities and check for correlations with some kind of ground-truth about sense-emergence, such as first-citation date in the OED. However, they do not pursue this possibility very much. Possibly this is due to the fact the focus is not exclusively, or even primarily, on sense *emergence*, but also on sense *change* – the idea that a sense is constantly there but is itself changing.

Another of their evaluation approaches attempts to mimic the WordNet-based approach of Mitra et al. [2014, 2015] outlined above. In place of comparing the sense-representing clusters at T_1 and T_2 for 'birth' candidates they develop a *novelty* score for a sense k at T_2. This, curiously, is not based on the contrast of the probability of the sense at T_1 and T_2, but is instead based on (i) the word distribution for the sense at T_2 and (ii) a set of top-1000 words W most distinguishing of T_2 vs T_1, defined essentially by the ratio of their frequencies in the data from these times. Their novelty score for a sense is $\sum_{w \in W} P(w|S = k, t = T_2)$, and for a word is the maximum over the sense-based scores. They then also have a mapping from the word-distributions associated with a sense k to WordNet ids. Following an analogous logic to Mitra et al. [2014, 2015], for words with a high novelty score at T_2 they then check whether for the responsible sense k at T_2 that it is also the case that the mapped image of k (at T_2) is also distinct from the mapped images of all the other senses k' (at T_1). Again it is not clear that this straightforwardly verifies the *dating* of a possible sense emergence rather than confirming that the systems sense distinctions can correlate with WordNet distinctions. In the end they observe that they obtain a success rate 'in the same ballpark' as Mitra et al. [2014, 2015], attaching to it the caveat that *scores are not directly comparable due to differences in training corpora, focus and reference times, and candidate words*. This fair caveat will be a recurring them in comparisons between different pieces of research on this topic.

Briefly noting their other approaches to evaluation, in a further experiment they refer to a list of 100 words which were rated by human annotators in Gulordava and Baroni [2011] for

their degree of semantic *change* between the 1960s and 1990s: they see if the above-noted novelty scoring of words yields a ranking which correlates with human ratings for change. Note the annotation concerns change, and not sense-emergence per-se. They mention also one further 'evaluation' which used their model in a particular way to participate in the Diachronic Text Evaluation[7] (DTE) task conducted by SemEval 2015. However, this task consists of determining the period when a text was written, so does not measure success in identifying sense emergences.

2.4 Discussion

Possibly because of the relative novelty of the area of applying machine learning to sense emergence, the research works reviewed in the preceding sections do not use the same data-sets, test items, notions of ground truth or performance criteria. Nor do they propose models or algorithms which are consciously put forward as evolutions of each other. This makes a traditional critique in terms of quantitative performance and/or relations amongst the models or algorithms difficult. Nonetheless there are some main points to be made about how the work to be proposed here relates to this prior work.

A noteworthy characteristic in almost all of the work just reviewed in the preceding sections is that they apply some form of sense induction (call this SI) algorithm which is *time-unaware*. One design option is to *pool* all training data for the SI phase, then assign the likeliest senses to examples, and then to finally check for a correlation with time. In this design option, the sense assignments are static in nature as the SI algorithm is time-unaware. The system discussed in Lau et al. [2012] and Cook et al. [2013, 2014] used precisely this option. Another *split-then-relate* design option is to separate the data into eras, perform independent SI on each subset and then seek to consider how the sense representations from each era may (or may not) be identified with each other. Mitra et al. [2014, 2015] used precisely this option. These two pieces of work represent just one way to instantiate these two strategies for exploiting a time-unaware SI system, and one simple research direction in sense emergence systems could be simply to plug other time-unaware SI systems into these *pool* or *split-then-relate* approaches.

This is *not* the direction that will be followed here. Instead we will be introducing a *time-aware* probabilistic model, a crucial feature of which will be that it has parameters which represent the probability of a sense at a time, $P(S|Y)$. The aim will be to evaluate by considering the estimated values of this parameter. Such an approach has not been much explored. The principal related piece of work is [Frermann and Lapata, 2016]. As noted in section 2.2 this work is best seen as a possible logical development of the model discussed here, even if it was not developed in this way. As such we have left the discussion of this model to section 3.5, after the presentation of this thesis' model in section 3.2, so that a meaningful comparison can be made.

Ground-truth and evaluation is difficult for a sense emergence system. The review of the

[7] `http://alt.qcri.org/semeval2015/task7/`

prior work in the preceding sections seems to indicate a lack of consensus about the best or proper way to go about this. For this reason chapter 4 constitutes a dedicated discussion of this topic, something which arguably has not been done before. It reviews all possible approaches to ground-truth, noting their strengths and weaknesses. Additionally in section 4.1.3 we also propose an additional so-called 'tracks' method, on the basis of which to establish a ground-truth about sense emergence in a particular corpus.

The 'pseudo-word' technique was introduced by Schütze [1998] as a way to test word-sense discrimination model without requiring large scale annotation. As an additional contribution in the area of evaluating sense emergence systems a development of this will be proposed in section 6.2 to give a diachronic version which will be called 'pseudo-neologisms'. It could be applied to any sense emergence system, and will be applied to the system proposed here.

In the experiments that are conducted the data-set used represents an extended sequence of *year-by-year* data. This contrasts to the approach of Cook et al. [2013, 2014], Lau et al. [2012], which has two eras of data, and to the approach of Mitra et al. [2014, 2015], who have 8 eras of ever diminishing size. This presents some new issues in the identification of a sense emergence date in a time series. These will be addressed in section 4.1.4.

Chapter 3

Research theory

In the previous chapter 2, we discussed the various models and evaluation schemes used in the prior work. We try to address the shortcomings from the prior works in our current work. Before we actually introduce our probabilistic model, we provide a brief discussion on the different 'parameter estimation essentials' (section 3.1) such as 'Parameter estimation approaches', 'Dirichlet priors' followed by an outline to 'Expectation Maximization' (EM) and 'Gibbs sampling' procedures involved in estimating model parameters. The actual model, a so-called 'diachronic model' is introduced in section 3.2 that can discover neologism sense from a given dataset. For the introduced diachronic model, the parameter updates are derived for EM and Gibbs sampling procedures in sections 3.3 and 3.4; and algorithms using these updates are also provided. In section 3.5 there is a detailed discussion of the model of Frermann and Lapata [2016], identifying its points of overlap and contrast with the model presented here.

3.1 Parameter estimation essentials

In this section various parameter estimation approaches are discussed in brief. Further there is discussion on the Dirichlet distribution, which is often used as a parameter prior. Additionally, there is a brief discussion providing the outline of Expectation Maximization (EM) and Gibbs sampling parameter estimation schemes.

3.1.1 Parameter estimation approaches

Consider a dataset $\boldsymbol{D} = \{d^1, d^2, \ldots, d^n\}$ with n data items, where each d^i is defined by some set of variables and there is a model $P(\boldsymbol{D}; \Theta)$ where Θ is some parameter setting of the model (eg. probabilities of various settings of various variables). There are a variety of approaches to estimate the parameter values [Heinrich, 2004]. It is a natural idea to seek to find values for Θ that *maximizes the likelihood* $P(\boldsymbol{D}; \Theta)$. This is the so called Maximum Likelihood Estimate (MLE)

$$\Theta_{MLE} = \arg\max_{\Theta} P(\boldsymbol{D};\Theta) \tag{3.1}$$

The Bayesian approach to parameter estimation assumes the parameters are a further kind of unknown having their own distribution, its *prior* $P(\Theta;\gamma_\Theta)$, where γ_Θ is some hyper-parameter setting for Θ. There is then the joint probability $P(\boldsymbol{D};\Theta) \times P(\Theta;\gamma_\Theta)$ and via this a posterior on Θ given the data $P(\Theta|\boldsymbol{D})$, given by

$$P(\Theta|\boldsymbol{D};\gamma_\Theta) = \frac{P(\boldsymbol{D}|\Theta) \times P(\Theta;\gamma_\Theta)}{P(\boldsymbol{D};\gamma_\Theta)} \tag{3.2}$$

Another natural estimate is the *mode* of this posterior $P(\Theta|\boldsymbol{D};\gamma_\Theta)$. The denominator $P(\boldsymbol{D};\boldsymbol{\gamma}_\theta)$ given by $\int_\Theta \{P(\boldsymbol{D}|\Theta)P(\Theta;\gamma_\Theta)\}$ is a normalizing constant. So the mode is also just the maximizer of the numerator. The so-called Maximum A Posteriori (MAP) estimate is defined by Θ_{MAP}:

$$\Theta_{MAP} = \arg\max_{\Theta} P(\Theta|\boldsymbol{D};\gamma_\Theta) = \arg\max_{\Theta} P(\boldsymbol{D};\Theta) \times P(\Theta;\gamma_\Theta) \tag{3.3}$$

In other words a MAP estimate is equivalent to maximizing the product of likelihood and prior. In the case of uniform prior, the MAP estimate works out to be a Maximum Likelihood estimate.

Given the existence of the posterior $P(\Theta|\boldsymbol{D};\gamma_\Theta)$ it also natural to consider its *mean*, as another point estimate[1]:

$$mean(\Theta) = \int_\Theta P(\Theta|\boldsymbol{D};\gamma_\Theta) \times \Theta \tag{3.4}$$

All three types of estimates (equations 3.1, 3.3, 3.4) have some motivation. A practical question however is whether in any given case are there ways to determine their values. In certain simple cases there are closed form formulae for these but in most cases of interest there are not. Instead there are various algorithms which compute approximations of them, principally Expectation Maximization (section 3.1.3) for MLE and MAP estimates and Gibbs sampling (section 3.1.4) for mean estimates. These parameters can be computed by iterative re-estimation procedures.

Where does prior come from? The prior is a non-negligible part of Bayesian inference, while it is at user's discretion to chose a type of prior. A prior distribution is supposed to represent knowledge about the distribution of parameters prior to the experiment outcomes. But how to choose a prior for the experiment? When there is no or little knowledge about the distribution of parameters prior to the experiment outcomes, one can choose a *non-informative prior*, where the prior distribution is uniform. As noted above, this makes MAP estimate equivalent to a ML estimate.

[1] One may also want to consider representations of the spread about the mean.

There are certain prior distributions in use, expressible by a parameterized formula. One of these is Dirichlet which seems to allow a variety of intuitions to be expressed and has desirable computational properties – some of those are described in the following section 3.1.2.

3.1.2 Dirichlet priors

The Dirichlet distribution can be considered as the multi-variate generalization of the *beta* distribution. The density function of beta distribution is defined in (equation 3.5), where x is the random variable and α, β are the parameters.

$$f(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{3.5}$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is a beta function.

We can see the Dirichlet distribution as a distribution over distributions. Let us consider a K dimensional vector of probabilities $\boldsymbol{x} = [x_1, x_2, \ldots, x_K]$ where $\sum_{k=1}^{K} x_k = 1$. The distribution is parameterized by $\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_K]$. The density function of a Dirichlet is given by

$$\boxed{f(x_1, x_2 \ldots x_K; \alpha_1, \alpha_2 \ldots \alpha_K) = \frac{1}{\beta(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1}} \tag{3.6}$$

where,
(i) $\beta(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K}\alpha_k)}$, (ii) $\sum_{k=1}^{K} x_k = 1$ and $0 < x_k < 1$, (iii) $\Gamma(n)$ denotes the gamma function, and (iv) $\boldsymbol{\alpha}$ is a vector of parameters and each value in the vector is greater than 0. One of the definitions of the gamma function is that for real numbers $x > 0$, $\Gamma(x) = \int_0^\infty y^{x-1} e^{-y} dy$, from which follows a recurrence property that $\Gamma(x) = (x-1)\Gamma(x-1)$ and $\Gamma(1) = 1$ that will be used below[2].

It is important to note that the 'normalizing' constant $\beta(\boldsymbol{\alpha})$ appearing in the definition of the Dirichlet is the integral of the main product term ie.

$$\beta(\boldsymbol{\alpha}) = \int_{\boldsymbol{x}} \prod_{i=1}^{k} x_i^{\alpha_i - 1} \tag{3.7}$$

Following is a discussion of some of the properites of of the Dirichlet distribution, including its mean and mode.

3.1.2.1 Mean

Before the expectation or mean of the Dirichlet distribution is formulated, the expectation for the beta distribution is derived as the Dirichlet distribution is considered the multi-variate

[2]When applied to integers the recurrence property entails that the gamma functoin can be expressed in terms of the factorial function ie., $\Gamma(n) = (n-1)!$

generalization of beta distribution[3].

If X has a beta distribution the expectation of X is given by

$$\mathbb{E}[X] = \int_0^1 \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} x.dx$$

where $B(\alpha,\beta)$ is a beta function given by $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

$$\mathbb{E}[X] = \frac{1}{B(\alpha,\beta)} \int_0^1 x^{\alpha}(1-x)^{\beta-1} dx$$

$$= \frac{1}{B(\alpha,\beta)} \int_0^1 x^{(\alpha-1)+1}(1-x)^{\beta-1} dx$$

$$= \frac{1}{B(\alpha,\beta)} B(\alpha+1,\beta) \text{ (by integral representation of beta function)}^4$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \text{ (from definition of beta function)}$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+1)} \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \text{ (re-arranging from previous step)}$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta)(\alpha+\beta)} \frac{\Gamma(\alpha)(\alpha)}{\Gamma(\alpha)} \text{ (gamma's property } \Gamma(z) = \Gamma(z-1)(z-1))$$

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$$

Expectation for Dirichlet: The Dirichlet is a distribution over K dimensional vectors. The (the j-th component) of the expectation or mean of the distribution is given as,

$$\mathbb{E}[x_j] = \frac{\boldsymbol{\alpha_j}}{\sum_k \alpha_k} \tag{3.8}$$

This derivation is similar to the one used to derive the expectation of the beta distribution, and

[3]The proof closely follows the proof from `http://www.statlect.com/beta_distribution.htm` – last executed on Jun 10, 2016.

[4]The integral representation of beta function is provided at `http://www.statlect.com/subon2/betfun1.htm` – last executed on Jun 10, 2016.

is given below for x_1.

$$\mathbb{E}[x_1] = \int_{\boldsymbol{x}} \frac{1}{\beta(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1} x_1 . d\boldsymbol{x}$$

$$= \frac{1}{\beta(\boldsymbol{\alpha})} \int_{\boldsymbol{x}} \prod_{k=2}^{K} x_k^{\alpha_k - 1} x_1^{(\alpha_1 + 1) - 1} . d\boldsymbol{x}$$

consider $\boldsymbol{\alpha'}_{k \neq 1} = \boldsymbol{\alpha_k}$ and $\boldsymbol{\alpha'_1} = \boldsymbol{\alpha} + 1$

$$\mathbb{E}[x_1] = \frac{1}{\beta(\boldsymbol{\alpha})} \int_{\boldsymbol{x}} \frac{\beta(\boldsymbol{\alpha'})}{\beta(\boldsymbol{\alpha'})} \prod_{k=1}^{K} x_k^{\alpha'_k - 1} . d\boldsymbol{x}$$

$$= \frac{\beta(\boldsymbol{\alpha'})}{\beta(\boldsymbol{\alpha})} (1)$$

$$= \frac{\prod_{k=2}^{K} \Gamma(\alpha_k) \Gamma(\alpha_1 + 1)}{\Gamma(\sum_k \alpha_k + 1)} \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)}$$

using gamma's property $\Gamma(z) = (z-1)\Gamma(z-1)$ a few times we arrive at

$$\mathbb{E}[x_1] = \frac{\alpha_1}{\sum_k \alpha_k}$$

As nothing crucially depended on choosing x_1 in this, we obtain in general the mean of the Dirichlet distribution

$$\boxed{\mathbb{E}[x_j] = \frac{\alpha_j}{\sum_k \alpha_k}}$$

3.1.2.2 Mode

The mode is the value of the distribution where the density is the highest. Its formula for the Dirichlet is (assuming all $\alpha_k > 1$):

$$x_k = \frac{\alpha_k - 1}{\sum_k (\alpha_k - 1)}, \tag{3.9}$$

Here, the mode for the beta distribution is first derived and then extended to the Dirichlet distribution as the Dirichlet distribution is the generalization of the beta distribution.

To find maxima or minima of the beta distribution, we get the derivative of the density function (equation 3.5) with respect to x, set it to zero and solve for x.

$$\frac{\partial}{\partial x} f(x; \alpha, \beta) = \frac{\partial}{\partial x} \frac{1}{B(\alpha, \beta)} x^{\alpha - 1} (1 - x)^{\beta - 1} = 0$$

Considering the beta function $\frac{1}{B(\alpha,\beta)}$ as a constant C,

$$\frac{\partial}{\partial x} f(x;\alpha,\beta) = C\left((\alpha-1)x^{\alpha-2}(1-x)^{\beta-1} + x^{\alpha-1}(-1)(\beta-1)(1-x)^{\beta-2}\right) = 0$$

$$\Rightarrow (\alpha-1)x^{\alpha-1}x^{-1}(1-x)^{\beta-1} - x^{\alpha-1}(\beta-1)(1-x)^{\beta-1}(1-x)^{-1} = 0$$

$$\Rightarrow x^{\alpha-1}(1-x)^{\beta-1}\left[\frac{\alpha-1}{x} - \frac{\beta-1}{(1-x)}\right] = 0$$

$$\Rightarrow x^{\alpha-1}(1-x)^{\beta-1}\left[\frac{(1-x)(\alpha-1)-(\beta-1)x}{x(1-x)}\right] = 0$$

$$\Rightarrow (1-x)(\alpha-1)-(\beta-1)x = 0$$

Now, solving for x, we get

$$x = \frac{\alpha-1}{\alpha+\beta-2}$$

For example, for $\alpha = \beta$, this gives a solution at $x = \frac{1}{2}$, and if $\alpha > 1, \beta > 1$, this is a maximum, that is the *mode* of beta distribution[5].

Analogous derivations can be carried out for Dirichlet distribution. The beta distribution has the two parameters α and β. Analogous derivations can be made for the Dirichlet with its $K \geq 2$ parameters in $\boldsymbol{\alpha}$, leading to an analogous expression for a stationary point:

$$\boxed{\boldsymbol{x}_k = \frac{\boldsymbol{\alpha}_k - 1}{\sum_k(\alpha_k - 1)}}$$

As with the beta distribution, if all components of $\boldsymbol{\alpha}$ are equal and greater than 1, the above is a maximum, that is, the mode of the distribution[6].

3.1.2.3 Dirichlet as conjugate prior

By now we know that the Dirichlet distribution is a generalization of the beta distribution. As beta distribution is used as prior for the binomial distribution, it is a good idea to use Dirichlet distribution as prior for multi-nomial distribution. Following is the formulation to prove multi-nomial and Dirichlet distributions form conjugate prior[7]. The multinomial distribution is

[5] If $\alpha < 1, \beta < 1$ this is minimum.

[6] If all are equal and less than 1, it is a minimum. Some of the possibilities are illustrated in Figure 3.2

[7] A conjugate prior of a likelihood function is the prior when both posterior and prior distributions are of the same distribution

defined by,

$$\text{multi}[x;\theta] = \frac{(\sum_{k=1}^{K} x_k)!}{\prod_{k=1}^{k}(x_k!)} \prod_{k=1}^{K} \theta_k^{x_k} \tag{3.10}$$

where the parameters θ are the probability values to get into one of the K-categories. Now, the posterior probability function is defined to be:

$$p(\theta|x) = \frac{p(x;\theta)p(\theta)}{p(x)} \tag{3.11}$$

But, as $p(x)$ will act as a normalizing constant, this can be excluded and the posterior probability distribution is rewritten to be

$$\begin{aligned} p(\theta|x) &= p(x;\theta)p(\theta) \\ &= \text{multi}[x;\theta]\ \text{Dir}(\theta|x) \\ &\approx \prod_{k=1}^{K} \theta_k^{x_k} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \\ &\approx \prod_{k=1}^{K} \theta_k^{(x_k + \alpha_k - 1)} \end{aligned}$$

$$\boxed{p(\theta|x) = \text{Dir}(x+\alpha)}$$

This section followed the report by Huang [2005]. There are the formulae for the mode and mean (sections 3.1.2.2 and 3.1.2.1) of the posterior Dirichlet's. So in the case of data in the form of counts amongst K alternatives then the mode and mean of the posterior on $\boldsymbol{\theta}$ (discussed in section 3.1) assuming a Dirichlet prior for $\boldsymbol{\gamma}_\theta$ can be just calculated exactly. This kind of neat arrangement disappears once hidden variables are introduced in the model.

3.1.2.4 Impact of Dirichlet prior

Now, consider a collection of documents with each document consisting of a few words and the total vocabulary size is 10. One could assign probability of the words appearing in each document through random draws from a Dirichlet distribution.

A short demonstration of Dirichlet distribution for the said document collection is discussed here with plots. It is difficult to visualize plots that are greater than 2 or 3 dimensional. Therefore, two dimensional and three dimensional plots are provided to depict the behavior of Dirichlet distribution. Figure 3.1 provides 15 random draws at each of four different α settings and the α's are considered symmetric for these plots. The graph has the words plotted on x-axis and the probability of the words on the y-axis. From the plots, the following can be observed (i) as the alpha goes below 1, the graph gets uneven and sparse (ii) for each draw, the same distribution is maintained but the proportion of words are not always same for each draw. This

property will be utilized while modeling 'Diachronic model' with prior, discussed in section 3.2.

Figure 3.2 provides three-dimensional plots of Dirichlet distribution with α's at different settings. The α settings will be called as hyper-parameters in the later sections, The plots provided in this figure will help in understanding the Gibbs sampling algorithm (section 3.1.4) for a diachronic model (introduced later in section 3.2) better.

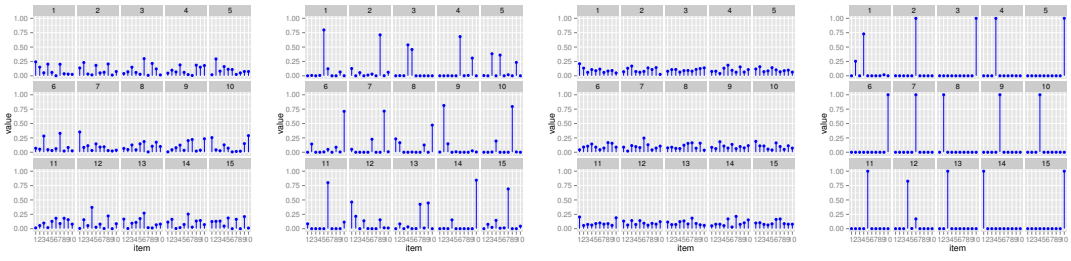


Figure 3.1 – 2D plots of Dirichlet distribution, (from left to right), α at different settings: 1. $\alpha = 1$, 2. $\alpha = 0.1$, 3. $\alpha = 5$, 4. $\alpha = 0.01$. The figure provides 15 random draws at each of four different α settings. The graph has the number of topics plotted on x-axis and the probability of the topic on the y-axis.

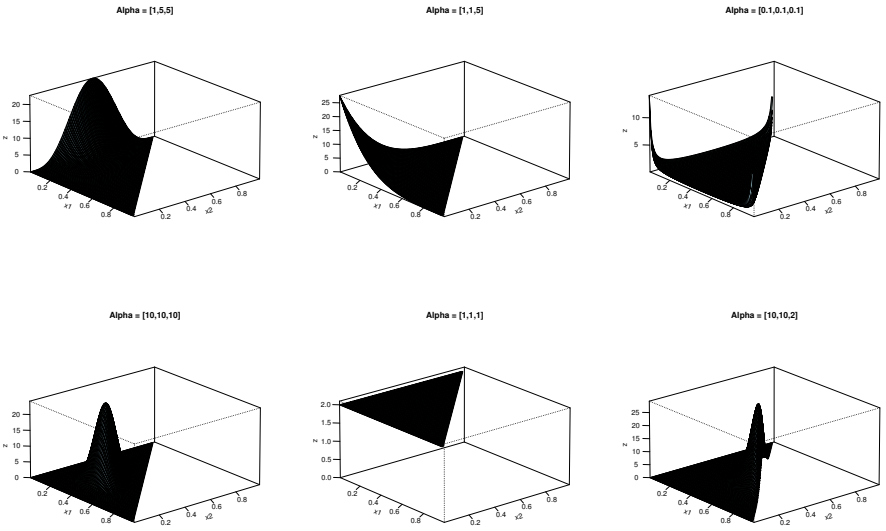


Figure 3.2 – 3D plots of Dirichlet distribution. Each of these plots are produced at dimension $k = 3$ and the respective alpha settings are titled for each plot. It can observed that when $\alpha = 1$ (symmetric), the distribution is uniform, while $\alpha = 10$ (symmetric), the distribution has a mode at the maximum and the hump is dense. But when we have α's set asymmetric, the distributions are skewed.

3.1.3 Expectation Maximization outline

Expectation Maximization (EM) introduced by Dempster et al. [1977] is a point estimation technique with iterative procedure used to find the maximum likelihood estimate (MLE) [Fisher, 1959, Sprott, 2000] of some parameters belonging to a parameter distribution Θ from incomplete data. In MLE, the idea is to estimate the model parameter(s) for which the observed data will have the maximum likelihood.

Consider a large dataset $\boldsymbol{D} = \{d^1, d^2, \ldots, d^n\}$ of size n. For each d^i in $\boldsymbol{D}$ is defined by settings of unobserved variable S that takes one of K values and some observed variables $\boldsymbol{x}$. Given a joint probability model $P(S, \boldsymbol{x}; \Theta)$ with Θ providing a representation for all the parameters, EM makes sequence of estimates $\Theta^0 \rightarrow \Theta^1 \rightarrow \ldots \rightarrow \Theta^n$. The unknown or hidden variable S enters the likelihood via the sum $\sum_s P(S, \boldsymbol{x}; \Theta)$. Taking the product of this across all data items leads to a problem which cannot just be solved by calculus arriving at a closed form solution. The EM algorithm seeks to find the MLE by iteratively applying the following two steps:

> **(E-step)** For each training instance d^i in $\boldsymbol{D}$ consider all its possible completions with setting S, $(S = k, \boldsymbol{x}^d)$, computing for each its conditional probability $P(S = k|\boldsymbol{x}^d; \Theta^n)$ under the current estimates Θ^n. Let $\gamma^d(k)$ represent this conditional probability.
>
> **(M-step)** Treating $\gamma^d(k)$ as if they were genuine counts, apply Maximum Likelihood Estimation to the virtual corpus of completions to derive new estimates for Θ^{n+1}.

The EM procedure derives useful estimates in the sense that it increases data likelihood over iterations. It theoretically guarantees that as EM iterates through, the $(n+1)^{th}$ estimate will never be less likely than the n^{th} estimate ie., $\prod_i(\sum_s P(S, \boldsymbol{x}; \Theta^n)) <= \prod_i(\sum_s P(S, \boldsymbol{x}; \Theta^{n+1}))$.

3.1.4 Gibbs sampling outline

Gibbs sampling [Bishop, 2006, Gelfand and Smith, 1990, Resnik and Hardisty, 2010] is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution (i.e. from the joint probability distribution of two or more random variables), when direct sampling is difficult. The idea behind using this algorithm is to get the desired posterior distribution after iterating through a number of sampling steps from the conditional distribution.

Consider a probability distribution $P(\boldsymbol{Z}) = P(z_1, z_2, \ldots, z_n)$, from which sampling can be done. Gibbs sampling is used to generate a sequence of samples from such a probability distribution. The Gibbs sampling procedure can work with some initial state. So we initialize state values for the variables $z_1, z_2, \ldots, z_n$. Each step of the Gibbs sampling would involve replacing value for one variable z_i with a value sampled from a distribution of the variable conditioned on the remaining variables from the distribution ie., $P(z_i|\boldsymbol{z}_{-i})$. This way, one Gibbs sample is obtained after sampling for all the variables in the distribution $P(\boldsymbol{Z})$. This procedure is defined in the following pseudo-code.

Gibbs sampling

1. Initialize $\{z_i : i = 1,\ldots,n\}$

2. for $\tau = 1,\ldots,T$:

– Sample $z_1^{\tau+1} \sim P(z_1|z_2^{(\tau)}, z_3^{(\tau)}, \ldots, z_n^{(\tau)})$

– Sample $z_2^{\tau+1} \sim P(z_2|z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots, z_n^{(\tau)})$

$\vdots$

– Sample $z_M^{\tau+1} \sim P(z_M|z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{n-1}^{(\tau+1)})$

This procedure is repeated a number of times until the samples begin to converge to what would be sampled from the true distribution. Although the number of sampling steps required to get a desired (stationary) distribution is not known, it is theoretically proved that Gibbs sampling method will reach the desired distribution after a sufficiently large number of sampling steps.

As introduced in section 3.1.1, Bayesian analysis have hyper-parameters for the prior. Continuing with the model introduced in section 3.1.3, suppose we have a model of joint probability with the hyper-parameters for all data, is $P(\boldsymbol{S}^{1:D}, \boldsymbol{x}^{1:D}, \Theta; \gamma_\Theta)$, where Θ is the model parameters and γ_Θ is hyper-parameter of the prior distribution of these parameters. Given this, there must exist a posterior $P(\boldsymbol{S}^{1:D}, \Theta|\boldsymbol{x}^{1:D}; \gamma_\Theta)$. In this, $\boldsymbol{S}^{1:D}$ and Θ play the role of $z_1, z_2, \ldots, z_n$ seeking to make samples of $\boldsymbol{S}^{1:D}$ and Θ from sampling formulas of these conditional probabilities: $P(S^d|\boldsymbol{S}^{-d}, \Theta, \boldsymbol{x}^{1:D}; \gamma_\Theta)$ and $P(\Theta^j|\boldsymbol{S}^{1:D}, \Theta^{-j}, \boldsymbol{x}^{1:D}; \gamma_\Theta)$. This way, a number of samples for each member of Θ are generated and the *mean* of these samples are considered. With enough samples, this procedure provides an accurate estimate of this mean.

3.2 Proposed Diachronic model

For this work, consider a large dataset[8] D containing a number of data items. Each data item[9] d for a target T includes a time-stamp Y. Consider the text snippets for the targets *bricked* and *smashed it* in table 3.1 for two different targets from different times. This will be used for a discussion on the models.

Now, to get a model the data is formulated to be: For the target $T =$ *bricked*, let $\boldsymbol{w}$ be a sequence of context words [10] – whose first l elements are the l words to the left of T and whose last r elements are the r words to the right of T. The time Y is the year in which the particular data item was authored. Consider S to range over K senses that T can take, but the target T in each d is assumed to have just one sense from K choices. Here, S is a hidden variable and the data will provide values only for Y and $\boldsymbol{w}$.

Based on the said formulations, and considering that the target entities *bricked* and *smashed it* can take two different senses, the data representation is provided below for the texts provided in the table 3.1 considering $r = l = 5$. In this case when the number of context words available

[8] can also be referred to as a *collection of documents* or a *corpus*.

[9] also referred to as a *document*

[10] The words to the left and right of the target T provide the context to the model

Senses	Text	Year
Sense 1	...In 1611 she was ***bricked*** into one of the rooms ...	2001
Sense 2	I've tried to flash a custom ROM and now I think I've ***bricked*** my phone	2011
Sense 1	the wind lifted his three-car garage and ***smashed it*** to the ground.	1995
Sense 2	sensational group CEO, totally ***smashed it*** in the BGT	2013

Table 3.1 – Example sentences with the target entities

to the left or right in the text snippet is less than 5, the vector has been padded with *L* to the left and *R* to the right[11].

$$Y = 2001, S = 1, \boldsymbol{w} = \{\text{L, In, 1611, she, was, into, one, of, the, rooms}\}$$
$$Y = 2011, S = 2, \boldsymbol{w} = \{\text{and, now, I, think, I've, my, phone, R, R, R}\}$$
$$Y = 1995, S = 1, \boldsymbol{w} = \{\text{lifted, his, three-car, garage, and, to, the, ground., R, R}\}$$
$$Y = 2013, S = 2, \boldsymbol{w} = \{\text{L, sensational, group, CEO,, totally, in, the, BGT, R, R}\}$$

Thus each data item can be considered to represented by the three variables Y, S and $\boldsymbol{w}$. The joint probability distribution of Y, S and $\boldsymbol{w}$ without loss of generality is given by,

$$p(Y, S, \boldsymbol{w}) = P(Y) \times P(S|Y) \times p(\boldsymbol{w}|S, Y) \tag{3.12}$$

$P(Y)$ intuitively reflects the relative abundance of data items with year Y within the entire corpus. $P(S|Y)$ directly expresses the idea that the sense varies with time (years). The final term expresses a dependency of context words on *S and Y*. Some independence assumptions are now made.

First assumption is that the context words $\boldsymbol{w}$ are conditionally independent of time Y given sense S, so $p(\boldsymbol{w}|S, Y)$ is reduced to $P(\boldsymbol{w}|S)$. This assumption reflects a reasoning that whenever a particular concept is being invoked by a word, the expected accompanying vocabulary is rather stable. In illustration, for the unambiguous word *pray* Figure 3.3 shows its corpus frequency[12] between 1840 to 2000, along with the conditional probabilities for certain neighboring words, with each trajectory of probabilities normalized by its mean. Its clear that while *pray*'s probability has changed substantially (fallen 4-fold), the conditional context probabili-

[11]This padding aspect is not particularly central but has simplifying consequence that all vectors have the same length. It is perfectly possible to not do this padding, and have vectors of varying length; the programs developed later allow for this option.

[12]This plot is based on data downloaded via Google-Ngrams API `https://github.com/econpy/google-ngrams`. This is discussed further in section 5.5.

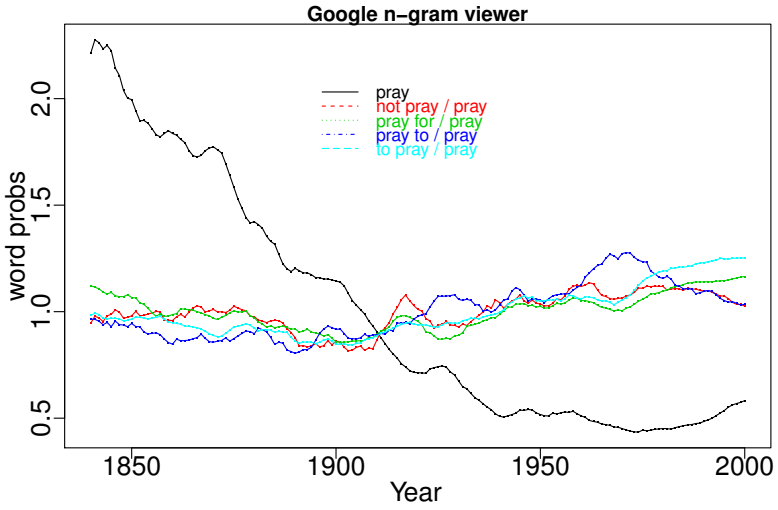


Figure 3.3

ties have changed far less. The viability of this assumption is really only for the experiments to reveal. The assumption certainly simplifies the learning problem, removing the need to learn a word probability distribution for every sense k and time t *combination*.

The second independence assumption is that the context words $\boldsymbol{w}$ given sense S are independent of each other; this assumption further reduces the third factor to $\prod_i p(w_i|S)$. The model is rewritten to be,

$$\begin{aligned} p(Y,S,\boldsymbol{w}) &= P(Y) \times P(S|Y) \times P(\boldsymbol{w}|S) \\ &= P(Y) \times P(S|Y) \times \prod_i p(w_i|S) \end{aligned} \tag{3.13}$$

Parameter notations: To fix a notation for model's parameters (i) for every time t let $\boldsymbol{\pi}_t$ be a length K vector of sense probabilities (ii) for every sense k, let $\boldsymbol{\theta}_k$ be a length V vector of context word probabilities for target sense k, where V is the size of the vocabulary encountered in all the data – and (iii) let $\boldsymbol{\tau}$ be a vector of time probabilities of length N, where N is the number of different time stamps. These parameters are effectively 2-D tables or 1-D sequences.

The model equation with the parameters for one data item is given by:

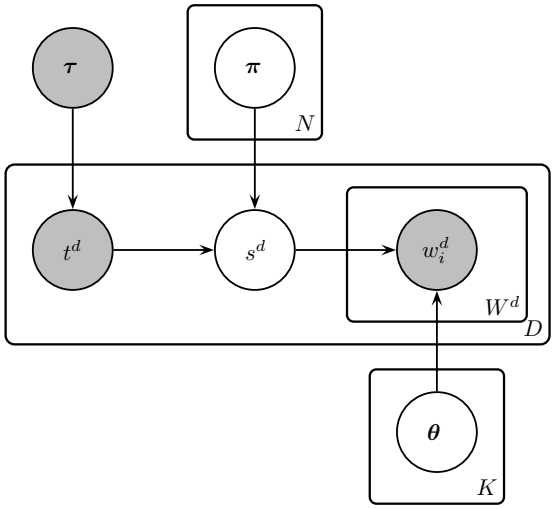
$$P(Y,S,\boldsymbol{w};\boldsymbol{\tau}_{1:N},\boldsymbol{\pi}_{1:N},\boldsymbol{\theta}_{1:K}) = P(Y;\boldsymbol{\tau}_{1:N}) \times P(S|Y;\boldsymbol{\pi}_{1:N}) \times \prod_i p(w_i|S;\boldsymbol{\theta}_{1:K}) \tag{3.14}$$

The entire data corpus is essentially a sequence of triples

$$\langle Y,S,\boldsymbol{w}\rangle^1 \ldots \langle Y,S,\boldsymbol{w}\rangle^d \ldots \langle Y,S,\boldsymbol{w}\rangle^D$$

It will be convenient to use $\boldsymbol{t}^{1:D}$,$\boldsymbol{s}^{1:D}$ and $\boldsymbol{w}^{1:D}$ to refer to all the D values of Y, S and $\boldsymbol{W}$, respectively in the D items, and to use $P(\boldsymbol{t}^{1:D}, \boldsymbol{s}^{1:D}, \boldsymbol{w}^{1:D})$ in place of $P(\langle Y, S, \boldsymbol{w}\rangle^1 \ldots \langle Y, S, \boldsymbol{w}\rangle^d \ldots \langle Y, S, \boldsymbol{w}\rangle^D)$ for the probability of the entire corpus[13]. An expression for this probability, including the parameters is:

$$P(\boldsymbol{t}^{1:D}, \boldsymbol{s}^{1:D}, \boldsymbol{w}^{1:D}; \boldsymbol{\tau}_{1:N}, \boldsymbol{\pi}_{1:N}, \boldsymbol{\theta}_{1:K}) = \prod_d \left[P(t^d; \boldsymbol{\tau}_{1:N}) P(s^d | t^d; \boldsymbol{\pi}_{1:N}) \prod_i P(w_i^d | s^d; \boldsymbol{\theta}_{1:K}) \right] \quad (3.15)$$



$\boldsymbol{\pi}_t$: parameter for senses (sense probs at t)
s^d : sense label for document d
w_i^d : word i in document d
W^d : words in document d
D : number of documents
K : total number of senses (classes)
$\boldsymbol{\theta}_k$: parameter for words (word probs at sense k)
$\boldsymbol{w}^{1:D}$: set of all documents
$\boldsymbol{\tau}$: a vector of all time periods
τ_t : probability for a particular time t
V : all words in the vocabulary
N : number of unique time instances

Figure 3.4 – Plate diagram - No prior

Figure 3.4 shows a plate diagram for the model in equation 3.15, which does not have any prior assumption over the model parameters. In the plate diagram, only the word vectors $\boldsymbol{w}$ and times (years) t are observed variables and are shown in shaded circles. Everything else is not observed and is represented as non-shaded circles. A number is drawn on the plate to represent the number of repetitions. For example, D in the corner of the plate indicates that the variables inside the plate are repeated D times, once for each document in the corpus. The directed edges between variables indicate dependencies between the variables. Consider the directed edge between t and $\boldsymbol{\tau}$. This indicates that every document's time t^d is dependent on the vector of probabilities $\boldsymbol{\tau}$ for all years, which is indicated in the first product of equation 3.15. Each document's sense s^d depends on the time t^d and sense parameter $\boldsymbol{\pi}_{t^d}$, similarly each word w_i^d in the document depends on the current document's sense assignment s^d and word parameters $\boldsymbol{\theta}_{s^d}$ – these are represented in the second and third products of the equation, which are depicted in the plate diagram using directed edges.

If the values for the parameters in equation 3.15 will be inferred from the data without any prior assumptions about the parameters, the parameter estimate woud be the Maximum Likelihood Estimate (MLE). As described in the preceding sectins, however, for Bayesian analysis,

[13]This will be particular convenient in developing the formulae in the later discussion of Gibbs sampling

a prior distribution over the parameters would be assumed.

For such a Bayesian formulation, let the $\boldsymbol{\pi}_t$ sense probability vectors have a K-dimensional Dirichlet prior with parameter γ_π and let the $\boldsymbol{\theta}_k$ word probability vectors have a V-dimensional Dirichlet prior with parameter γ_θ. So giving the model equation with the hyper-parameters for the document collection [14] included we have:

$$P(\boldsymbol{t}^{1:D}, \boldsymbol{s}^{1:D}, \boldsymbol{w}^{1:D}, \boldsymbol{\pi}_{1:N}, \boldsymbol{\theta}_{1:K}; \boldsymbol{\gamma}_\pi, \boldsymbol{\gamma}_\theta, \boldsymbol{\tau}) = \prod_t \text{Dirich}(\boldsymbol{\pi}_t; \boldsymbol{\gamma}_\pi) \times \prod_k \text{Dirich}(\boldsymbol{\theta}_k; \boldsymbol{\gamma}_\theta) \times \prod_d \left[P(t^d; \boldsymbol{\tau}_{1:N}) P(s^d | t^d; \boldsymbol{\pi}_{1:N}) \prod_i P(w_i^d | s^d; \boldsymbol{\theta}_{1:K}) \right] \tag{3.16}$$

Figure 3.5, provides a plate diagram of the model in equation 3.16 – with prior over the

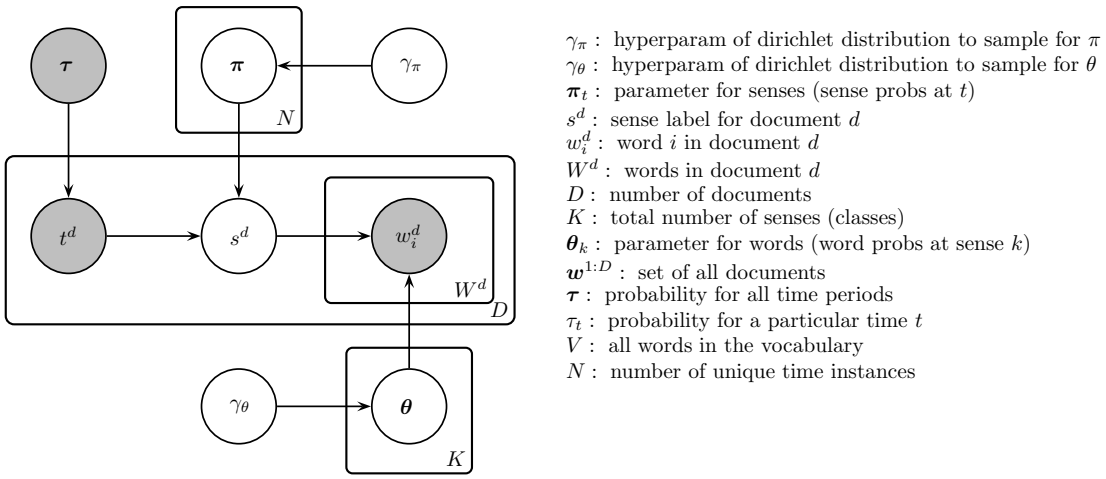


Figure 3.5 – Plate diagram - with prior over parameters

parameters[15]. The model parameters are multinomial in nature, so the prior distribution of the parameters is chosen from Dirichlet distribution as Dirichlet is the conjugate prior[16] of Multi-nomial distribution. The benefit of choosing the same family of distributions is that the posterior distribution is easy to compute. Recall the choice of Dirichlet as prior was discussed in section 3.1.2.3. The only difference in equation 3.16 from 3.15, is that the model's parameters have got a prior distribution – this is also shown in the plate diagram 3.5 with directed edges from hyper-parameters γ_π and γ_θ connected to π and θ respectively.

Equation 3.16 represents the proposed diachronic model, which is dynamic in sense of having the sense probabilities vary with the time. The following section brief discusses a static alternative.

[14] Hyper-parameter is the parameter of prior distribution; this term is used just to distinguish itself from the model parameters

[15] The priors γ_π and γ_θ are the hyper-parameters that allows to choose a prior distribution over the sense parameter $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ respectively.

[16] When a posterior distribution and prior distribution is in the same family of distributions, they are considered to be conjugate distributions

3.2.1 Static model alternative

The static model can be further derived from the diachronic model by considering a further simplifying assumption on equation 3.13. The second factor $P(S|Y)$ in equation 3.13 can be further reduced to $p(S)$ by considering that sense is independent of time. Now, the joint probability can be expressed as:

$$p(Y,S,\boldsymbol{w}) = p(Y) \times p(S) \times p(\boldsymbol{w}|S) \tag{3.17}$$

A static model presented in 3.17 can also be used in place of the diachronic model, but the model is just not dynamic as the inference procedure would involve getting the sense assignments for all the data items in the dataset and then work out the sense probabilities for each year separately. The static model can be compared to a simple Naive Bayes model assuming that the years are equally probable. Such a model can be applied by first training avoiding all time-stamps in the data, then second assigning most probable sense-labels to data items – still ignoring time-stamps, and then third by examining whether the assigned labels show any particular tendencies to be assigned more often in particular times. Further systems which work in this fashion are discussed in section 2.2.

3.2.2 Alternative choices

The context considered for the diachronic model is a bag-of-words, which closely resembles an uni-gram language model. However there are many possibilities that can be considered to be the context. One such possibility is a bi-gram model, where the context-words would be words dependent on previous word ie., $w_i|w_{i-1}$. For such a case, the model can be re-written to be,

$$P(Y,S,\boldsymbol{w};\boldsymbol{\tau},\boldsymbol{\pi},\boldsymbol{\theta}) = P(Y;\boldsymbol{\tau}_{1:N}) \times P(S|Y;\boldsymbol{\pi}_{1:N}) \times \prod_i P(w_{i-1},w_i|S;\boldsymbol{\theta}) \tag{3.18}$$

where w_{i-1},w_i term in the last factor is the context representing the current word and its previous word together. This form would increase the number of word parameters two-fold. Similarly one could extend the model to n-grams, which would further complicate the inference process. Evidently a bag of words is a simple model, but the idea is if one can infer the neologism sense using such a simple model, there is no need for such complicated model in 3.18.

Also there are other possibilities discussed in the prior art (section 2.2) which are not simple compared to the diachronic model. But the most scientific thing to do is to first consider the simplest model and go for complicated models only when the former fails.

3.3 EM estimation for Diachronic model

For the proposed diachronic model discussed in 3.2, two different parameter estimation schemes will be developed. This section develops a Expectation Maximization that derives a point ML estimate. Section 3.4 will develop an alternative Gibbs sampling approach.

The idea behind EM algorithm, as discussed in section 3.1.3, is to repeatedly calculate the expected completions of the incomplete data, and derive new parameters by maximum likelihood estimation of the expected completions (E-step and M-step). The model for a single data item provided in 3.14 to get an ML estimate of parameters considering t, s, w are the values of Y, S and W respectively, can be rewritten as:

$$P(d;\boldsymbol{\tau}_{1:N},\boldsymbol{\pi},\boldsymbol{\theta}) = P(t^d;\boldsymbol{\tau}_{1:N}) \times P(s^d|t^d;\boldsymbol{\pi}_{1:N}) \times \prod_i P(w_i^d|s^d;\boldsymbol{\theta}_{1:K}) \tag{3.19}$$

This can be expressed compactly using the parameter notations as

$$P(d;\Theta) = \tau_t \times \pi_{t,k} \times \prod_{i=1}^{W^d} \theta_{k,w_i^d} \tag{3.20}$$

where τ_t is the probability of time t in the corpus, $\pi_{t,k}$ is the probability of sense k at time t and θ_{k,w_i^d} is the probability of word w at i^{th} position in the document d with sense k and use Θ for all the model's parameters. All the notations are also summarized in the plate diagram available from figure 3.4.

Now the idea is to find the parameter setting Θ that maximizes the probability of the observed data -– a corpus of time-stamped occurrences for the given target T ie., $\arg\max_\Theta \prod_d P(d;\Theta)$. The EM steps are:

> **(E-step)** For each training instance $(Y^d, \boldsymbol{w}^d)$ from the dataset $\boldsymbol{D}$ consider all its possible completions with a setting of S, $(Y^d, S = k, \boldsymbol{w}^d)$, computing for each its conditional probability $P(S = k|Y = t, \boldsymbol{w}^d)$ under the current estimate $\Theta_n(\tau_t, \boldsymbol{\pi}_t, \boldsymbol{\theta}_k)$. Let $\gamma^d(k)$ represent this conditional probability.
>
> **(M-step)** treating the $\gamma^d(k)$ as if they were genuine counts of the alternative completions, apply maximum likelihood estimation to the virtual corpus of completions to derive new estimates $\Theta_{n+1}(\tau_t, \boldsymbol{\pi}_t, \boldsymbol{\theta}_k)$. This amounts to solving
>
> $$\arg\max_\Theta \prod_d \left[\prod_k P(Y^d, S = k, \boldsymbol{w}^d;\Theta)^{\gamma^d(k)} \right]$$
>
> or in log terms
>
> $$\arg\max_\Theta \sum_d \left[\sum_k \gamma^d(k) \times log(P(Y^d, S = k, \boldsymbol{w}^d;\Theta) \right]$$

On each iteration of the *E-step*, the conditional probabilities $P(S = k|Y = t^d, \boldsymbol{w} = \boldsymbol{w}^d)$ are

computed for all the training instances in D where $k \in \{1\ldots,K\}$. Let γ be a table of size $sizeOf(\boldsymbol{D}) \times sizeOf(k)$, where $\boldsymbol{D}$ is the number of data items in the corpus and k is the number of senses. The value for each table entry $\gamma[d][k]$ is provided by,

$$\begin{aligned}\gamma[d][k] &= \frac{P(Y=t^d, S=k, \boldsymbol{w}=\boldsymbol{w}^d)}{\sum_{S=k'} P(Y=t^d, S=k', \boldsymbol{w}=\boldsymbol{w}^d)} \\ &= \frac{\tau_t \times \pi_{t,k} \times \prod_{i=1}^{W^d} \theta_{k,w_i^d}}{\sum_{S=k'} \tau_t \times \pi_{t,k'} \times \prod_{i=1}^{W^d} \theta_{k',w_i^d}}\end{aligned}$$

(terms not dependent on k' are moved out)

$$= \frac{\tau_t \times \pi_{t,k} \times \prod_{i=1}^{W^d} \theta_{k,w_i^d}}{\tau_t \times [\sum_{S=k'_i} \pi_{t,k'} \times \prod_{i=1}^{W^d} \theta_{k',w_i^d}]}$$

(EM update after cancelling out common terms)

$$\gamma[d][k] = \frac{\pi_{t,k} \times \prod_{i=1}^{W^d} \theta_{k,w_i^d}}{\sum_{S=k'} \pi_{t,k'} \times \prod_{i=1}^{W^d} \theta_{k',w_i^d}} \tag{3.21}$$

Once table γ is filled, the M-step has to find the Θ which maximizes the probability of the virtual data corpus where each completion has a virtual frequency of $\gamma^d(k)$:

$$\arg\max_{\Theta} \sum_d \left[\sum_k \gamma^d(k) \times log\ P(Y^d, S=k, \boldsymbol{w}^d; \Theta) \right]$$

The solution to this can be derived use calculus, in conjunction with a so-called Lagrange multipliers – the derivations for the updates are worked out in section 3.3.1. The solutions work out to be as follows:

For each year t, $\boldsymbol{\pi}_t$, the parameter defining $P(S=k|Y=t;\boldsymbol{\pi}_t)$ is (for each sense k),

$$\boldsymbol{\pi}_{t,k} = \frac{\sum_d (\text{if } Y^d = t \text{ then } \gamma[d][k] \text{ else } 0)}{\sum_d (\text{if } Y^d = t \text{ then } 1 \text{ else } 0)} \tag{3.22}$$

For each sense k, $\boldsymbol{\theta}_k$, the parameter defining $P(w|S=k;\boldsymbol{\theta}_k)$ is (for each word w)

$$\boldsymbol{\theta}_{k,w} = \frac{\sum_d (\gamma[d][k] \times freq(w \in \boldsymbol{w}^d))}{\sum_d (length(\boldsymbol{w}^d))} \tag{3.23}$$

For each year t, τ_t, the parameter defining $P(t^d;\tau_t)$ is given by,

$$\tau_t = \frac{\sum_d (\text{if } t^d = t \text{ then } 1)}{\sum_d (1)} \tag{3.24}$$

Though this is included as an update for completeness, the settings for parameter τ will not

vary through iterations.

The above updates can be re-expressed further. If $\mathcal{S}_{t,k}$ is sum of the pseudo-counts of sense k for documents with time-stamp t. (ie. $\sum_d$(if $Y^d = t$ then $\gamma[d][k]$ else 0), we can also write:

$$\boldsymbol{\pi}_{t,k} = \frac{\mathcal{S}_{t,k}}{\sum_{k'} \mathcal{S}_{t,k'}} \tag{3.25}$$

Also if $\mathcal{V}_{k,v}$ is sum of the pseudo-counts of word v in sense k cases (ie. $\sum_d[\gamma^d(k) \times freq(w \in \boldsymbol{w}^d)]$), we can also write

$$\boldsymbol{\theta}_{k,w} = \frac{\mathcal{V}_{k,w}}{\sum_{w'}(\mathcal{V}_{k,w'})} \tag{3.26}$$

The procedure is summarized in algorithm 1.

as in text, assume data as
context words $\boldsymbol{w}^{1:D}$, time-stamps $\boldsymbol{t}^{1:D}$
assume K is a supplied no of senses
create $\gamma[D][K]$ `// data sense probs`
create $\mathcal{S}[T][K]$ `// see text`
create $\mathcal{V}[K][V]$ `// see text`
for *itr:=1 to no-iterations* **do**
  set $\mathcal{S}[t][k] = \mathcal{V}[k][v] = 0$ for all t,k,v
  `//` **E-step** `starts here`
  for *d:=1 to D* **do**
    for *k:=1 to K* **do**
      compute $\lambda_d[k]$ as in equation 3.21
    end
    for *k:=1 to K* **do**
      $\mathcal{S}[t^d][k]$ += $\lambda_d[k]$ `// incr count`
    end
    for *i:=1 to len*$(\boldsymbol{w}^d)$ **do**
      $\mathcal{V}[k][w_i^d]$ += $\lambda_d[k]$ `// incr count`
    end
  end
  `//` **M-step** `starts here`
  for *t:=1 to N* **do**
    compute $\boldsymbol{\pi}_t$ according to Eqn 3.22
  end
  for *k:=1 to K* **do**
    compute $\boldsymbol{\theta}_k$ according to Eqn 3.23
  end
end

Algorithm 1: EM estimation

3.3.1 Deriving EM updates

Let Θ^m be some setting of all parameters of the model and let d be a data item with unspecified S variable. The following function $Q^d(\Theta, \Theta^m)$ is formulated to be,

$$Q^d(\Theta,\Theta^m) = \mathop{\mathbb{E}}_{S|Y^d\ \boldsymbol{W}^d;\Theta^m}\Big[log\ P(S,Y^d,\boldsymbol{W}^d;\Theta)\Big] \tag{3.27}$$

$$= \sum_S \Big[P(S|Y^d,\boldsymbol{W}^d;\Theta^m) log\ P(S,Y^d,\boldsymbol{W}^d;\Theta)\Big] \tag{3.28}$$

which for the data item d, gives an *expectation* of the log probabilities of its completions at parameter setting Θ, with the expectations taken with respect to $P(S|Y^d,\boldsymbol{W}^d;\Theta^m)$, the conditional probability at parameter setting Θ^m of the completion given what is known.

This can be summed over all d, and it can be shown that the updates given in (3.22) (3.23) and (3.24) make Θ^{m+1} that value of Θ which, amongst all possible values satisfying the constraints on Θ, is the *maximizing* value of $\sum_d Q^d(\Theta,\Theta^m)$, that is

$$\Theta^{m+1} = \arg\max_\Theta \sum_d Q^d(\Theta,\Theta^m) \tag{3.29}$$

$$= \arg\max_\Theta \sum_d \left[\mathop{\mathbb{E}}_{S|Y^d\ \boldsymbol{W}^d;\Theta^m}\Big[log(P(S\ Y^d\ \boldsymbol{W}^d;\Theta))\Big]\right] \tag{3.30}$$

$$= \arg\max_\Theta \sum_d \sum_S \Big[P(S|Y^d\ \boldsymbol{W}^d;\Theta^m) log(P(S\ Y^d\ \boldsymbol{W}^d;\Theta))\Big] \tag{3.31}$$

The components of Θ are all effectively 2-D or 1-D tables, defining probability distributions. To formulate the appropriate Lagrangian it is useful to have names for each position in these tables and sequences. For this formulation, the following notations $\pi_{k|y}$ for $\boldsymbol{\pi}_{y,k}$, $\theta_{w|k}$ for $\boldsymbol{\theta}_{k,w}$ and τ_y for every τ_y will be used. In considering the space all possibly parameter settings, there are the constraints that for each y, $\sum_k \pi_{k|y} = 1$, that for each k, $\sum_w \theta_{w|k} = 1$ and that $\sum_y \tau_y = 1$.

To solve the constrained maximization problem (3.31) the Lagrangian is formulated to be

$$\mathcal{L} = \sum_d Q^d(\Theta,\Theta^m) + \sum_y \lambda_{S|y}(\sum_k \pi_{k|y} - 1) + \sum_k \lambda_{W|k}(\sum_w \theta_{w|k} - 1) + \lambda_Y(\sum_y \tau_y - 1)$$

where we have Lagrange multipliers $\lambda_{S|y}$ for each y, $\lambda_{W|k}$ for each s, and also on further multiplier, λ_Y. Besides these, $\mathcal{L}$ is over the variables $\pi_{k|y}$ (one for each k,y), $\theta_{w|k}$ (one for each w,k) and τ_y (one for each y). The values of the $\pi_{k|y}$, $\theta_{w|k}$ and τ_y variables that are stationary points of $\mathcal{L}$ are stationary points of $\sum_d Q^d(\Theta,\Theta^m)$ that satisfy the constraints. To find the stationary points of $\mathcal{L}$, we must find all partial derivatives of $\mathcal{L}$ and set to 0.

Using the earlier defined notation γ_m^d for $P(S|Y^d\ \boldsymbol{W}^d;\Theta^m)$, the multiplicative definition of the model and that these turn into additions under taking logs, $Q^d(\Theta,\Theta^m)$ can be expressed

$$Q^d(\Theta,\Theta^m) = \sum_k \gamma_m^d \left[log(\pi_{k|y}) + (\sum_{w\in V} \#(w,\boldsymbol{W}^d) log(\theta_{w|k})) + log(\tau_y)\right]$$

For each $\pi_{k|y}$, collecting the relevant terms of $\mathcal{L}$ we have $(\sum_{d:Y^d=y} \gamma_m^d(k) \left[log(\pi_{k|y})\right]) + \lambda_{S|y}\pi_{k|y}$, so for the partial derivative:

$$\frac{\partial \mathcal{L}}{\partial \pi_{k|y}} = \left[\sum_{d:Y^d=y} \gamma_m^d(k)[\frac{1}{\pi_{k|y}}]\right] + \lambda_{S|y} = 0 \implies \pi_{k|y} = \frac{\sum_{d:Y^d=y} \gamma_m^d(k)}{-\lambda_{S|y}} \tag{3.32}$$

Setting $\frac{\partial \mathcal{L}}{\partial \lambda_{S|y}}$ to 0 enforces the constraint that $\sum_{k'} \theta_{k'|y} = 1$, hence

$$\frac{\left[\sum_{d:Y^d=y}[\gamma_m^d(1) + \ldots + \gamma_m^d(K)]\right]}{-\lambda_{S|y}} = 1 \implies -\lambda_{S|y} = \sum_{d:Y^d=y}(1) \tag{3.33}$$

Putting these two equations together gives the update equation for $\pi_{k|y}$ given in (3.22) in the definition of the algorithm.

For each $\theta_{w|k}$ collecting the relevant terms of $\mathcal{L}$ we have $(\sum_d \gamma_m^d(k) \left[\#(w, \boldsymbol{W}^d) log(\theta_{w|k}))\right]) + \lambda_{W|k}\theta_{w|k}$ so for the partial derivative:

$$\frac{\partial \mathcal{L}}{\partial \theta_{w|k}} = \left[\sum_d \gamma_m^d(k)[\frac{\#(w, \boldsymbol{W}^d)}{\theta_{w|k}}]\right] + \lambda_{W|k} = 0 \implies \theta_{w|k} = \frac{\sum_d \gamma_m^d(k)\#(w, \boldsymbol{W}^d)}{-\lambda_{W|k}} \tag{3.34}$$

Setting $\frac{\partial \mathcal{L}}{\partial \lambda_{W|k}}$ to 0 enforces the constraint that $\sum_{w'} \theta_{w'|s} = 1$, and this implies

$$\begin{aligned}
-\lambda_{W|k} &= \sum_{w'}\sum_d \gamma_m^d(k)\#(w', \boldsymbol{W}^d) && (3.35)\\
&= \sum_d\sum_{w'} \gamma_m^d(k)\#(w', \boldsymbol{W}^d) && (3.36)\\
&= \sum_d \gamma_m^d(k)\sum_{w'}\#(w', \boldsymbol{W}^d) && (3.37)\\
&= \sum_d \gamma_m^d(k) \times length(\boldsymbol{W}^d) && (3.38)
\end{aligned}$$

Putting this expression for $\lambda_{W|k}$ into (3.34) gives the update formula for $\theta_{w|k}$ that was given in (3.23) in the statement of the algorithm.

Finally for each τ_y collecting the relevant terms of $\mathcal{L}$ we have $(\sum_{d:Y^d=y}\sum_s \gamma_m^d(k) \left[log(\tau_y)\right]) + \lambda_Y\tau_y$ and so for the partial derivative:

$$\frac{\partial \mathcal{L}}{\partial \tau_y} = \left[\sum_{d|Y^d=y}\sum_s \gamma_m^d(k)[\frac{1}{\tau_y}]\right] + \lambda_Y = 0 \implies \tau_y = \frac{\sum_{d|Y^d=y}(1)}{-\lambda_Y} \tag{3.39}$$

Setting $\frac{\partial \mathcal{L}}{\partial \lambda_Y}$ to 0 enforces the constraint that $\sum_{y'} \theta_{y'} = 1$, hence

$$\frac{\sum_{d|Y^d=1}(1) + \ldots \sum_{d|Y^d=|Y|}(1)}{-\lambda_Y} = 1 \implies -\lambda_Y = \sum_d(1) \tag{3.40}$$

Putting this value for λ_Y into (3.39) gives the update equation for τ_y that was given in (3.24) in the definition of the algorithm.

3.3.2 Proof of Monotone increase in likelihood

Let $l(\mathbf{d};\Theta)$ be the log probability of all the data at parameter setting Θ, ie.

$$l(\mathbf{d};\Theta) = \sum_d \left[log\left(\sum_k P(S=k, Y^d, \boldsymbol{W}^d;\Theta) \right) \right]$$

We need to show that EM iterations increase this, that is

$$l(\mathbf{d};\Theta^m) \leq l(\mathbf{d};\Theta^{m+1})$$

It can be shown that (i) $l(\mathbf{d};\Theta)$ has a *lower bound* in terms of $\sum_d Q^d(\Theta,\Theta^m)$ and a constant not involving Θ, ie. $l(\mathbf{d};\Theta) \geq \sum_d Q^d(\Theta,\Theta^m) + h(\Theta^m)$ and that (ii) this lower bound is actually *equal* to the likelihood when $\Theta = \Theta^m$. Since maximizing $\sum_d Q^d(\Theta,\Theta^m)$ is maximizing this lower bound which is tight at $\Theta = \Theta^m$, the update must increase the total log probability.

(i) **Lower bound for** $l(\mathbf{d};\Theta)$

$$\begin{aligned}
l(\mathbf{d};\Theta) &= \sum_d \left[log\left(\sum_S P(S, Y^d, \boldsymbol{W}^d;\Theta) \right) \right] && (3.41)\\
&= \sum_d \left[log\left(\sum_S P(S|Y^d, \boldsymbol{W}^d;\Theta^m) \frac{P(S, Y^d, \boldsymbol{W}^d;\Theta)}{P(S|Y^d, \boldsymbol{W}^d;\Theta^m)} \right) \right] \text{ multiply and divide} && (3.42)\\
&= \sum_d \left[log\left(\mathop{\mathbb{E}}_{S|Y^d,\boldsymbol{W}^d;\Theta^m} \left[\frac{P(S, Y^d, \boldsymbol{W}^d;\Theta)}{P(S|Y^d, \boldsymbol{W}^d;\Theta^m)} \right] \right) \right] \text{ re-expressed as an expectation} && (3.43)\\
&\geq \sum_d \left[\mathop{\mathbb{E}}_{S|Y^d,\boldsymbol{W}^d;\Theta^m} \left[log\left(\frac{P(S, Y^d, \boldsymbol{W}^d;\Theta)}{P(S|Y^d, \boldsymbol{W}^d;\Theta^m)} \right) \right] \right] \text{ Jensen's inequality} && (3.44)\\
&= \sum_d \left[\mathop{\mathbb{E}}_{S|Y^d,\boldsymbol{W}^d;\Theta^m} \left[log(P(S, Y^d, \boldsymbol{W}^d;\Theta)) \right] + \mathop{\mathbb{E}}_{S|Y^d,\boldsymbol{W}^d;\Theta^m} \left[-log(P(S|Y^d, \boldsymbol{W}^d;\Theta^m)) \right] \right] && (3.45)\\
&= \sum_d \left[Q^d(\Theta,\Theta^m) \right] + h(\Theta^m) && (3.46)
\end{aligned}$$

(ii) **this lower bound is actually equal to the likelihood when** $\Theta = \Theta^m$

this means

$$l(\mathbf{d};\Theta^m) = \sum_d \left[Q^d(\Theta^m,\Theta^m) + h(\Theta^m) \right]$$

Backing up to the line where Jensen's inequality is used

$$\sum_d \left[\mathop{\mathbb{E}}_{S|Y^d,\boldsymbol{W}^d;\Theta^m} \left[log(\frac{P(S,Y^d,\boldsymbol{W}^d;\Theta)}{P(S|Y^d,\boldsymbol{W}^d;\Theta^m)}) \right] \right]$$

If Θ is set to Θ^m, looking just at the term whose expectation is being taken:

$$\begin{aligned} log(\frac{P(S,Y^d,\boldsymbol{W}^d;\Theta^m)}{P(S|Y^d,\boldsymbol{W}^d;\Theta^m)}) &= log(\frac{P(S,Y^d,\boldsymbol{W}^d;\Theta^m)}{P(S,Y^d,\boldsymbol{W}^d;\Theta^m)/P(Y^d,\boldsymbol{W}^d;\Theta^m)}) \\ &= log(P(Y^d,\boldsymbol{W}^d;\Theta^m)) \end{aligned}$$

As this term does not contain S it can be taken out the summation:

$$\begin{aligned} \mathop{\mathbb{E}}_{S|Y^d,\boldsymbol{W}^d;\Theta^m} \left[log(\frac{P(S,Y^d,\boldsymbol{W}^d;\Theta)}{P(S|Y^d,\boldsymbol{W}^d;\Theta^m)}) \right] &= \sum_S P(S|Y^d,\boldsymbol{W}^d;\Theta^m) log(P(Y^d,\boldsymbol{W}^d;\Theta^m)) \\ &= log(P(Y^d,\boldsymbol{W}^d;\Theta^m)) \sum_S P(S|Y^d,\boldsymbol{W}^d;\Theta^m) \\ &= log(P(Y^d,\boldsymbol{W}^d;\Theta^m) \end{aligned}$$

and this means that the expression giving the lower-bound for $l(\mathbf{d};\Theta)$ is actually tight for $\Theta = \Theta^m$.

3.3.3 MAP estimation

The parameter updates that were just derived are for the ML estimation. However EM can easily be extended to produce a Maximum A Posteriori (MAP) estimate taking account of a prior over the parameters. Whereas previously the M-step considered just the probability of the virtual data corpus, the MAP version considers this in combination with the prior on the parameters, solving the maximization:

$$\begin{aligned} \mathop{\arg\max}_{\boldsymbol{\pi}_{1:N},\boldsymbol{\theta}_{1:K}} \left(\prod_d \left[\prod_k P(Y^d, S=k, \boldsymbol{w}^d; \boldsymbol{\tau}_{1:N}, \boldsymbol{\pi}_{1:N}, \boldsymbol{\theta}_{1:K})^{\gamma^d(k)} \right] \right) &\times \prod_t \text{Dirich}(\boldsymbol{\pi}_t; \boldsymbol{\gamma}_\pi) \\ &\times \prod_k \text{Dirich}(\boldsymbol{\theta}_k; \boldsymbol{\gamma}_\theta) \end{aligned}$$

or in log terms

$$\begin{aligned} \mathop{\arg\max}_{\boldsymbol{\pi}_{1:N},\boldsymbol{\theta}_{1:K}} \left(\sum_d \left[\sum_k \gamma^d(k) \times log(P(Y^d, S=k, \boldsymbol{w}^d; \boldsymbol{\tau}_{1:N}, \boldsymbol{\pi}_{1:N}, \boldsymbol{\theta}_{1:K} \right] \right) &+ \sum_t log\left[\text{Dirich}(\boldsymbol{\pi}_t; \boldsymbol{\gamma}_\pi)\right] \\ &+ \sum_k log\left[\text{Dirich}(\boldsymbol{\theta}_k; \boldsymbol{\gamma}_\theta)\right] \end{aligned}$$

This then works out to give these updates. For each year t and for each sense k

$$\pi_{t,k} = \frac{\sum_d(\text{if } Y^d = t \text{ then } \gamma[d][k] \text{ else } 0) + \gamma_\pi[k] - 1}{\sum_d(\text{if } Y^d = t \text{ then 1 else } 0) + \sum_k(\gamma_\pi[k] - 1)} = \frac{\mathcal{S}_{t,k} + \gamma_\pi[k] - 1}{\sum_k(\mathcal{S}_{t,k} + \gamma_\pi[k] - 1)} \tag{3.47}$$

For each sense k

$$\theta_{k,w} = \frac{\sum_d(\gamma[d][k] \times freq(w \in \boldsymbol{w}^d)) + \gamma_\theta[w] - 1}{\sum_d(length(\boldsymbol{w}^d)) + \sum_w(\gamma_\theta[w] - 1)} = \frac{\mathcal{V}_{k,w} + \gamma_\theta[w] - 1}{\sum_w(\mathcal{V}_{k,w} + \gamma_\theta[w] - 1)} \tag{3.48}$$

Referring back to section 3.1.2.2 concerning the *mode* of the Dirichlet, it is worth noting that the above equations make $\boldsymbol{\pi}_t$ the mode of $\text{Dir}(\boldsymbol{\pi}_t; \mathcal{S}_t + \boldsymbol{\gamma}_\pi)$ and $\boldsymbol{\theta}_k$ is the mode of $\text{Dir}(\boldsymbol{\theta}_k; \mathcal{V}_k + \boldsymbol{\gamma}_\theta)$. Practically, $\gamma_\theta[w] - 1$ turns out to be a smoothing parameter, – call this 'wc_min'. For implementation, 'wc_min' was used and by default this was set to zero which is equivalent to the Dirichlet prior $\gamma_\theta = 1$ (uninformative prior).

The EM procedure is implemented in c++ – for the *E-step*, equation 3.21 is used to compute the γ table for all data items and on computing this, the table of γ values are considered as virtual counts for the *M-step*, computed using equations 3.22 and 3.23.

3.3.4 Parameter initialization

Concerning initialization, for an experiment on a target T having a corpus of occurrences $corp$, we initialize $P(w|S)$ to $(1-\lambda)P_{corp} + \lambda P_{ran}$, where P_{corp} are the word probabilities in $corp$, P_{ran} is a random word distribution and λ is a mixing proportion, here set to 10^{-5}. Also initially the per-year sense distributions $P(S|Y)$ values are set to the same as each other. These start values thus are very far from representing the senses as being drastically different to each other or having any time variation at all. The outcomes of the EM experiments are discussed in chapter 6.

EM is one of the widely used parameter estimation technique for sense induction tasks. Emms [2013], Emms and Jayapal [2014, 2015] have used this technique to study how words and multi-word expressions respectively have changed over time. Additionally, Choe and Charniak [2013] and Jin et al. [2010] have used EM for the WSI tasks, where the former used WSI to compare this method with Gibbs sampling (section 3.1.4).

3.4 Gibbs sampling estimation for Diachronic model

Given the model provided in the equation 3.16, we have the following posterior over the unknowns:

$$P(\boldsymbol{s}^{1:D}, \boldsymbol{\pi}_{1:N}, \boldsymbol{\theta}_{1:K} | \boldsymbol{t}^{1:D}, \boldsymbol{w}^{1:D}, \boldsymbol{\tau}_{1:N}; \boldsymbol{\gamma}_\pi, \boldsymbol{\gamma}_\theta)$$

A Gibbs sampler is derived here, to produce repeated samples of $(\boldsymbol{s}^{1:D}, \boldsymbol{\pi}_{1:N}, \boldsymbol{\theta}_{1:K})$ from this posterior. Having generated the samples, the *mean* of sampled $\boldsymbol{\pi}_{1:N}$ and $\boldsymbol{\theta}_{1:K}$ values will be

considered, thus providing a different type of point estimate provided by the EM algorithm. As outlined in section 3.1.4, to do this sampling we need to compute a conditional probability to sample a new value for each variable in a vector conditioned on the other values of the variables.

The sampling distributions are first stated before proceeding to derive them. To give their definitions, it is first necessary to define two count vectors, $\mathbb{S}_{t,k}$ is the number of data items with time-stamp t and sampled sense k; $\mathbb{V}_{k,v}$ is the number of times word v occurs in data items with sampled sense k.

The sampling formulas work out to be:

$$P(s_d|\boldsymbol{s}_{-(d)},\boldsymbol{t}^{1:D},\boldsymbol{w}^{1:D},\boldsymbol{\tau}_{1:N},\boldsymbol{\pi}_{1:N},\boldsymbol{\theta}_{1:K};\boldsymbol{\gamma}_\pi,\boldsymbol{\gamma}_\theta) = \frac{\pi_{t,k}\prod_{i=1}^{W^d}\theta_{k,w_i^d}}{\sum_{k'}\pi_{t,k'}\prod_{i=1}^{W^d}\theta_{k',w_i^d}} \tag{3.49}$$

$$P(\boldsymbol{\pi}_t|\boldsymbol{\pi}_{-(t)},\boldsymbol{s}^{1:D},\boldsymbol{w}^{1:D},\boldsymbol{t}^{1:D},\boldsymbol{\tau}_{1:N},\boldsymbol{\theta}_{1:K};\boldsymbol{\gamma}_\pi,\boldsymbol{\gamma}_\theta) = \mathrm{Dir}(\boldsymbol{\pi}_t;\boldsymbol{\gamma}_\pi+\mathbb{S}_t) \tag{3.50}$$

$$P(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{-(k)},\boldsymbol{s}^{1:D},\boldsymbol{w}^{1:D},\boldsymbol{t}^{1:D},\boldsymbol{\tau}_{1:N},\boldsymbol{\pi}_{1:N};\boldsymbol{\gamma}_\pi,\boldsymbol{\gamma}_\theta) = \mathrm{Dir}(\boldsymbol{\theta}_k;\boldsymbol{\gamma}_\theta+\mathbb{V}_k) \tag{3.51}$$

The algorithm for generating Gibbs samples is provided in Algorithm 2

3.4.1 Deriving Gibbs sampling distributions

Sample for sense labels: To sample for a sense label for each document, compute the conditional probability – $P(s_d|\boldsymbol{s}_{-(d)},\boldsymbol{t}^{1:D},\boldsymbol{w}^{1:D},\boldsymbol{\tau}_{1:N}\boldsymbol{\pi}_{1:N},\boldsymbol{\theta}_{1:K};\boldsymbol{\gamma}_\pi,\boldsymbol{\gamma}_\theta)$, which samples for a new sense label for the current document based on the values of all other variables – $\boldsymbol{s}_{-(d)}$ is used to represent the sense values for all documents except the current document. This is given by,

$$P(s_d|\boldsymbol{s}_{-(d)},\boldsymbol{t}^{1:D},\boldsymbol{w}^{1:D},\boldsymbol{\tau}_{1:N},\boldsymbol{\pi}_{1:N},\boldsymbol{\theta}_{1:K};\boldsymbol{\gamma}_\pi,\boldsymbol{\gamma}_\theta) = \frac{P(\boldsymbol{s}^{1:D},\boldsymbol{w}^{1:D},\boldsymbol{t}^{1:D},\boldsymbol{\tau}_{1:N},\boldsymbol{\pi}_{1:N},\boldsymbol{\theta}_{1:K};\boldsymbol{\gamma}_\pi,\boldsymbol{\gamma}_\theta)}{P(\boldsymbol{s}_{-(d)},\boldsymbol{t}^{1:D},\boldsymbol{w}^{1:D},\boldsymbol{\tau}_{1:N},\boldsymbol{\pi}_{1:N},\boldsymbol{\theta}_{1:K};\boldsymbol{\gamma}_\pi,\boldsymbol{\gamma}_\theta)}$$

After canceling out terms that are not dependent on sense label, we get

$$\begin{aligned} P(s_d|\boldsymbol{s}_{-(d)},\boldsymbol{t}^{1:D},\boldsymbol{w}^{1:D},\boldsymbol{\tau}_{1:N},\boldsymbol{\pi}_{1:N},\boldsymbol{\theta}_{1:K};\boldsymbol{\gamma}_\pi,\boldsymbol{\gamma}_\theta) &= \frac{P(k|\tau_t;\boldsymbol{\pi}_t)\prod_{i=1}^{W^d}P(w_i^d|k;\boldsymbol{\theta}_{1:K})}{\sum_{k'}P(s_d|\tau_t;\boldsymbol{\pi}_t)\prod_{i=1}^{W^d}P(w_i^d|s_d;\boldsymbol{\theta}_{1:K})} \\ &= \frac{\pi_{t,k}\prod_{i=1}^{W^d}\theta_{k,w_i^d}}{\sum_{k'}\pi_{t,k'}\prod_{i=1}^{W^d}\theta_{k',w_i^d}} \end{aligned} \tag{3.52}$$

A sense label has to be sampled for each document (set of contexts) from the document set (set of contexts). The conditional probability in equation 3.52 turns to be just the same equation as in the 'E-step' provided in equation 3.21 of EM. The further differences between EM and Gibbs sampling parameter estimation schemes are discussed in section 3.4.4.

Sample for sense parameter: To sample for a sense parameter π, compute the conditional

as in text, assume data as
context words $\boldsymbol{w}^{1:D}$, time-stamps $\boldsymbol{t}^{1:D}$
assume K is a supplied no of senses
create $S[D]$ // data sense labels
create $\mathbb{S}[T][K]$ // see text
create $\mathbb{V}[K][V]$ // see text
for *itr:=1 to no-iterations* **do**
  set $\mathbb{S}[t][k] = \mathbb{V}[k][v] = 0$ for all t,k,v
  for *d:=1 to D* **do**
    for *k:=1 to K* **do**
      compute $\lambda_d[k]$ as in equation 3.52
    end
    $k \sim \text{Discrete}(\boldsymbol{\lambda}_d)$
    $S[d] = k$
    $\mathbb{S}[t^d][k]$ += 1 // incr count
    for *i:=1 to len*$(\boldsymbol{w}^d)$ **do**
      $\mathbb{V}[k][w_i^d]$ += 1 // incr count
    end
  end
  for *t:=1 to N* **do**
    $\boldsymbol{\pi}_t \sim \text{Dirichlet}(\boldsymbol{\pi}_t; \boldsymbol{\gamma}_\pi + \mathbb{S}[t])$
  end
  for *k:=1 to K* **do**
    $\boldsymbol{\theta}_k \sim \text{Dirichlet}(\boldsymbol{\theta}_k; \boldsymbol{\gamma}_\theta + \mathbb{V}[k])$
  end
  Save $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ values
end
return mean of $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ from all iterations

Algorithm 2: Gibbs sampling estimation

probability – $P(\boldsymbol{\pi}_t|\boldsymbol{\pi}_{-(t)}, \boldsymbol{s}^{1:D}, \boldsymbol{w}^{1:D}, \boldsymbol{t}^{1:D}, \boldsymbol{\tau}_{1:N}, \boldsymbol{\theta}_{1:K}; \boldsymbol{\gamma}_\pi, \boldsymbol{\gamma}_\theta)$. This is given by,

$$P(\boldsymbol{\pi}_t|\boldsymbol{\pi}_{-(t)}, \boldsymbol{s}^{1:D}, \boldsymbol{w}^{1:D}, \boldsymbol{t}^{1:D}, \boldsymbol{\tau}_{1:N}, \boldsymbol{\theta}_{1:K}; \boldsymbol{\gamma}_\pi, \boldsymbol{\gamma}_\theta) = \frac{P(\boldsymbol{\pi}_t, \boldsymbol{\pi}_{-(t)}, \boldsymbol{s}^{1:D}, \boldsymbol{w}^{1:D}, \boldsymbol{t}^{1:D}, \boldsymbol{\tau}_{1:N}, \boldsymbol{\theta}_{1:K}; \boldsymbol{\gamma}_\pi, \boldsymbol{\gamma}_\theta)}{\int_{\boldsymbol{\pi}_t} P(\boldsymbol{\pi}_t, \boldsymbol{\pi}_{-(t)}, \boldsymbol{s}^{1:D}, \boldsymbol{w}^{1:D}, \boldsymbol{t}^{1:D}, \boldsymbol{\tau}_{1:N}, \boldsymbol{\theta}_{1:K}; \boldsymbol{\gamma}_\pi, \boldsymbol{\gamma}_\theta)}$$

The numerator is ,

$$\prod_d \left[P(\tau_t; \boldsymbol{\tau}) P(s_d|\tau_t; \boldsymbol{\pi}_t) \prod_{i=1}^{W^d} P(w_i^d|s_d; \boldsymbol{\theta}_{1:K}) \right] \times \text{Dir}(\boldsymbol{\pi}_t; \boldsymbol{\gamma}_\pi) \times \prod_{-(t)} \text{Dir}(\boldsymbol{\pi}_{-(t)}; \boldsymbol{\gamma}_\pi) \times \prod_{1:K} \text{Dir}(\boldsymbol{\theta}_k; \boldsymbol{\gamma}_\theta)$$

For only some d, $t_d = t$. Also, the integral over $\boldsymbol{\pi}_t$ in the denominator refers to only two terms mentioning $\boldsymbol{\pi}_t$. Because of this, the fraction can be rewritten as:

$$\frac{\prod_{d:t_d=t} [\boldsymbol{\pi}_{t,s_d}] \times \text{Dir}(\boldsymbol{\pi}_t; \boldsymbol{\gamma}_\pi)}{\int_{\boldsymbol{\pi}_t} \left[\prod_{d:t_d=t} [\boldsymbol{\pi}_{t,s_d}] \times \text{Dir}(\boldsymbol{\pi}_t; \boldsymbol{\gamma}_\pi)\right]}$$

Re-expressing the numerator with $\mathbb{S}$ and using the definition of Dirichlet, we get

$$\prod_k \boldsymbol{\pi}_{t,k}^{\mathbb{S}_{t,k}} \times \frac{1}{\beta(\boldsymbol{\gamma}_\pi)} \prod_k \boldsymbol{\pi}_{t,k}^{\boldsymbol{\gamma}_\pi[k]-1} = \frac{1}{\beta(\boldsymbol{\gamma}_\pi)} \prod_k \boldsymbol{\pi}_{t,k}^{\mathbb{S}_{t,k}+\boldsymbol{\gamma}_\pi[k]-1}$$

Hence the fraction can be written as:

$$\frac{\prod_k \boldsymbol{\pi}_{t,k}^{\mathbb{S}_{t,k}+\boldsymbol{\gamma}_\pi[k]-1}}{\int_{\boldsymbol{\pi}_t}[\prod_k \boldsymbol{\pi}_{t,k}^{\mathbb{S}_{t,k}+\boldsymbol{\gamma}_\pi[k]-1}]} = \frac{1}{\beta(\mathbb{S}_t+\boldsymbol{\gamma}_\pi)} \prod_k \boldsymbol{\pi}_{t,k}^{\mathbb{S}_{t,k}+\boldsymbol{\gamma}_\pi[k]-1}$$

where the last step uses the fact noted in equation 3.7 concerning the normalizing constant in the Dirichlet. Hence we finally obtain

$$P(\boldsymbol{\pi}_t|\boldsymbol{\pi}_{-(t)}, \boldsymbol{s}^{1:D}, \boldsymbol{w}^{1:D}, \boldsymbol{t}^{1:D}, \boldsymbol{\theta}_{1:K}; \boldsymbol{\gamma}_\pi, \boldsymbol{\gamma}_\theta) = \text{Dir}(\boldsymbol{\pi}_t; \boldsymbol{\gamma}_\pi + \mathbb{S}_t)$$

i.e., a particular posterior Dirichlet.

Sample for word parameter: To sample for a value for the word parameter θ, compute the conditional probability – $P(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{-(k)}, \boldsymbol{s}^{1:D}, \boldsymbol{w}^{1:D}, \boldsymbol{t}^{1:D}, \boldsymbol{\tau}_{1:N}, \boldsymbol{\pi}_{1:N})$ is given by

$$P(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{-(k)}, \boldsymbol{s}^{1:D}, \boldsymbol{w}^{1:D}, \boldsymbol{t}^{1:D}, \boldsymbol{\tau}_{1:N}, \boldsymbol{\pi}_{1:N}; \boldsymbol{\gamma}_\pi, \boldsymbol{\gamma}_\theta) = \frac{P(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{-(k)}, \boldsymbol{s}^{1:D}, \boldsymbol{w}^{1:D}, \boldsymbol{t}^{1:D}, \boldsymbol{\tau}_{1:N}, \boldsymbol{\pi}_{1:N}; \boldsymbol{\gamma}_\pi, \boldsymbol{\gamma}_\theta)}{\int_k P(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{-(k)}, \boldsymbol{s}^{1:D}, \boldsymbol{w}^{1:D}, \boldsymbol{t}^{1:D}, \boldsymbol{\tau}_{1:N}, \boldsymbol{\pi}_{1:N}; \boldsymbol{\gamma}_\pi, \boldsymbol{\gamma}_\theta)}$$

The numerator is

$$\prod_d \left[P(\tau_t; \boldsymbol{\tau}) P(s_d|\tau_t; \boldsymbol{\pi}_t) \prod_{i=1}^{W^d} P(w_i^d|s_d; \boldsymbol{\theta}_{1:K}) \right] \times \prod_{1:N} \text{Dir}(\boldsymbol{\pi}_t; \boldsymbol{\gamma}_\pi) \times \text{Dir}(\boldsymbol{\theta}_k; \boldsymbol{\gamma}_\theta) \times \prod_{-(k)} \text{Dir}(\boldsymbol{\theta}_{-(k)}; \boldsymbol{\gamma}_\theta)$$

For only some d, $s_d = k$. Also, the integral over $\boldsymbol{\theta}_k$ in the denominator refers to only two terms mentioning $\boldsymbol{\theta}_k$. Because of this, the fraction can be written to be

$$\frac{\prod_{d:s_d=k} \left[\prod_{i=1}^{W^d} \theta_{k,w_i^d} \right] \times \text{Dir}(\boldsymbol{\theta}_k; \boldsymbol{\gamma}_\theta)}{\int_{\boldsymbol{\theta}_k} \left\{ \prod_{d:s_d=k} \left[\prod_{i=1}^{W^d} \theta_{k,w_i^d} \right] \times \text{Dir}(\boldsymbol{\theta}_k; \boldsymbol{\gamma}_\theta) \right\}}$$

Re-expressing the numerator with $\mathbb{V}$ and using the definition of Dirichlet we get,

$$\prod_v \boldsymbol{\theta}_{k,v}^{\mathbb{V}_{k,v}} \times \frac{1}{\beta(\boldsymbol{\gamma}_\theta)} \prod_v \boldsymbol{\theta}_{k,v}^{\boldsymbol{\gamma}_\theta[v]-1} = \frac{1}{\beta(\boldsymbol{\gamma}_\theta)} \prod_v \boldsymbol{\theta}_{k,v}^{\mathbb{V}_{k,v}+\boldsymbol{\gamma}_\theta[v]-1}$$

Hence the fraction can be rewritten to be

$$\frac{\prod_v \boldsymbol{\theta}_{k,v}^{\mathbb{V}_{k,v}+\boldsymbol{\gamma}_\theta[v]-1}}{\int_{\boldsymbol{\theta}_k} \prod_v \boldsymbol{\theta}_{k,v}^{\mathbb{V}_{k,v}+\boldsymbol{\gamma}_\theta[v]-1}} = \frac{1}{\beta(\mathbb{V}_k+\boldsymbol{\gamma}_\theta)} \prod_v \boldsymbol{\theta}_{k,v}^{\mathbb{V}_{k,v}+\boldsymbol{\gamma}_\theta[v]-1}$$

where the last step uses the fact noted in equation 3.7 concerning the normalizing constant in

the Dirichlet. Hence we finally obtain,

$$P(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{-(k)}, \boldsymbol{s}^{1:D}, \boldsymbol{w}^{1:D}, \boldsymbol{t}^{1:D}, \boldsymbol{\tau}_{1:N}, \boldsymbol{\pi}_{1:N}; \boldsymbol{\gamma}_\pi, \boldsymbol{\gamma}_\theta) = \mathrm{Dir}(\boldsymbol{\theta}_k; \boldsymbol{\gamma}_\theta + \mathbb{V}_k)$$

i.e., a particular posterior Dirichlet.

The update equations derived for building a Gibbs sampler are used in the algorithm provided in Algorithm 2 and the procedure is further used for c++ implementation. A number of Gibbs samples are collected executing this algorithm for a number of times to get the actual posterior of the parameter distribution. There is no recommendation on the number of samples to be obtained to get the desired posterior – in other words the desired posterior can be considered to be a stable distribution. For experiment purposes, $10,000$ samples were collected and the first $1,000$ were considered to be 'burn-in' iteration.

3.4.2 Why 'burn-in'?

The *burn-in* tries to address: say we start at a random point, say at x and the Gibbs chain is executed for n iterations, from which first b iterations are ignored considering them to be non-stationary. The b iterations is the burn-in period. After the distributions obtained after b is considered to be a stationary distribution and can be used for further computation.

3.4.3 Parameter initialization

The algorithm is not dependent on the initial values for the parameters, so they can be assigned in a random fashion. Concerning the initialization two possibilities are made available through command-line parameters; the first option is to initialize the parameters ($\boldsymbol{\pi}$ and $\boldsymbol{\theta}$) by sampling from Dirichlet distribution based on the hyper-parameters ($\boldsymbol{\gamma}_\pi$ and $\boldsymbol{\gamma}_\theta$), obtained from user-inputs; and the second option is to initialize $\boldsymbol{\theta}$ using the corpus word probabilities P_{corp} with a random word distribution P_{ran} (ie., $(1-\lambda)P_{corp} + \lambda P_{ran}$, where λ is a mixing proportion, here set to 10^{-5}), and initialize $\boldsymbol{\pi}$ with per-year sense distributions. Further it is to be noted here that symmetric priors are assigned for the hyper-parameters.

3.4.4 How is a Gibbs sampler different from the EM?

The aim of the Gibbs sampling algorithm here is to get a 'mean' of the parameter estimate, while the EM algorithm in this work is used to get the MLE and MAP parameter estimates. The EM algorithm provides a point estimate, while Gibbs sampling algorithm produces a number of posterior samples which can be used to make further inferences.

For EM, in the 'E-step' we compute the conditional probability (in equation 3.21) and treat these values to be virtual counts; while in Gibbs sampling estimation the conditional probability in equation 3.52 turns to be just the same equation as in the 'E-step' of EM, but instead of considering them to be virtual counts, a sense label is sampled from the conditional probability distribution. Also, the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ are sampled from Dirichlet distribution

based on the sampled sense labels assigned to each document; while in EM, an MLE estimate is made in the 'M-step' as an equivalent to the Dirichlet samples generated for the parameters.

3.4.5 Label switching

Label switching is a phenomenon identified in Monte Carlo Markov Chain (MCMC) sampling techniques on mixture models that does not happen always. Gibbs sampling, being an MCMC sampling technique it is expected to identify this phenomenon during experiments. In this section, we discuss this problem briefly.

To explain this phenomenon, consider a dataset with occurrences of the word *play* having two different senses. *Sense 1* may represent a sentence or document with the word *play* being used in the context of a game, while *Sense 2* may represent a sentence or document with the word *play* being used in the context of music. A sense discrimination model using Gibbs sampling may not distinguish all the game related documents as *Sense 1* and all the music related documents as *Sense 2*. For some series of samples, the model might label music related documents as *Sense 1*. This is called the problem of label switching. Therefore, it would be unwise to consider the mean or mode of the labels from multiple samples.

According to Stephens [2000], the label switching problem arises when a symmetric prior is used in Bayesian approaches to parameter estimation. In our Gibbs parameter estimation scheme, symmetric priors (allowing the prior distributions $\boldsymbol{\gamma}_\pi$ and $\boldsymbol{\gamma}_\theta$ to have the same value) are assigned to the latent variables $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$. So label switching is expected to happen in this case as well.

For testing purposes, a mock-dataset was considered for a target T with 8 data items containing two senses 'sense 0' and 'sense 1' with just 2 vocabulary items A & B. The mock-dataset is provided in table 3.2 where each data is provided with the sense annotations.

A A A B A T A B A A A	1990 0	A A A A B T A A A A A	1990 0
A A B A A T A A A A B	1990 0	A A A A A T A B A A A	1990 0
A B A A A T B A B A A	1990 0	B B B B A T B B B A B	1990 1
A A A A B T A B A A B	1990 0	B B B B A T B B B B B	1990 1

Table 3.2 – Mock dataset used to test label switching for a target T with window 5 containing A and B as vocabulary items from a single year 1990. The $0's$ and $1's$ to the right are appended providing a sense annotation for each data item.

The dataset is effectively a sample from a diachronic model with $P(S|Y)$ parameters $\pi_{1990}[0] = 0.75$, $\pi_{1990}[1] = 0.25$ and $P(W|S)$ parameters $\theta_0[A] = 0.8$, $\theta_0[B] = 0.2$, $\theta_1[A] = 0.15$ and $\theta_1[B] = 0.85$, where $\pi_{1990}[0]$ is the year 1990 prob for 'sense 0' and $\theta_0[A]$ is the probability for vocabulary item B in 'sense 0'. The parameters values stated are the maximum likelihood values for this sample. The unsupervised Gibbs sampling estimation procedure was run on this sample for 10,000 and 50,000 iterations without the supplied sense choices. For these runs, the sense and word distributions ($\boldsymbol{\pi}_{1990}[k]$ and $\boldsymbol{\theta}_k[w]$) for 10k samples and the first 50k samples are plotted in figures 3.6 and 3.7. For all the figures reported in this section, every 5th sample was

considered for plotting.

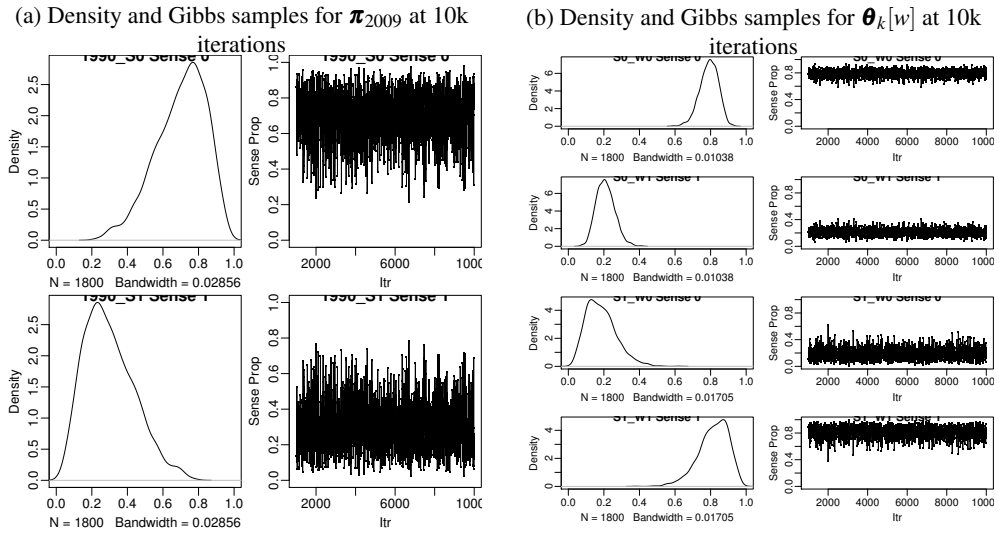


Figure 3.6 – The Gibbs samples of sense $\boldsymbol{\pi}_t[k]$ and word $\boldsymbol{\theta}_k[w]$ produced in with a 10k run are plotted in (a) and (b); the left hand plots in (a) and (b) shows the density of 10k Gibbs samples while the right hand plots show the Gibbs samples in 10k iterations.

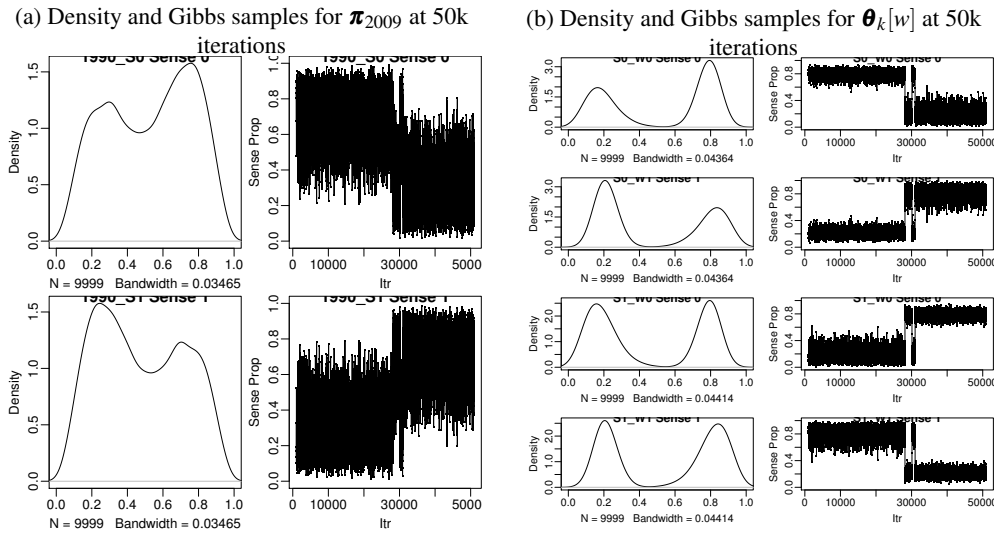


Figure 3.7 – The Gibbs samples of sense $\boldsymbol{\pi}_t[k]$ and word $\boldsymbol{\theta}_k[w]$ produced in with a 50k run are plotted in (a) and (b); the left hand plots in (a) and (b) shows the density of 50k Gibbs samples while the right hand plots show the Gibbs samples of 50k samples.

In figures 3.6(a) and (b), the plots in the first column show a density plot[17] based on $\pi_{1990}[k]$ samples obtained from 10k iterations and the second column shows a sequence of samples of $\theta_k[w]$ for vocabulary w from sense k. In (a), the plots in the first row correspond to $\pi_{1990}[0]$ Gibbs samples and the plots in the second row correspond to $\boldsymbol{\pi}_{1990}[1]$ samples, whereas in (b)

[17]The density plot was made using the 'density' function available from 'zoo' package in 'R'.

each row of plots correspond to Gibbs samples for $\theta_0[A]$, $\theta_0[B]$, $\theta_1[A]$ and $\theta_1[B]$. All density plots in this figure show a single mode, signalling no label switching. Similar to this, figures in 3.7 (a) and (b) shows plots for 50k samples. When the algorithm is run for a longer number of iterations (50k in this case), label switching happens and this can be seen from the density plots in the form of two modes in each chain of samples (see left hand plots in figures 3.7 (a) and (b)). Further label switching is also evident when the samples are plotted over iterations (see right hand plots in figures 3.7 (a) and (b)). During label switching, considering the mean of the samples will not be appropriate. The rest of this shows non-occurrence of label switching on a larger mock data set.

A A A B A T A B A A A	1990 0	A A A A B T A A A B A	2000 0
A A B A A T A A A A B	1990 0	A A A A B T A A A A A	2000 0
A B A A A T B A B A A	1990 0	A B A A B T A A A B A	2000 0
A A A A B T A B A A B	1990 0	B B B A B T B A B B B	2000 1
A A A A B T A A A A A	1990 0	B B A B B T B B B B A	2000 1
A A A A A T A B A A A	1990 0	B A B B B T A B A B B	2000 1
B B B B A T B B B A B	1990 1	B B B B A T B A B B A	2000 1
B B B B A T B B B B B	1990 1	B B B B A T B A B B B	2000 1
B A B B A T B B B A B	1990 1	B B B A B T B A B B B	2000 1
A A A B A T A B A A A	1990 0	A A A A B T A A A B A	2000 0
A A B A A T A A A A B	1990 0	A A A A B T A A A A A	2000 0
A B A A A T B A B A A	1990 0	A B A A B T A A A B A	2000 0

Table 3.3 – Mock dataset used to test label switching for a target T with window 5 containing A and B as vocabulary items from two years 1990 and 2000. The $0's$ and $1's$ to the right are appended providing a sense annotation for each data item.

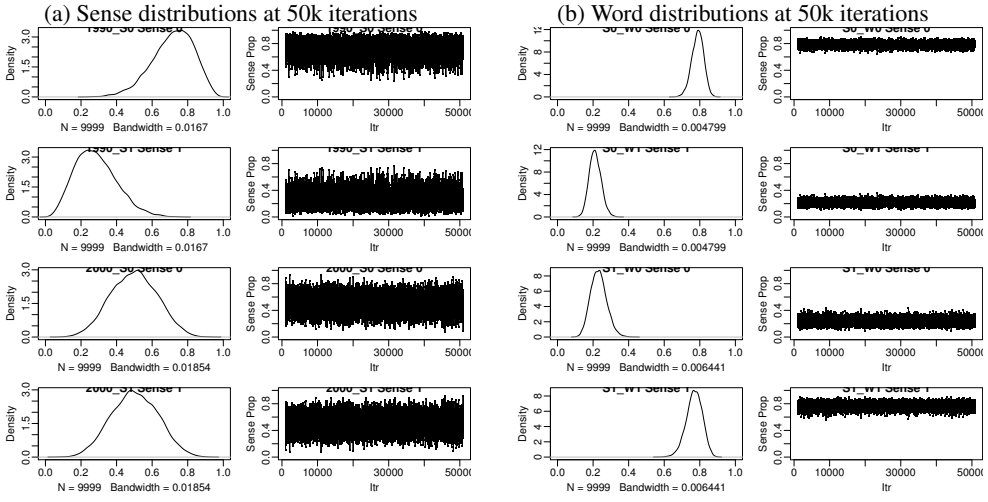


Figure 3.8 – The Gibbs samples of sense and word distributions produced on a larger dataset with 24 items are plotted in (a) and (b); the left hand plots in (a) and (b) shows the density of 50k Gibbs samples from 50k iterations and the right hand plots show the Gibbs samples of 50k samples.

Table 3.3 gives a larger mock dataset, with two year data. The Gibbs sampling algorithm for the diachronic model was also executed on this dataset for 50k iterations with uninformed

prior over the parameters. Figure 3.8(a) shows the sense distributions $\boldsymbol{\pi}_t[k]$ and (b) the word distributions $\boldsymbol{\theta}_k[w]$ for 50k Gibbs samples. In this case, no label-switching is apparent even after 50k iterations. The plots suggest that as the data size increases, there is less possibility of seeing a label switching in the earlier part of the Gibbs chain. In all experiments reported later the above types of plots are produced and can be consulted to check for evidence of label switching.

3.4.6 Credible interval

Gibbs sampling produces a number of samples from the posterior on the parameters. It is possible to consider a point estimate in the form of *mean*. A Bayesian *credible interval* (CI) is an interval in the domain of a posterior probability distribution that provides a range where a given percentage of posterior distribution lies [Box and Tiao, 1992]. For a given percentage there could be lots of such intervals. According to Box and Tiao [1992], a HPD interval has the property that a particular percentage is included and every point included in the interval has higher posterior density than every point excluded. Equivalently the interval can be thought of as what is found if a horizontal line of a density value is raised until the area under the curve between the points where it meets the density is the required size. This is shown using a sample density plot in figure 3.9 with 90% HPD interval.

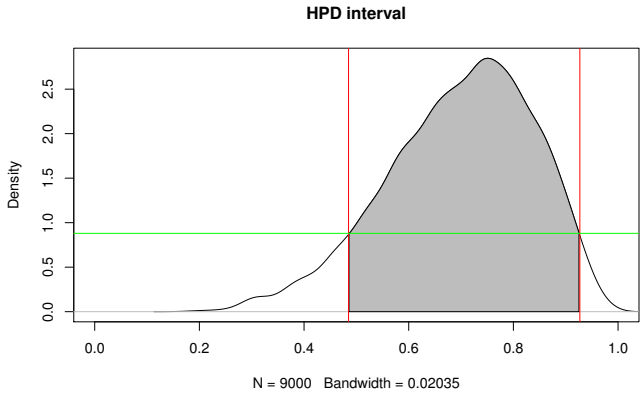


Figure 3.9 – A Density plot showing 90% HPD interval with area under the curve shaded.

There are a number of algorithms to compute HPD (discussed in CHEN and SHAO [1999]), but we used an *R* package called *LaplaceDemon*[18] to compute the HPD intervals for plotting purposes (can be seen from the various plots in chapter 6).

In this section, the Gibbs sampling updates were first derived for the diachronic model, then there was a discussion about the different parameters that will be used to produce Gibbs samples, then the label switching problem that is expected to occur with Gibbs estimation and finally about the credible intervals. The outcomes of different experiments based on the Gibbs

[18] Available from http://www.bayesian-inference.com/software – last accessed on Jan 21, 2016

sampling algorithm are discussed in chapter 6.

3.5 The model of Frermann and Lapata [2016]

Earlier in section 2.4, we mentioned of a related proposal from Frermann and Lapata [2016]. It is a generative probabilistic model, and like the model just described, given a time t, a sense k is chosen, and for each t there is a parameter $\boldsymbol{\pi}_t$ for $P(S|Y)$. Then given a sense k and a time t, the sequence of words $\boldsymbol{w}$ is chosen, and for each sense k *and* time t, they have a parameter $\boldsymbol{\theta}_{t,k}$ for $P(\boldsymbol{w}|S,Y)$ (they have different notations for the parameters, but are adapted here for the convenience of comparisons that will be made with the current proposal). Thus one key difference to the proposal made here is that they *do not* assume the conditional independence. $P(\boldsymbol{w}|S,Y) = P(\boldsymbol{w}|S)$ as we do.

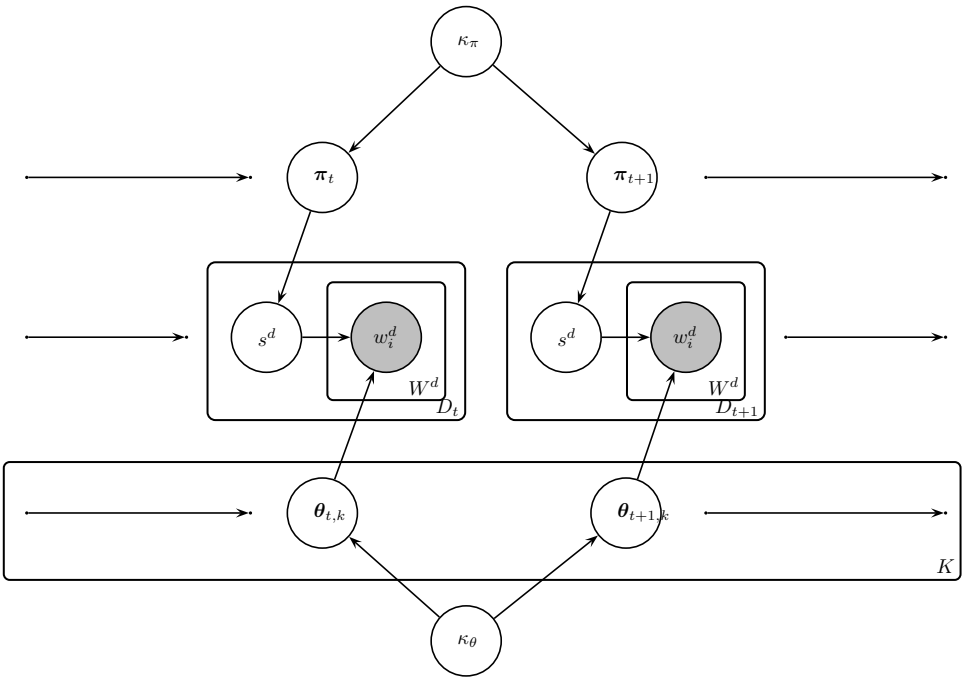


Figure 3.10 – Plate diagram for the model used in Frermann and Lapata [2016]

Their model is shown in figure 3.10 as a plate diagram[19]. The succession of the plates with repetition count D_t and D_{t+1} represent the same generative story as that used in our model 3.2. As in our model, as one proceeds through time, at each time t a particular sense parameter $\boldsymbol{\pi}_t$ is assumed. In contrast to our model, the word distribution for a sense k are not assumed independent of time, and so as one proceeds through time, at each time, t and for each sense k, there is a particular word probability parameter $\boldsymbol{\theta}_{k,t}$.

They also make different assumptions concerning priors on parameters. They do not adopt assume the sequences of $\boldsymbol{\pi}_t$ and $\boldsymbol{\theta}_{t,k}$ values are draws from Dirichlet. Instead for each they wish to have a prior which, in their words

[19]It is adapted from Figure 1 of Frermann and Lapata [2016]

> *encourages smooth change of parameters at neighboring times, in terms of a first order random walk on the line*

To do this first (see section 3.1 of their paper) they treat the $\boldsymbol{\pi}_t$ and $\boldsymbol{\theta}_{t,k}$ vectors (K and V dimensional) as images under the so-called *logistic transformation* of other underlying vectors normal. For example, this means that a K dimensional $\boldsymbol{\pi}_t$, each dimension of which is between 0 and 1, is seen as the image under the logistic transformation, LN, of a K dimensional vector $\boldsymbol{\eta}_t$, each dimension of which is between $-\infty$ and $+\infty$: $\pi_t[k] = LN(\eta_t[k]) = \frac{e^{\eta_t[k]}}{\sum_{k'} e^{\eta_t[k']}}$. These underlying vectors are then seen as draws from multivariate Gaussian. Technically this means that each $\boldsymbol{\pi}_t$ and $\boldsymbol{\theta}_{t,k}$ may be seen as drawn from a so-called *logistic normal* prior. It is via assumptions about how values of the dimensions of parameters of these Gaussian may vary over time that the modeling of smoothness comes in (see section 3.2 of Frermann and Lapata [2016]). This is done via the use of *intrinsic Gaussian Markov Random Fields* (iGMRFs) Mimno et al. [2008], Rue and Held [2005]. Without attempting to give all the details of iGMRFs, where $\Phi_1 \ldots \Phi_N$ is a sequence of length N (the number of time points), an iGMRF with precision hyper-parameter κ may be defined with the effect that when sequences are drawn from this prior, the *differences* between successive values will be distributed according to a Gaussian with mean 0 and variance κ^{-1}

$$\Phi_t - \Phi_{t-1} \sim \mathcal{N}(0, \kappa^{-1})$$

Thus a *high* κ favors *small* changes in successive values. At the same time this prior does *not* impose any overall mean on the values in the sequence.

They assume such an iGMRF prior for the means of the Gaussian underlying the $\boldsymbol{\pi}_t$ values, a prior with precision parameter κ_π. This is shown in the upper part of the plate diagram in figure 3.10[20]. Similarly they assume an iGMRF prior for the means of the Gaussian underlying the $\boldsymbol{\theta}_t$ values for each k, a prior with precision parameter κ_θ. This is shown in the lower part of the plate diagram in 3.10.

Concerning parameter estimation for such a model, they point out (see section 3.3 of Frermann and Lapata [2016]) that

> *The logistic normal prior is not conjugate to the multinomial distribution. This means that the straightforward parameter updates known for sampling standard, Dirichlet-multinomial, topic models do not apply.*

They then proceed to describe a technique due to Mimno et al. [2008] by which it is nonetheless possible to obtain a Gibbs sampling method for estimation.

In applying this model, for the succession of $\boldsymbol{\theta}_{t,k}$ values, they set κ to a high value, so that although $\boldsymbol{\theta}_{t,k}$ does not have to be constant over time, only small variation is anticipated by the prior. The succession of $\boldsymbol{\pi}_t$ values is allowed greater variation.

[20] The diagram thus omits a few details of the pathway from κ_π to the $\boldsymbol{\pi}_t$ values

It seems fair to say that the model we have proposed is a simpler one than that of Frermann and Lapata [2016], and that in many respects the model of Frermann and Lapata [2016] is a logical conceptual development of the model proposed here, though it was not developed in that way. In a certain sense, from the perspective of their model, our model is what would be arrived at by (i) letting κ for $\boldsymbol{\theta}_{tk}$ tend to ∞, preventing any change of word-given-sense probabilities in successive times and (ii) letting κ for $\boldsymbol{\pi}_t$ tend to 0, allowing arbitrary change of sense-given-time probabilities.

Chapter 4

Evaluation & analysis options for neologism

An emerging sense can possibly be inferred using the diachronic model introduced in chapter 3. But evaluating this work is challenging as there is no pre-existing sense-annotated dataset that can be used for evaluation purposes. So, in section 4.1 we consider a range of possible approaches to ground truth for sense emergence that have been used, noting their advantages and disadvantages. A new approach is also put forward, and following on from this, the specifics of how emergence times will be extracted from time-lines in this thesis. Further in section 4.2 there is a discussion on the different analysis schemes used to analyze the parameter outcomes in ascertaining the neologistic sense.

4.1 Ground truth for sense emergence

To evaluate the proposed diachronic model we need to know the date at which a neologistic sense emerged in a particular corpus – call this C_0 – this is the time at which the neologistic sense for the word departed from close to zero and continued to climb thereafter (as shown in figure 4.1).

Given a large-scale, time-stamped and sense-labelled corpus it would be easy to determine C_0, a time at which the new sense for a word first departed from zero frequency. But this kind of corpus does not exist, which has also been observed by Cook et al. [2013], Lau et al. [2012]. Therefore it is a hard problem in establishing the ground truth concerning sense emergence, against which to evaluate the outputs of any sense emergence system.

Without a labelled data-set, the simplest of all possibilities is to rely on native speaker intuition about the timing of sense emergences. This is at best applicable to recent innovations and clearly the subjectivity of such an approach is far from ideal. This approach to ground-truth was adopted by Mitra et al. [2014]. This section discusses some more objective possibilities.

Firstly in sections 4.1.1 and 4.1.2 two possible approaches that exploit dictionaries are discussed, both of which have been adopted in prior work, and their advantages and drawbacks

are noted. Then in section 4.1.3 we will propose a novel 'tracks-plot' approach to providing evidence of the location of C_0. Then in section 4.1.4 we discuss specifically how the time-lines of inferred sense-probabilities will be assessed to determine whether a neologism has been detected, and if so what its emergence time is. The same processes are involved in determining a reference time from the time-lines involved in a 'tracks-plot'.

4.1.1 Dictionary first citation

A native speaker may be confident about the recent lexical innovations in a language, but in pursuit of being objective and considering innovations that are less recent, it is natural to consider dictionaries.

Historically oriented dictionaries such as the Oxford English Dictionary strive to maintain the earliest citation date[1] of a word-sense pair. Call this D_0^c. D_0^c is not the same as C_0 (shown in figure 4.1) as the latter indicates the time at which the word's novel sense started to catch on, which may well be quite a time after its very first use. It is a very reasonable assumption, however, that D_0^c is a *lower*-bound for C_0 i.e., $D_0^c \leq C_0$. This could only fail to be so if the dictionary compilers have failed to do their work properly.

We will adopt this as one kind of test on the inferred emergence date: if a system were to give an inferred emergence date substantially *earlier* then D_0^c that would be grounds for counting the system as having made a mistake.

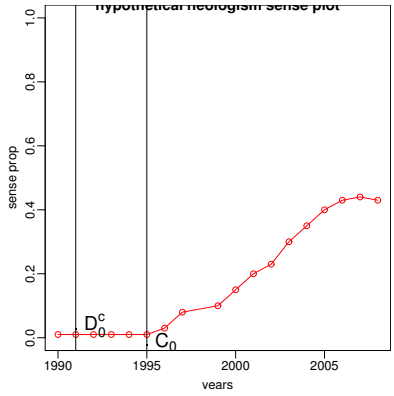


Figure 4.1 – A hypothetical sense emergence plot - Dictionary and corpus emergence dates (D_0^c & C_0) annotated

For example consider the target *mouse* in its *pointing device* sense. For this sense D_0^c, the dictionary first citation date is 1965, and the citation comes from a research paper published in 1965. The mouse computer peripheral only became popular considerably later: according to Wikipedia [2016] its use took off in the early 1980s. So it would not be surprising if C_0, the date at which this use of the term *mouse* departed and continued to climb from zero in the n-grams books based data were to be substantially later than 1965. The 'tracks-plot' technique discussed further in the next section 4.1.3 provides compelling evidence of exactly this.

[1] brief information on how OED updates its online dictionary is provided in `http://www.oxforddictionaries.com/words/questions-about-dictionaries`

4.1.2 Dictionary first inclusion

A number of dictionaries have a sequence of editions and so a particular word-sense pairing will have an earliest appearance in such a sequence. So a close inspection of a succession of dictionaries can provide a date based on first inclusion of a word-sense pair – call this D_0^i.

It seems dictionary compilers weigh up a complex combination of subtle considerations in making decisions on whether or not to include a new sense entry in a new dictionary edition [Barnhart, 2007, Sheidlower, 1995, Simpson, 2000] some of them commercial. So some caution is probably appropriate in relating D_0^i, a dictionary first inclusion date, to C_0, the 'true' time of emergence of the particular word-sense pairing. It seems likely that C_0 will be *earlier* than D_0^i, as dictionary compilers wait to see if something novel persists or fades away.

There are some disadvantages in seeking to make reference to a dictionary first inclusion date, D_0^i. One is the low time resolution, as it follows the publication cycle of the dictionaries. New editions do not appear on a yearly basis, but at longer and varying time intervals, so that 5, 10, 15 year intervals, or even longer are quite possible. A second difficulty is just the practical one of gaining access to all of the different versions of a dictionary, and the somewhat labor intensive process of consulting them all to find out whether a target word-sense pair is found in them.

Using dictionary first inclusion date, D_0^i as the main source of ground-truth about sense emergence was the approach taken by [Cook et al., 2013, 2014, Lau et al., 2012], [Rohrdantz et al., 2011] and [Tang et al., 2015], as was noted earlier in section 2.3. It will not be made the main such source in the current thesis work, with the emphasis placed instead on the approach described in section 4.1.3. Nonetheless in section 6.3 where the experimental outcomes for particular targets are discussed, some D_0^i dating information will be noted.

4.1.3 Tracks-plots

In the previous section 4.1.1, we discussed one way to test an inferred emergence date by checking that it is later than D_0^c, the dictionary first citation date. In this section a proposal is made for a method that uses a so called 'tracks-plot', to establish C_0, the actual sense emergence time in the corpus, against which an inferred corpus emergence time can be compared.

For a target word T, there are words which it is intuitive to expect in the vicinity of T in its neologistic sense of T, but not in its vicinity in its other senses. For example, consider the target *mouse* (introduced in the previous section 4.1.1). When used in its neologistic sense it is likely to co-occur with words like *click, button, pointer* and *drag*. The idea behind this 'tracks-plot' is if the per-year probabilities for such words in the dataset $P(w|Y)$ are plotted, they are expected to be close to 0 during an initial period and take off at C_0 – the corpus emergence date. For this example, the per-year probabilities $P(w|Y)$ for the words *click, button, pointer* and *drag* in the *mouse* dataset are plotted in figure 4.2. For this plot, the per-year word probability values are re-scaled to a new arbitrary range 0 and 1. Additionally, a *moving average* of these probability values are computed (3 year values are averaged) to get a *smooth* plot without jaggedness.

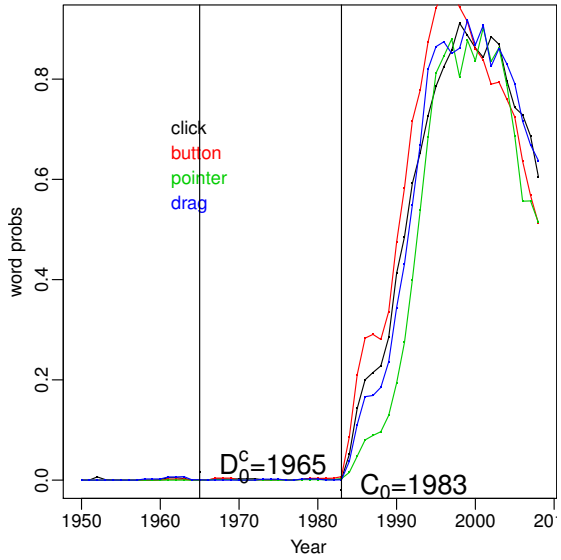


Figure 4.2 – Tracks plot for *mouse* - True C_0 annotated

In the plot seen in figure 4.2, all the words associated with neologistic sense of the target *mouse* starts at near 0 and takes off in the year 1983. This is a strong indication that C_0, the true corpus emergence date for this sense, is 1983. Also plotted for comparison is D_0^c, which is 1965 in this case, and considerably earlier than the apparent C_0.

For this technique to establish a value for C_0 for a given sense of a target word it is necessary to use intuition to establish which words might be especially prominent with the given sense. But after that the corpus data itself indicates a time at which these words have sudden increase in probability of co-occurring with the target. So this is an improvement over relying purely on speaker intuition to establish the time of a sense's emergence ie. the very simplest of all possibilities mentioned at the beginning of 4.1.

4.1.4 Emergence time detection

We have suggested the use of a so called 'tracks-plot', to establish C_0, the actual sense emergence time in the corpus, against which an inferred corpus emergence time can be compared.

This suggests the following approach (actually followed in COLING paper Emms and Jayapal [2016])

> *The algorithm produce a time-series of probabilities for senses and a plot of these is made and visually inspected. This might reveal one sense to have a neologistic pattern, being 0 or near zero for a certain span of time, and then climbing away from this. Based on visual inspection, a time (or time-period), is determined as the apparent emergence time inferred by the algorithm. In a similar fashion, by visual inspection of an appropriate 'tracks-plot', an apparent true sense emergence time, C_0, can be determined. These two times can be compared.*

The visual inspection aspect of this is perhaps not ideal: someone can object that the subjectivity can be exploited to give an overly favourable evaluation. So an automatic, objective procedure by which to calculate an emergence time (if any) from a time series would be preferable.

The desire to detect changes in a time series has arisen in a wide variety of application areas, such as quality control in manufacturing processes, navigation system monitoring, seismic data processing, medical condition monitoring, climate change detection, audio segmentation, human activity analysis among others [Basseville and Nikiforov, 1993], [Adams and Mackay, 2007], [Aminikhanghahi and Cook, 2017]. There is quite a substantial literature concerning this, where it is often referred to as 'change point detection'. This is *prima facie* relevant to the topic of this section, the objective determination of a sense emergence time, even if 'change point detection' has a much wider scope. Some of this work is briefly reviewed below. We go on, however, to propose a simple procedure of our own for detecting an emergence time from our time series.

A lot of work in this area uses a 'mean-shift' model, originally considered by Page [1955], and developed by many since. The following outline exposition is based on Basseville and Nikiforov [1993] and Granjon [2013]. One supposes there is a time-series of (real-valued) observations $\boldsymbol{x}_{1:N}$, and then considers two possibilities for their generation. One possibility (H_A) is that for some $j \leq N$ there is an initial part $\boldsymbol{x}_{1:j-1}$ drawn from one distribution, say a Gaussian, with mean μ_0, and that the remnant $\boldsymbol{x}_{j:N}$ is drawn from a second distribution, with mean μ_1. The other possibility (H_0) is that all are drawn from a single distribution, with mean μ_0. The aim is to determine whether the observations are best explained by supposing there actually was a shift between distributions (ie. H_A), and then if there was, to determine the value of j. Algorithms have been developed concerning this based on the (log) of the ratio between the probability of the data under the two hyptheses, which comes to

$$L_j = \sum_{i=j}^{i=N} log(\frac{p(x_i;\mu_1)}{p(x_i;\mu_0)}) \quad \text{or} \quad \sum_{i=j}^{i=N} s(x_i) \quad \text{where } s(x_i) = log(\frac{p(x_i;\mu_1)}{p(x_i;\mu_0)})$$

If the two distributions are known, the basic algorithm proposes decision H_A vs H_0 can be based on the size of the *generalised likelihood*, G, defined to be $max_{1\leq j\leq N} L_j$, and that a change time $\hat{j}$ can be estimated via the maximising point $\arg\max_{1\leq j\leq N} L_j$. When a *cumulative sum* function S_n is defined as $\sum_{i=1}^{n} s(x_i)$, these can be re-expressed

$$G = S_N - \min_{1\leq j\leq N} S_{j-1} \quad \text{and } \hat{j} = \arg\min_{1\leq j\leq N} S_{j-1}$$

leading to the approach being known as the *cumulative sum* or CUSUM algorithm. Figure 4.3 replicates an example of this procedure taken from Granjon [2013]

This rather unlikely situation of the two distributions being completely known is the starting point for further developments, such as when there are two Gaussians of unknown means. One line of development generalises the likelihood ratio further, replacing μ_0 and μ_1 with their

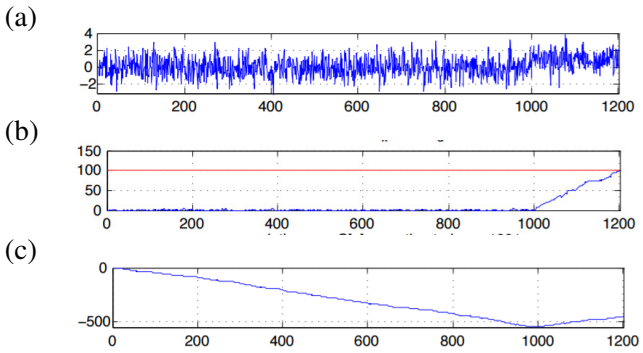


Figure 4.3 – example taken from Granjon [2013]. (a) data, which is samples from successive Gaussians with means $\mu_0 = 0$, $\mu_1 = 1$, changing at sample 1000. (b) generalised likelihodd G, which reaches a threshold value of 100 at sample 1201 (c) cumulative sum S_n giving estimated change-point at sample 1001

maximum likelihood estimates for each conjectured change-point j. For a range of variants of this style of approach there is theoretical work concerning optimality, in various senses, [Lorden, 1971] [Csörgö and Horváth, 1997]

It seems that in the most standard application of change point detection algorithms, the time series considered are values of an *observed* variable: in the above mean-shift model, they are treated as samples from an unknown underlying sequence of Gaussians. On these grounds alone it seems fair to say that the assumptions of this mean-shift model do not transfer, at least straightforwardly, to the problem which is the subject of this section, that of detecting emergence in an estimated time-series $\pi_{1:N}$ resulting from running our EM and GS procedures. The values in our time-series $\pi_{1:N}$ are *sense probabilities*, and in the theoretical development of the algorithms are seen as modes or means of particular posterior Dirichlets, not as sequence of samples from a sequence of Gaussians.

It may well be that one of the other extant techniques in the general area of change point detection (surveyed in Aminikhanghahi and Cook [2017]) is perfectly suited to our sense emergence task. We make no claim to have exhaustively considered the possibilities. In this regard the following quote from Aminikhanghahi and Cook [2017] is worth nothing

> *Several artificial and real-world datasets have been used to measure the performance of CPD algorithms. It is important to notice that an objective comparison of the performance of different CPD methods is very difficult due to the use of these different datasets.*

This suggests that even with considerably greater review of extant work it would still be difficult to identify what technique is the most compelling candidate to be adopted for this sense-emergence dating task.

We turn now to specifying the simple procedure which we will use. Informally we have described a neologistic pattern as being for an initial period zero or near zero, and then departing from this and continuing to climb away from this. We fix some criteria encapsulating this.

Function `EmergeTime(`π`)`**:**
  Surges $= \emptyset$
  for *n:=1 to N-r* **do**
    if `SurgeStart(`n,π`)` **then** Surges = Surges $\cup\{n\}$
  end
  if Surges $\neq \emptyset$ **then return** min(Surges)
  else return $\emptyset$

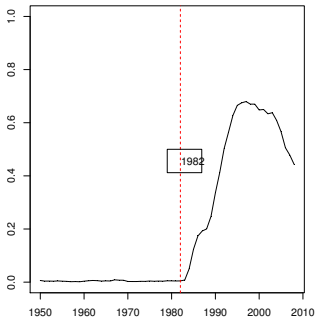
Function `SurgeStart(`n,π`)`**:**
  h = false, l = false
  if $|\{n' : n < n' < n+r$ **and** `Step(`n',π`)` $\}| \geq 85\%\ of\ r$ **then** h = true
  if $|\{n' : n' < n$ **and** $\pi[n'] \leq 0.1\}| > 80\%\ of\ n$ **then** l = true
  if *h* **and** *l* **then**
    return true
  else
    return false
  end

Function `Step(`n',π`)`**:**
  if $\pi[n'] - \pi[n'-1] \geq\ 2/3\%\ of\ max(\pi)$ **then return** true
  else return false

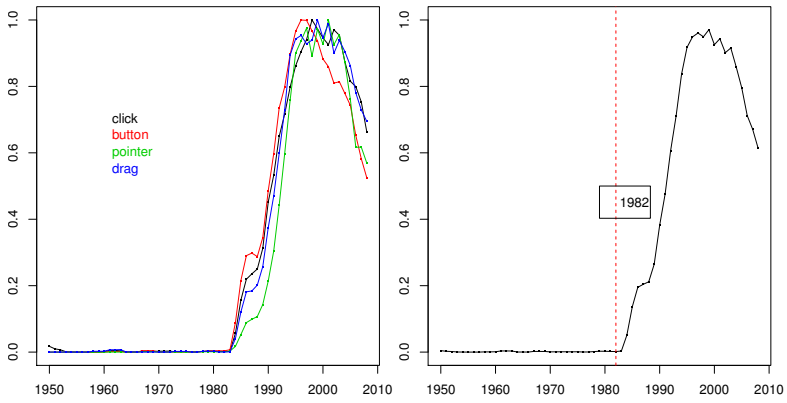
Algorithm 3: `EmergeTime`: input is sequence of sense probabilities for some particular sense, r is the required length of a run of year of sustained increase and in all the later experiments this was set to 10.

To find a sense emergence time from the inferred sense distributions $P(S|Y)$, we look for any year n which is the beginning of a run of years of sufficient year-on-year increase in probability. We set the run of years to be length 10, and define sufficient increase to require 85% of the years in the run show a climb of $2/3$% towards the time series' maximum value. The time series might climb over a prolonged period of time so that several years count as being at the beginning of a period of steady growth, in which case we seek the earliest such year. Additionally such a year ought to be such that 80% of its predecessors are are lower than 0.1. This is stated in pseudo-code in algorithm 3.

For example, applied to one of the inferred sense probability series for *mouse* obtained using EM, the years 1982 1983 1984 1985 1986 1987 1988 qualify as starting 10 year runs of sufficiently strong growth and the earliest of these would be indentified by the procedure as an emergence time, which is depicted in the plot below.



Then to have a reference date based on a 'tracks' plot the same procedure is applied to the single time series obtained by taking a mean of the separate tracks. Below to the left a tracks plot for some words thought especially indicative of the neologistic sense of *mouse* is shown, *click*, *button*, *pointer* and *drag* and to the right is shown the mean of these together with an emergence time which is based on the above procedure.



4.2 Analyzing parameter outcomes

Earlier in this chapter, the different ways to evaluate the neologistic sense identified by the proposed diachronic model were discussed. As discussed earlier, the expectation for a neologism

sense is to start at close to zero and depart in the later years. The OED and 'tracks' plots are the evaluation schemes used to evaluate the identified neologistic sense, but before even going for the evaluation, it would be better to confirm that the inferred neologism sense is the expected neologism sense – for this, further analysis schemes are proposed in this section.

4.2.1 *gist* words

There are a number of changes such as world changes and opinion changes (discussed earlier in section 1.6) but not all these are considered as language changes. When the EM and/or the Gibbs sampling inference procedures appears to have inferred a neologism sense – as suggested by the parameter plot made with $\boldsymbol{\pi}$ – it is still open to doubt whether or not the 'inferred' trend is strongly related to an anticipated sense emergence. One way to try to confirm or dis-confirm this is to look into the word-parameter $\boldsymbol{\theta}$ outcomes[2]. Consider the following illustrations for world changes and opinion changes:

Suppose since 1980 there was a substantial increase in mice population, which in turn brought a lot of diseases. One might expect a parameter component θ_k with increasing probability after 1980. If the high probable words are analyzed in θ_k, one would expect words such as *disease, population* and *deaths*.
Now, let us suppose since 1980 humans found that mice are very intelligent. For such a parameter component θ_k, one might still find increasing probability after 1980. When the high probable words are analyzed, one might expect words such as *feeling, emotion* and *anxious*.

So world or opinion changes might lead to a component being inferred, whose prevalence varies in a distinctive way over time, but such variation should not be taken as a language change. This suggests that one inspect in some way the word probabilities of a particular component.

As for each k, $\boldsymbol{\theta}_k$ is a length V vector of probabilities, where V is the size of the vocabulary, it is something of an issue how best to 'inspect' $\boldsymbol{\theta}_k$.

For a given sense k, we propose to extract a set of top ranked words from $\boldsymbol{\theta}_k$ that are particularly representative of the distribution. Let us call these the *gist* words for a particular sense k. Simply ranking by decreasing probability is not going to be revealing as the top positions will be dominated by generically frequent words (the, and etc). Two different ways of ranking to get the *gist* words are discussed below.

Ranking by comparing a sense-specific word distribution to the corpus word distribution: One could extract *gist* words from the parameter values $\boldsymbol{\theta}_k$ representing $P(w|S=k)$ by computing the ratio of $P(w|S=k)$ to $P_{corp}(w)$, where $P_{corp}(w)$ is the probability of the word w in the corpus. From this ratio, the words can be ranked in the descending order to get the top words associated with each sense. The result of such top 30 *gist* words ranked in descending order made from the parameter outcomes for the target *mouse* example introduced in section

[2]Recall from section 3.2, $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ are the parameter representations for the diachronic model.

4.1.1 is provided in table 4.1.

$P(w\|S0)/P_{corp}(w)$	$P(w\|S1)/P_{corp}(w)$	$P(w\|S2)/P_{corp}(w)$
$cat\|S0 = 3.2008$	$button\|S1 = 2.48445$	$cells\|S2 = 3.99281$
$a\|S0 = 3.08358$	$pointer\|S1 = 2.39148$	$in\|S2 = 3.58596$
$rat\|S0 = 3.02426$	$left\|S1 = 2.32885$	$embryo\|S2 = 2.98432$
$as\|S0 = 2.87897$	$right\|S1 = 2.21559$	$of\|S2 = 2.90259$
$"\|S0 = 2.83585$	$release\|S1 = 2.18371$	$mammary\|S2 = 2.8724$
$keyboard\|S0 = 2.78903$	$over\|S1 = 2.17976$	$embryos\|S2 = 2.7731$
$-\|S0 = 2.6673$	$down\|S1 = 2.13141$	$brain\|S2 = 2.62479$
$,\|S0 = 2.55959$	$move\|S1 = 2.12581$	$cell\|S2 = 2.59001$
$anti\|S0 = 2.55087$	$your\|S1 = 2.10488$	$model\|S2 = 2.52385$
$game\|S0 = 2.52766$	$to\|S1 = 2.02401$	$tumor\|S2 = 2.46576$
$like\|S0 = 2.24782$	$drag\|S1 = 2.02375$	$\_END\_\|S2 = 2.44903$
$or\|S0 = 2.14626$	$you\|S1 = 2.01798$	$development\|S2 = 2.40309$
$and\|S0 = 2.10188$	$subjected\|S1 = 1.97732$	$virus\|S2 = 2.40251$
$such\|S0 = 2.10085$	$hold\|S1 = 1.92721$	$/\|S2 = 2.35402$
$rabbit\|S0 = 2.00737$	$on\|S1 = 1.92396$	$house\|S2 = 2.33416$
$In\|S0 = 1.98873$	$when\|S1 = 1.84758$	$.\|S2 = 2.31284$
$little\|S0 = 1.94476$	$click\|S1 = 1.83455$	$(\|S2 = 2.30812$
$not\|S0 = 1.85856$	$then\|S1 = 1.82935$	$gene\|S2 = 2.28547$
$clicks\|S0 = 1.84501$	$Release\|S1 = 1.80291$	$:\|S2 = 2.24056$
$was\|S0 = 1.82027$	$use\|S1 = 1.78597$	$from\|S2 = 2.20633$
$have\|S0 = 1.80394$	$cursor\|S1 = 1.76889$	$bone\|S2 = 2.1937$
$human\|S0 = 1.78778$	$the\|S1 = 1.7626$	$Mus\|S2 = 2.13185$
$field\|S0 = 1.77676$	$clicking\|S1 = 1.72484$	$marrow\|S2 = 2.10241$
$IgG\|S0 = 1.77356$	$is\|S1 = 1.6685$	$skin\|S2 = 2.10138$
$that\|S0 = 1.7491$	$Move\|S1 = 1.66106$	$embryonic\|S2 = 2.07304$
$but\|S0 = 1.74615$	$position\|S1 = 1.65671$	$adult\|S2 = 2.00754$
$than\|S0 = 1.73704$	$\_START\_\|S1 = 1.64541$	$during\|S2 = 1.89012$
$'\|S0 = 1.72155$	$press\|S1 = 1.62483$	$early\|S2 = 1.88799$
$quiet\|S0 = 1.71718$	$Click\|S1 = 1.58942$	$)\|S2 = 1.88013$
$--\|S0 = 1.70512$	$while\|S1 = 1.58375$	$musculus\|S2 = 1.8767$

Table 4.1 – Top 30 *gist* words for the target *mouse* given each sense listed here are ranked by computing the ratio of $P(W|S=k)$ to $P_{corp}(W)$.

Ranking by two sense-specific word distributions: Another possibility to gain insight into $\boldsymbol{\theta}_k$, the word distribution for a particular k, would be to compare to the other $\boldsymbol{\theta}_{k'}$, making a succession of rankings according to $\frac{P(w|S=k)}{P(w|S=k')}$

From table 4.1, the *gist* words obtained by computing a ratio of $P(w|S=k)$ to $P_{corp}(w)$ seem very consistent with the neologistic sense 'sense 1' for the target *mouse* representing a 'computer pointing device'. Although the other senses may not be of interest for our work, it can be observed that 'sense 0' consistently represents the 'animal' sense and 'sense 2' represents the 'biological experiments' sense of the target *mouse*.

4.2.2 Sense examples

Getting sense examples can be considered as an additional analysis option to be clear about what the neologistic sense represents. The sense examples can be obtained by computing the conditional probability $P(S|Y,\boldsymbol{w})$ for each data item, given by

$$P(S=k|Y=t^d,\boldsymbol{w}=\boldsymbol{w}^d) = \frac{P(S=k,Y=t^d,\boldsymbol{w}=\boldsymbol{w}^d)}{\sum_{S=k'} P(S=k',Y=t^d,\boldsymbol{w}=\boldsymbol{w}^d)}$$

Data items with $P(S = SENSE\ 1 \mid Y = 1990, \boldsymbol{w})$			
L L Drag the T down the R R	1	succession without moving the T R R R R	1
L L dragging the T over them R R	1	L just drag the T pointer R R R	1
L L drag the T pointer through R R	1	L just release the T button R R R	1
L L Drag the T pointer to R R	1	dialog box with the T R R R R	1
L L Drag the T down and R R	1	L displayed when the T pointer R R R	1
L L Drag the T to highlight R R	1	direction you drag the T R R R R	1
L L drag the T pointer down R R	1	L L distance the T has moved R R	1
L L drag the T to move R R	1	L directly with the T . R R R	1
L L drag the T pointer across R R	1	displayed , move the T R R R R	1
L L roll the T to the R R	1	Enter or click the T R R R R	1
row height with the T R R R R	1	highlighted , release the T R R R R	1
L row with the T , R R R	1	highlighting it with the T R R R R	1
L L roll the T on your R R	1	highlight it with the T R R R R	1
L L roll the T over the R R	1	pixel coordinates of the T R R R R	1
sure to release the T R R R R	1	L shrill command the T controls R R R	1

Table 4.2 – Top 30 sense examples words for the target *mouse* given 'sense 1' from the year *1990* with their probabilities listed here are ranked by computing $P(S|Y, \boldsymbol{w})$.

Joint probabilities of the numerator and denominator are computed as in equation 3.14 and use the inferred parameter values to get the conditional probability value for each data item for each sense S. These conditional probabilities for each data item are ranked in descending order to get the most probable sense examples for each sense. Such top examples for each sense in different years can be produced for comparison. It is expected that there are no sense examples for the years before the corpus emergence date C_0 for the target T.

As a further note, the sense examples can be used to compare the examples from two different years – years before and after C_0 of the neologistic sense – consider them to be $Y^o_{c_0}$ and $Y^n_{c_0}$ respectively. The intuition here is, there is no expectation of sense examples from the neologistic sense in $Y^o_{c_0}$ as $P(S|Y)$ before C_0 is expected to be close to 0; and there is a number of sense examples in $Y^n_{c_0}$.

Consider the *mouse* example, for which the *gist* words were produced in the previous section 4.2.1 – here we will see the sense examples for the neologistic sense (sense 1) from 1990 – a time after the true emergence date. In table 4.2, a list of top 30 sense examples for the neologistic sense 'sense 1 from the year 1990 (after C_0) is produced. From the examples listed, it can be observed that all the 'sense 1' examples from the year 1990 (after 'true C_0') are representative of the 'computer pointing device' sense.

Chapter 5

Diachronic dataset & target possibilities

In chapter 3, a simple 'diachronic' model was introduced with a sense parameter $P(S|Y)$ and a word parameter $P(\boldsymbol{w}|S)$ to identify a novel sense for a given target T from a *time-stamped* dataset containing occurrences of T. The parameter updates for the estimation procedures to infer parameter estimates from the data were derived in sections 3.3 and 3.4. To infer values for the parameters of the diachronic model, there is a requirement for time-stamped datasets coming from a substantial time-period. Therefore it is important to establish the dataset possibilities for this work. In this chapter some of the different dataset possibilities for time-stamped data (sections 5.1, 5.2), and further details of the data sources which was actually chosen for experiments (section 5.3) is discussed. In this chapter there is also a discussion on how the target words - *semantic neologisms* were chosen for experiments (section 5.5).

5.1 Downloadable datasets

There are a number of time-stamped resources available for research purposes. In this section there is a discussion on such resources that are *downloadable* and may be useful for the novel sense detection task and during the discussion of each such resource, some information is given on the choices that were made concerning the resource. The two most relevant criteria influencing suitability for the experiments are:

1. length of time-line
2. amount of data per year for a given target.

Table 5.1 provides a corpora list and the first 5 in the list are downloadable resources and the rest are 'web-accessible' resources. Column 'time-line' in the table refers to the number of years of data the corpus holds, column 'Free' refers to whether the corpus is *web accessible, freely downloadable* or *not*, column 'size' provides the size of the corpus, column 'Target occurrences' provides the count of the target occurrences in the corresponding corpus and column 'Relevant work' provides the references of works that use the resource.

Name	Time-line	Free	Size	Target occurrences	Relevant work
TIPSTER	1988-1992	no	4.48×10^6 words	NA	
AQUAINT	1996-2000	no	3.75×10^6 words	NA	
NYT	1987-2007	no	1.8×10^6 articles	NA	Rohrdantz et al. [2011]
EUROPARL	1997-2011	yes	5.3×10^6 words	smashed it: 0 bricked: 1 crawled: 2 mouse: 87 surf: 24 site: 462	
Google 5-gram	1600-2008	yes	3.61×10^{11} words	mouse 2008:51k mouse 2000-09:529k surf 2008: 7k surf 2000-09: 49k bricked 2008:332 bricked:2000-09:1.3k	Wijaya and Yeniterzi [2011] Mitra et al. [2014, 2015] Kulkarni et al. [2014] Gulordava and Baroni [2011]
COCA	1990-2015	web	5.2×10^8 words	mouse 2008: 241 mouse 2000-09: 3k bricked 2008: 4 bricked 2000-09: 39 crawled 2008: 135 crawled 2000-09: 1196	
COHA	1810-2009	web	4×10^8 words	mouse 2008: 6 mouse:2000-09: 601 bricked 2008: 0 bricked 2000-09: 11 crawled 2008: 15 crawled 2000-09: 397	
Web Corp Live	see text	NA	see text	Bricked 2008: 0 Bricked 2000-09:542 mouse 2008: 0 mouse 2000-09: 1810	
Web Corp Diachronic	2000-2010	web	1.3×10^6 words	bricked 2008: 1 Bricked 2000-09: 19 mouse 2008: 288 mouse 2000-09: 2872	
Web corp Synchronic	2000-2010	web	1.27×10^6 words	bricked 2008:0 bricked 2000-09: 60 mouse 2008: 2173 mouse 2000-09: 10417	

Table 5.1 – A list of corpora explored for the Diachronic analysis with further details related to the corpora are provided here – see text for explanation on further details.

The matter of how targets were chosen – known semantic neologisms – is discussed in section 5.5. However to continue with the discussion of the data-sets in some context, it is necessary to mention a few targets. The word *mouse* has a 'pointing device' sense, which it has not always had. The OED gives a first citation of 1965, and according to Wikipedia [Wikipedia, 2016] this particular computer accessory really took off in the 1980s. The words *surf* and *crawl* have senses relating to 'moving through a network of sites making up the WWW (by user and search-engine respectively)' which date from the arrival of that technology, some time after

1995. The investigation of these targets has some implications for what will be relevant time-line. The word *bricked* has a 'render inert' sense and the phrase *smashed it* has a 'was excellent' sense, both of which seem to be quite recent, perhaps in the last 10 years.

Now, continuing with the discussion on data sources, consider the first data source in Table 5.1 TIPSTER[Harman and Liberman, 1993]. It is a downloadable dataset distributed by LDC for TREC (Text Research Collection) related workshops. It contains data from varied sources such as Associated Press, wall street journal, the Federal register and US Patents. It approximately has 448 million words with the data between 1988 – 1992. This time period is too short to be considered for this work.

Another similar corpus we had from LDC[1] is AQUAINT corpus, that has data from various Associated Press, New York Times (NYT) and Xinhua news services between years 1998 – 2000, 1998 – 2000 and 1996 – 2000 respectively. Again this gives a short overall time line. Rohrdantz et al. [2011] has used a 'complete' NYT corpus with newspaper articles that spans between 1987 and 2007 for their work on novel sense detection. It has the advantage of accurate time-stamps as every article has a publication date attached to it. But just a small chunk of the NYT corpus was found as a part of the AQUAINT corpus, so it was opted for this work.

Another possible time-stamped corpus is EUROPARL[2][Koehn, 2005] parallel corpus, which is used mainly for training Machine Translation (MT) systems. The corpus contains European parliament proceedings between 1997 and 2011. When searched for the words *crawled, mouse* and *surf*, the number of occurrences (see Table 5.1) are rather small. The potential targets *bricked* and *smashed it* do not occur at all.

Google N-gram corpus[Michel et al., 2011] is another downloadable dataset, containing N-gram counts from over 5 million digitized books with 361 billion English words, with each N-gram getting a per-year count, based on the year of publication of the book from which it comes. The data covers a very long time-line, from early 17^{th} century till 2008. The corpus has data from uni-grams to 5-grams and out of all these, 5-gram version provides the greatest amount of context words for any given target. For the target *mouse* there is a great deal of data over a long time for the year 1975 there are 9241 occurrences, for the years 1985, 2008 there are 18395 and 51448 occurrences. The per year occurrence of the words seem to be reasonable enough to consider them for the current work and it was adopted as data source. A number of further details of this corpus are discussed in section 5.3.

5.2 Web-accessible datasets

In the previous section 5.1, some downloadable time-stamped data sources were discussed. There are also further sources that may be accessed via the web but are not downloadable in

[1]LDC refers to Linguistic Data Consortium which holds data from various sources and distributed free for some LDC conducted workshops and in other times for some fee. See `https://www.ldc.upenn.edu` – Last accessed on July 23, 2016

[2]EUROPARL refers to European Parliament Proceedings Parallel Corpus 1996-2011. This corpus is available from `http://www.statmt.org/europarl/` – Last accessed on July 23, 2016

their entirety. The second half of table 5.1 provides a list of web-accessible corpora with their details. Particular targets that are semantic neologisms were chosen on such web-accessible sources to see if they could provide the kind of data needed.

Dr. Mark Davies of Brigham Young University has put together a set of widely used corpora and are made web-accessible. One such web-accessible corpus is *Corpus of Contemporary English* (COCA)[3] of size 520 million words coming from Contemporary American English. The range of times covered is 1990 – 2015. Another web-accessible corpus available is *Corpus of Historical American English* (COHA)[4] with 400 million words covering the years 1810 – 2009. For COCA for some of the targets already mentioned this seems an appropriate time-line – though not for *mouse*. For *bricked*, an apparently more recent neologism, the time-line seems appropriate but the returned amount of data is small for this word. For *crawled* the time-line is appropriate and the per-year amounts are possibly large enough. For COHA for the above-mentioned targets the per year amounts are smaller than COCA. The time-line is appropriate for *mouse* but the data amounts are rather small. For both COCA and COHA, a major set-back in using these two corpora is that although search outcomes are visible in the browser, the outcomes of searches are not downloadable as such. For a licensing fee then entirety of the corpus can be downloaded.

WebCorp[5] Live [Renouf et al., 2007] is an online service aiming to facilitate use of the World Wide Web as a corpus. Remarks below concern when it was considered as an option – Aug 2015. Some of its functionalities may have changed since then. When a word or phrase is entered, it forwards the search to a commercial search engine, then from the URLs in the returned hit lists it downloads linked-to documents and processes these further to make a concordance style output. The search-engines called on impose a max size of hit list eg. for Google this is just 64. WebCorp's processing includes dating result pages and there is a possibility to filter the results by date but this does not set a date range for searching. For example searching for 'bricked' returns 562 concordance lines from 64 different pages, but if the date is then set to 2008, no results remain. This is because the original hit list returned via Google contained nothing as old as 2008. So while WebCorp Live may have several features useful for enabling corpus linguistics via the Web, it does not really have the features to enable the construction of a large diachronic data set for a given word or phrase.

In addition to WebCorp Live, they also have a so-called WebCorp-Linguistic Search Engine, which has the ability to search through different corpora such as 'Synchronic English Web Corpus' and 'Diachronic English Web Corpus'. 'Synchronic English Web Corpus' is made from web-extracted texts with 467 million words covering data between 2000 – 2010 and 'Diachronic English Web Corpus' is a 130 million word corpus covering data between Jan 2000 – Dec 2010 with each month having 1 million words. For several of the targets (such

[3]COCA corpus is available online from `http://corpus.byu.edu/coca/` – Last accessed on July 23, 2016

[4]COCA corpus is available online from `http://corpus.byu.edu/coha/` – Last accessed on July 24, 2016

[5]accessible from `http://www.webcorp.org.uk/`

as *mouse*) the time line is not appropriate because of its limitations. For the recent sense of *bricked* the time line is appropriate but as can be seen from the table the number of occurrence is very small.

Lexis-Nexis is a large web-accessible database with data from legal, news and business sources. These are downloadable, but access to this database is *not* free of cost. A custom date based search is available to look for data from particular dates of interest, which allows us to download data belonging to a particular year. Although data from a wide time span is available for online searches, only up to 3000 results of them were available for user download at one time and not all results were made available exclusively for downloads. Further the slowness in rendering the results made it further difficult to download the data from this database.

5.3 Google 5-gram

This is a data set released by Google[6] giving per-year counts[7] of 5-grams in their digitized books holdings. From the entire 5-gram data-set it is possible to pull for a given target word T, a corpus of time-stamped 5-grams (with counts) containing T. Compared to the different corpora discussed in the previous section, Google 5-gram is far larger, covers more times, and should have more accurate time-stamps as they are dependent on the publishing date of a book in getting the time-stamps. One potential disadvantage is they never have a context of more than 4 words.

5-gram	count
diligence and patience the mouse	2
and patience the mouse ate	2
patience the mouse ate in	2
the mouse ate in two	2
mouse ate in two the	2

Table 5.2 – Sample 5-grams for target *mouse* from the year 1821

To explain further about Google 5-gram dataset, it is necessary to emphasize that it is not really a corpus rather a (per-year) frequency table for 5-gram *types*. It is a data-set giving time-stamped counts on 5-gram types arising by sliding a window over the original texts, a window in which a succession of token sequences appear; basically the window contents will contribute to a count if the tokens do not span certain boundaries such as sentence or paragraph endings. Consider the data provided in table 5.2 for the target *mouse* from 1821[8] – these are 5 counts for 5 types. These counts could come from 10 separate 1821 *tokens* of *mouse*, or from just 2 occurrences of[9] *diligence and patience the mouse ate in two the*. In general a token of T could have contributed to the counts of up to 5 different 5-grams. However for this work, the count

[6]There are also 1,2,3 and 4-gram data sets.

[7]They excludes 5-grams with total count < 40.

[8]They are not consecutive in this way in the original data.

[9]By doing date specific search at `http://books.google.com` it can be verified that the latter is true

n for a 5-gram is treated as if the *n* occurrences of the target contribute to no other 5-grams for that target, which amounts to treating the 5-grams as if they were independent miniature documents. There is no real practicable alternative to this.

The Google n-gram books dataset has been used for sense change detection in various works Kulkarni et al. [2014], Mitra et al. [2014, 2015], Wijaya and Yeniterzi [2011] for experiments.

In addition to all the feel-good factors about the Google n-gram dataset, there are also a few criticisms that may be noteworthy here: (i) the n-gram corpus has been composed by digitizing the books using a OCR, so there are chances of errors. However Google has claimed in their latest release[10] that, "Books with low OCR quality and serials were excluded" – here low OCR quality refers to 80% accuracy [Michel et al., 2011]. (ii) According to Michel et al. [2011], the publishing date of the books has some errors with the likelihood of error in a randomly sampled book from the corpus between 1800 – 2000 stands less than 6.2%. Such error analysis seems left untested for the period 2001 – 2008.

Even with these limitations, considering the volume of data available for analysis such errors are intuitively not expected to have a larger impact in the experiment outcomes.

5.4 Comparing eras

Rather than a real diachronic corpus, some works are based on texts representing just two eras. Few such examples are the works of Cook et al. [2013, 2014], Lau et al. [2012]. They have used British National Corpus (BNC) [Clear, 1993] corpus and UKWAC[Ferraresi et al., 2008] corpus for this work. Instead of considering the occurrences of target from different times (say a number of years), they have considered all the occurrences of a target from BNC corpus to be representing data from late 20th century. To compare with, they have used UKWAC corpus having data from 2008. In their work, they have just compared two times ie., 2008 to the later 20th century. As they have considered BNC corpus for comparing two era, the possibility of BNC was also considered for this work. BNC is a 100 million word corpus widely used for various text processing utilities with data between 1960 and 1993, while the large majority of text spans between 1985 and 1993. Given a very short time-span 1985 – 1993, the possibility of the usage of this corpus for our work is ruled out. However for consistency reasons, the number of occurrences for the words *mouse* and *bricked* were fetched, which produced 1728 and 32 occurrences of the words in the corpus. This kind of corpus may be ideal for comparing text from two different eras, as the majority of the data are from a very short time-span.

5.5 Choosing targets

To test the proposal we require words which are genuine semantic neologisms . As already noted, there is no gold-standard reference data-set to turn to. The starting point for the targets

[10]Their latest release happened in 2012 and we have considered this version of data for all our work.

chosen was native speaker intuition, and to some extent, some of the prior work.

For any such potential target with some supposed novel sense it was checked that the OED records this sense, and has a first citation date for it. It is also necessary that the novel sense be present in the 5-gram data. Although the Google 5-gram data is very large, it is not absolutely certain to contain examples of every sense of every word. For example, a 5-gram sub-corpus for the potential target *crawled* was extracted, containing 214048 5-grams between 1970 – 2008. Some intuitively associated words for its newer sense include *web, internet, page, url, index, indexing*, but far from having a point in time where they have increasing frequency in the sub-corpus, these do not occur even once. Similarly a sub-corpus for the potential target *bricked* contained 6096 5-grams between 1970 – 2008. However when words are taken which you might intuitively expect in the context of its newer 'render device inert' sense, such as *mobile, phone, OS, android, apple, software*, they did not occur in its 5-gram sub-corpus. For such items it seems the novel sense is not present in the 5-gram data, and it is not possible to establish a tracks-based date against which to compare any inferred sense-emergence date.

It is quite a time-consuming process to extract a sub-corpus for a particular target from the entire 5-gram corpus. To try to get a quick preliminary impression on the presence or absence of particular novel sense we made some use of the Google n-gram viewer, which offers possibilities to get a preliminary impression of what particular tracks-plots *would* be produced on the sub-corpus. For example, plots in figure 5.1 shows information obtained via the online N-gram viewer with respect to the targets *mouse* and *surfing*. First, let us consider

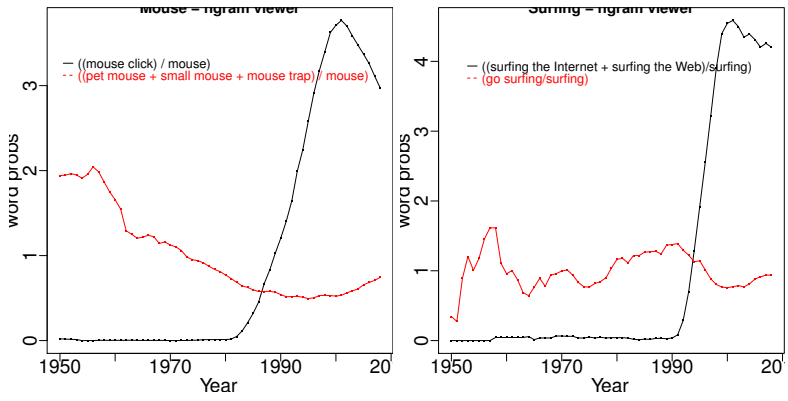


Figure 5.1 – Plots of words adjacent to *mouse* and *surf* are based on values from ratio queries given to the Google n-gram viewer API – values are normalized by their means over time. The black line in each case concerns adjacent words intuitively particularly associated with an emerging sense, the red lines concerns words associated with a long established sense.

the target *mouse* provided in the left hand plot of figure 5.1: for this, the intuitively associated word in the 'pointing device' sense is *click*. Online the Google n-gram viewer can be given the query *((mouse click) / mouse)*, which looks for 2-gram counts of *mouse click* and is divided by the count of *mouse*, which is indicative of the probability of the word *click* immediately next to *mouse*. The results for such a query can be downloaded using the Google N-gram viewer

API[11]. The plot in figure 5.1 is based on such downloaded numbers and the additional step has been taken of dividing the values by their mean (over the time-span) – the same applies to other plots shown earlier based on the Google N-gram viewer API. This division by the mean is to place different queries on a comparable scale. From the plot one can observe that the word *click* has an emerging trend (where the word *click* occurring with the target has close to zero occurrence during the initial period and takes off later), which provides a suggestion that the dataset has the emerging sense for the target *mouse*. To provide a further comparison with the emerging trend the intuitively associated words (*pet, small, trap*) with the other sense is also plotted. Similar to this, consider another target *surfing* provided in the right hand plot of figure 5.1: for this *the internet, the web* – intuitively associated words in the novel usage of 'activity of moving from site to site' are searched for by the Google N-gram viewer for the query *((surfing the Internet + surfing the Web)/surfing)* and the combined relative counts for the words *the Internet + the Web* are plotted. This shows the emerging trend for the novel sense, which confirms the reasoning that the word *surfing* is likely to occur in the 5-gram corpus.

Such plots were used to anticipate the behavior in a 5-gram sub-corpus. It is important to note that the plots created by accessing the Google N-gram viewer use ratios where the denominator is a count for some 1-gram x, and the numerator is based on counts for extending 2-grams and 3-grams x', and so do not reproduce exactly the conditional probabilities that will be obtained from a target-specific 5-gram sub-corpus[12].

Target	New sense	OED
mouse	computer pointing device	1965
gay	homosexual person	1941
strike	industrial action	1904
bit	basic unit of information	1958
compile	transform to machine code	1965
paste	duplicate text in computer edit	1981
surf	exploring internet	1992
boot	computer start up	1981
rock	genre of music	1960
stoned	under drug influence	1965
hip	up-to-date; smart, stylish	1904
export	to transmit data out of computer	1982
mirror	to copy data on to a different server	1993
domain	A subset of locations on the network	1982
high	under the influence of drugs	1932

Table 5.3 – The table provides the information for targets that are neologisms

In the end, the targets[13] *mouse, gay, strike, bit, compile, paste, surf, boot, rock, stoned, hip, export, mirror, domain, high* were chosen for experiments using the diachronic model in

[11]The API is available from `https://github.com/econpy/google-ngrams`

[12]Besides the difference in length of n-gram and the fact the 5-gram conditional probabilities are not position specific, there is also the fact that for each length of n-gram Google applies a frequency threshold of 40 to the released n-gram corpus.

[13]Explicit dictionary definitions and citation informaton from the online Oxford English Dictionary (OED) for the chosen targets are provided in the appendix A.1

identifying a novel sense associated with them from Google 5-gram corpus. This is provided in table 5.3.

In some cases, targets which were originally prompted by intuition coincide with targets considered in other work: [Wijaya and Yeniterzi, 2011] also considered *gay* and *mouse*. Some were prompted by the works [Cook et al., 2013, 2014, Lau et al., 2012]. They considered *domain, export, mirror, poster, worm, visit, feed, site, platform*. Of these, based on preliminary indicators via the n-gram viewer we adopted also *domain, export, mirror*. In Mitra et al. [2014, 2015] a number words are discussed which the system suggested to be semantic neologisms and which, according to the authors, truly are. The assessment of some of these as truly 'sense-births', relative to the dates discussed, seems wrong, going by the OED (eg. *hooker, amp, bum, sissy, thug, dude*). We did not adopt as targets any of their sense-birth words[14]

[14]For a number of others, although their claimed novelty was consistent with the OED, our above-mentioned preliminary investigations via the n-gram viewer suggested the absence of the proposed novel sense in the n-gram data (eg. *giants, donation, partition, passwords*)

Chapter 6

Neologism Experiments - Google 5-grams

Semantic neologisms are the existing words with some sense that takes a new sense at some time. The examples for this are discussed in chapter 1. As discussed in that chapter, it is not easy to find an emerging sense from a large data set containing occurrences of any given semantic neologism target with time-annotations. For this, a diachronic model was introduced in chapter 3 and the parameter updates are also derived for EM and Gibbs sampling parameter estimation schemes, that can supposedly infer the emerging sense from raw text.

For this model, experiments using EM and Gibbs sampling algorithms are conducted based on Google 5-gram datasets (section 5.3) for different targets to find the relevance of the diachronic model in finding a neologistic sense. These experiment outcomes are reported in this chapter. In section 6.2 of this chapter, a set of experiments are formulated to test the model with manipulated datasets and this is called as *Pseudo-neologism* tests. This work is carried out with motivation from the *pseudo-word* tests conducted by Schütze [1998] for unsupervised 'word sense discrimination' task. Then, EM and Gibbs sampling experiments for genuine neologism targets are reported in section 6.3, where further analysis based on the experiment outcomes are also performed. Then in section 6.4, following the methodology of [Cook et al., 2013, 2014, Lau et al., 2012] similar experiments are conducted for non-neologism targets to prove that the outcomes of 'genuine' neologisms are not accidental.

Also, following Cook et al. [2013, 2014], Lau et al. [2012] methodology, a new method is devised to discriminate neologism targets from the the non-neologism ones based on a so-called 'novelty-scores' (section 6.5). Further in section 6.6, the EM and Gibbs sampling experiments for the 'genuine' neologism targets that did *not* produce the expected outcomes are also reported.

In addition to the initial model tests (section 6.2) conducted, few other tests were also conducted: (i) to understand the impact of data on the estimation procedures, "Ablation" tests were conducted and those are reported in section 6.7.1 and (ii) in an attempt to infer the number of senses required to find a neologistic sense, "merge tests" were conducted and its outcomes

are reported in section 6.7.2.

6.1 Generating sub-corpus

Google 5-gram corpus[1] is a large collection[2] of per year counts of 5-grams organized in alphabetical order[3]. Therefore, it is not an easy task to pull all the occurrences of a target T from the complete collection during an experiment. So a sub-corpus was made for each T by collecting all the occurrences of T coming from different years – this was done using a script, which opened each zip file from the complete dataset, looked up for the occurrence of T in each line of the file and based on year of occurrence, the 5-gram gets into a separate file corresponding to the year; this way the sub-corpus had all the occurrences of T from all the years. As a single process, the extraction process takes approximately 36 hours for a target.

6.2 pseudo-neologism model test

The 'pseudo-word' technique was introduced by Schütze [1998] as a way to test un-supervised word-sense discrimination. It can be given a diachronic twist to furnish what might be called 'pseudo-neologisms' in the following way. Consider two unambiguous words T_1 and T_2 with T_1 in use throughout the time period, but T_2 emerging at time T_e in the period. If the 5-grams for T_1 and T_2 are then all treated as examples of the fake word '$T_1 - T_2$' this functions as an artificial semantic neologism, manifesting T_2's sense only from t_e onwards, and furthermore for all $t > t_e$, T_2's sense is present to the exact extent to which the T_2 5-grams contribute to the merged set of T_1 and T_2 5-grams for year t.

This technique is used, for the time-period 1850–2008, choosing *ostensible* for T_1, and *supermarket, genocide, byte* as possibilities for T_2. For each target, a separate dataset is extracted with which datasets T_1 and T_2 are combined considering them to be one $T_1 - T_2$ dataset. The dataset sizes for each target with their usage information is provided in table 6.1 – the 'Lines' column provides information on the number of data items (5-grams) associated with the target, 'OED' column provides the first citation date for the target from the online Oxford English Dictionary, while the next two columns are the emergence date information for the EM and Gibbs sampling algorithm outcomes. These dates are obtained by the emergence time detection algorithm discussed in section 4.1.4. Furthermore the proportion of T_2 n-grams relative to all n-grams (either T_1 or T_2) in a given year gives gold-standard for the created pseudo-neologism. The 'gold standard' column in table 6.1 gives the emergence time obtained from this sequence.

In this section, the EM and Gibbs sampling parameter outcomes for the said pseudo-neologisms are discussed – for all the EM and Gibbs sampling experiments in this section, the parameter distributions were assigned as discussed in sections 3.3.4 and 3.4.3 respectively.

[1] We use version 2 of the corpus generated in July 2012

[2] The whole collection of 5-grams occupies approximately 235 Gigabytes of storage in a compression format.

[3] 5-grams starting with different letters are compressed together in separate files

Target	Lines	usage	OED	EM Date	GS Date	gold standard
byte	119k	unit of information	1961	1965	1965	1965
genocide	94k	mass killing	1944	1946	1947	1946
supermarket	231k	a large self-service shop	1931	1949	1949	1949
ostensible	118k	appearing to be true	NA	NA	NA	NA

Table 6.1 – Words used to generate pseudo-neologisms

For EM, the parameter 'wc_min' (recall section 3.3.3) was set to 0 and for the Gibbs sampling experiments the hyper-parameters $\boldsymbol{\gamma}_\pi$ and $\boldsymbol{\gamma}_\theta$ were set to 1 – these settings assume an uniform (un-informative) prior over the parameters.

6.2.1 byte-ostensible

The *byte-ostensible* dataset is formed by combining data from *byte* and *ostensible* datasets and considered them to be one single dataset. The EM and Gibbs sampling experiments were conducted on the *byte-ostensible* dataset between years 1850 and 2008, where the *ostensible* dataset had data between 1850 - 2008, while *byte* had data in the period 1933 - 2008. The OED date for *byte* is 1961 (as can be seen from Table 6.1), but the dataset had data for *byte* from the year 1933 – on manually looking into the dataset, it was identified that there were small amount of data in the years before 1960, which may be accounted for wrong time annotation in the dataset. However all the data from the datasets were considered for the experiments.

The plot in figure 6.1 shows the inferred and empirical outcomes for *byte-ostensible* pseudo-neologism running with a sense setting $K = 2$, asking for the algorithms to produce two sense parameter outcomes. Here K is set to 2 as it is known that T_1 (*byte*) and T_2 (*ostensible*) are unambiguous and so has only one sense associated with each target. Both EM and Gibbs sampling inference procedures learns values for the parameter distributions $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$; the sense parameter parameter $\boldsymbol{\pi}$ has values for every time t in each sense k where $\sum_k \pi_t[k] = 1$ and; the word parameter $\boldsymbol{\theta}$ has values for every word w from vocabulary V in sense k where $\sum_w^V \theta_k[w] = 1$.

The first two plots are made from EM and Gibbs sampling algorithm's inferred $\boldsymbol{\pi}$ sense parameter outcomes – for each k, the succession of $\pi_t[k]$ values. EM, giving a point estimate, these values are produced by the algorithm, while the Gibbs sampling algorithm produces a series of Gibbs samples, from which the mean of the samples are considered for $\boldsymbol{\pi}$ and the plots are made. In the plots, the red line represents 'sense 0' made from inferred $\boldsymbol{\pi}[k = 0]$ and the blue line made from inferred $\boldsymbol{\pi}[k = 1]$ labeled as 'sense 1'. The third plot show the 'gold-standard' sense probabilities, i.e., the proportions of *byte* and *ostensible* in each year.

The plots made out of the inference outcomes are very alike and both the inference procedures infers 'sense 0' to have an apparent neologistic pattern, with the inflection seen in year 1965 – the emergence dating information are further shown as 'EM-Date' and 'GS-Date' in Table 6.1. These inferred dates are close to that apparent from the 'gold-standard' plots (and consistent with the OED citation date). It is expected from the inferred estimates that the 'red' line in the plot representing neologistic pattern is associated with the 'byte' sense, which can further be confirmed with the *gist* words – table 6.2, provides a list of top 30 *gist*

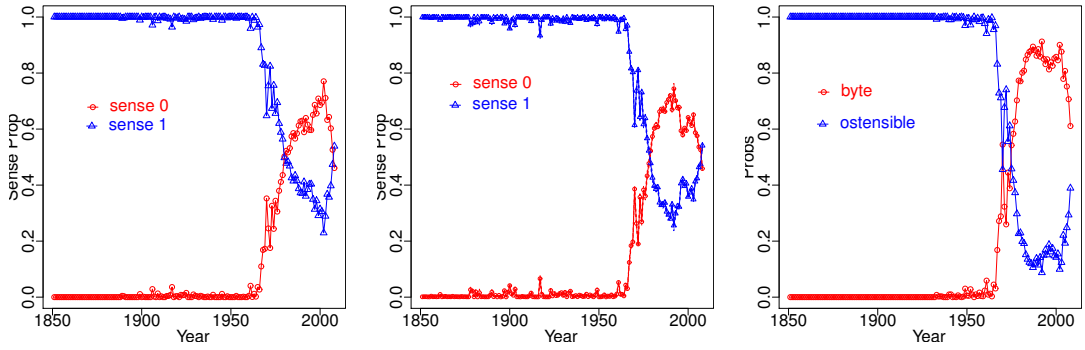


Figure 6.1 – The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *byte-ostensible* pseudo-neologism, and the third plot shows the known *byte* and *ostensible* proportions

gist words - EM outcome	**gist words - Gibbs sampler outcome**
gist(sense 0) byte: *-, a, (, [,],), 8, order, 1, bit, single, bits, 4, 2, one, word, _END_, ., time, A, two, byte, called, eight, by, character, array, short, low, field*	gist(sense 0) byte: *-, a, order, bits, significant, bit, 8, 1, word, single, ., _END_, 2, one, 4, time, A, low,), at, most, two, by, byte, least, called, eight, high, character, field*
gist(sense 1): *purpose, reason, was, The, object, first, for, the, second, cause, no, its, his, subject, of, last, their, this, _START_, means, that, with, which, next, ground, significant, least, motive, authority, aim*	gist(sense 1): *purpose, reason, was, [,], object, The, for, cause, no, his, its, subject, second, their, the, _START_, first, means, this, public, short, b, of, last, int, that, which, ground, motive*

Table 6.2 – Top 30 *gist* words for *byte-ostensible* pseudo-word

words obtained from the word parameter distributions $\boldsymbol{\theta}_k$ for 'sense 0' and 'sense 1' and they correspond to *byte* and *ostensible* related senses respectively. The words such as *8, bit, order* and *significant* are representative of the word *byte* – 'a unit of information' sense. These words are consistently seen from both the EM and Gibbs sampler outcomes except for a few changes in the word order in the top 30 'gist' words as seen from table 6.2. Additionally, it can also be observed that the top 30 'gist' words for 'sense 1' is consistent with the *ostensible* sense.

6.2.2 genocide-ostensible

Figure 6.2 shows the EM and Gibbs sampling inferred and empirical outcomes for *genocide-ostensible* pseudo-neologism running with sense setting $K = 2$. For the Gibbs sampling outcomes, *mean* of the Gibbs samples are plotted. From the inferred outcomes, 'sense 0' associated with *genocide* sense is identified to be a neologism as expected, further the 'EM-Date' and 'GS-Date' are very close to the OED citation date for *genocide* – the dating information are provided in table 6.1.

The top 30 'gist' words for 'sense 0' and 'sense 1' from EM and Gibbs sampling outcomes are provided in table 6.3. The 'gist' words such as *crime, war, humanity* and *commit* are representative of the word *genocide* – 'mass killing' sense, can be seen consistently in both EM and Gibbs sampling top 30 'gist' words list. Also, the 'gist' words associated with 'sense 1' are consistent with the *ostensible* word sense and this also confirms with the 'gist' words from *byte-ostensible* outcomes.

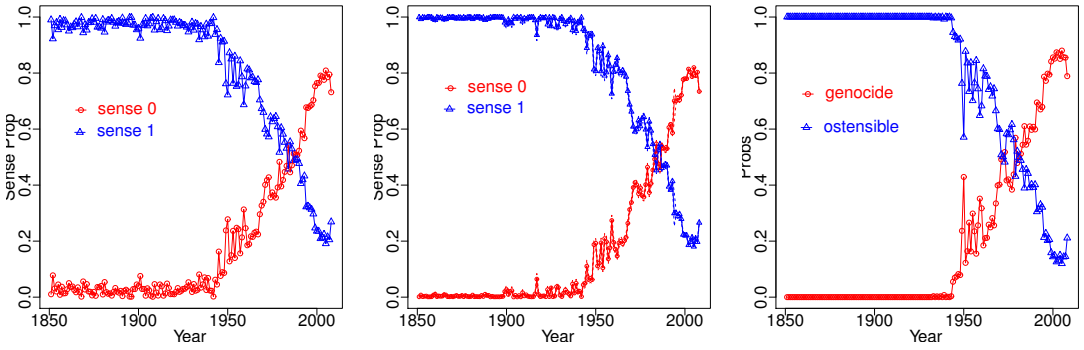


Figure 6.2 – The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *genocide-ostensible* pseudo-neologism, and the third plot shows the known *genocide* and *ostensible* proportions

gist words - EM outcome	**gist words - Gibbs sampler outcome**
gist(sense 0) genocide: *crimes, ", Rwanda, against, _END_, ., commit, humanity, and, war, ', 1994, in, Rwandan, Nazi, cultural, cleansing, Armenian, ethnic, crime, acts, term, as, form, word, victims, slavery, policy, ;, ,,*	gist(sense 0) genocide: *crimes, against, Rwanda, ", _END_, ., commit, humanity, in, war, and, ', 1994, Nazi, Rwandan, cultural, Armenian, cleansing, ethnic, Jews, term, during, crime, word, victims, slavery, as, policy, acts, form*
gist(sense 1): *purpose, reason, The, object, _START_, was, cause, no, for, his, means, subject, its, is, this, their, which, with, any, ground, that, aim, motive, than, authority, whose, without, Convention, has, His*	gist(sense 1): *purpose, reason, The, object, _START_, was, cause, no, for, with, any, means, his, that, is, subject, its, this, which, their, ground, aim, motive, authority, than, whose, without, Convention, has, His*

Table 6.3 – Top 30 *gist* words for *genocide-ostensible* pseudo-word

6.2.3 supermarket-ostensible

Figure 6.3 shows the inferred and empirical outcomes for *supermarket-ostensible* pseudo-neologism running with sense setting $K = 2$. As can be seen on the plots, the inferred outcomes are analogous with the empirical estimate. In the figure, it can be seen that the neologistic sense in the inferred outcomes are colored blue while in the empirical plot, it is colored red – this is because the inference procedure decides assigning the senses, however it gets clear when the 'gist' words are analyzed from the inferred outcomes.

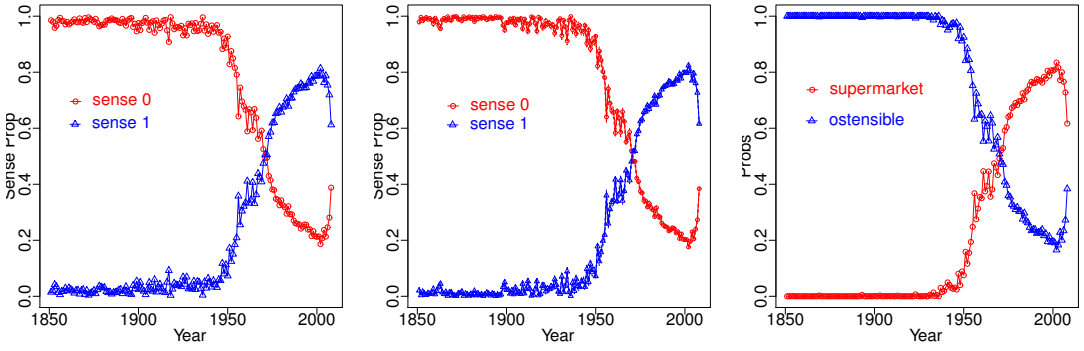


Figure 6.3 – The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *supermarket-ostensible* pseudo-neologism, and the third plot shows the known *supermarket* and *ostensible* proportions

gist words - EM outcome	gist words - Gibbs sampler outcome
gist(sense 0): *purpose, The, reason, object, was, _START_, for, cause, no, that, its, subject, his, of, which, this, is, means, their, an, ground, with, aim, motive, whose, His, without, Its, well, reasons*	gist(sense 0): *purpose, The, reason, object, was, _START_, for, cause, no, that, its, his, subject, this, which, of, their, is, means, with, ground, an, aim, motive, whose, His, without, Its, reasons, head*
gist(sense 1) supermarket: *at, _END_, a, ., local, in, go, your, or, from, checkout, chain, shelves, line, you, store, buy, shopping, to, parking, section, lot, ?, shelf, went, large, I, out, chains, department*	gist(sense 1) supermarket: *at, a, ., local, _END_, in, go, your, chain, checkout, shelves, from, or, line, you, store, buy, shopping, -, parking, to, section, lot, ?, shelf, went, chains, large, I, out*

Table 6.4 – Top 30 *gist* words for *supermarket-ostensible* pseudo-word

The 'gist' words of 'sense 1' such as *buy, checkout, chain* and *shelves* are consistent with the *supermarket* sense, similarly the 'gist' words with respect to *ostensible* sense are consistent with the earlier experiment outcomes.

Thus on these pseudo-neologisms, the proposed algorithm has been successful, identifying an emerging 'sense' in an unsupervised fashion. Moving on from this first test, the next section considers outcomes on authentic words. As a further note, for all the experiments mentioned in this section the value for K has been set to 2. This is because we know the number of senses involved for the 'pseudo-word' targets, but for the experiments that will be encountered further in this chapter we do not know the actual number of senses associated with the target words. Therefore, the required value for K cannot be determined before the experiments were conducted and may seem arbitrary.

6.3 Neologism targets

The 'pseudo-neologism' experiments provided evidence that the 'diachronic model' can detect sense emergence from artificial 'pseud-neologism' data. Experiments on genuine targets are presented in this section. The targets[4] *mouse, surf, gay, bit, boot, compile, strike, rock, stoned* are considered for the actual neologism experiments – these are the words which, relative to particular time periods, are known to exhibit sense emergence.

For each target, EM-inferred, Gibbs sampler inferred time-lines of $\boldsymbol{\pi}_t[k]$ values for each k will be plotted. The time-lines of the $\boldsymbol{\pi}_t[k]$ values give a visual impression of whether a sense emergence has been detected. The emergence time detection algorithm discussed in section 4.1.4 is applied to the time-lines of the $\boldsymbol{\pi}_t[k]$ values and for any time-line which the algorithm reports as showing an emergence, the detected time is reported.

Also for each target a 'tracks' plots is made for the anticipated novel sense – the co-occurring words used are discussed when the outcomes for each target are presented below. From this a tracks-based emergence date is calculated following the procedure described in section 4.1.4.

After a comparison of the emergence times based on EM and GS inferred $\boldsymbol{\pi}_t[k]$ time-lines with the tracks-based emergence date, there is further analysis of the inferred word-given-sense probabilities by considering 'gist' words and sense examples, as described in section 4.2.

[4] Explicit dictionary definitions and citation informaton from the online Oxford English Dictionary (OED) for the chosen targets are provided in the appendix A.1

Target	Years	Lines	Min Occs	Max Occs	New sense	Vocab size
mouse	1950-2008	910k	1263k	6318k	computer pointing device	5109
gay	1900-2008	1253k	1253k	5573k	homosexual person	6686
strike	1800-2008	5052k	3462k	17311k	industrial action	9088
bit	1920-2008	7393k	7407k	37037k	basic unit of information	13808
compile	1950-2008	214k	174k	873k	transform to machine code	1414
paste	1950-2008	318k	272k	1360k	duplicate text in computer edit	1965
surf	1950-2008	182k	137k	687k	exploring internet	1657
boot	1920-2008	1285k	907k	4538k	computer start up	5312
rock	1920-2008	4136k	2910k	14553k	genre of music	10163
stoned	1930-2008	12k	5k	28k	under drug influence	251

Table 6.5 – Google 5 gram dataset - the table provides the information for targets that are neologisms

Target	D_0^c **(OED)**	D_0^i **LDCOE/COE**	C_0 **(Tracks)**	**GS-Date**	$GS < 10\%$	**EM-Date**	$EM < 10\%$
mouse	1965	1987/1990	1982	1982	yes	1982	yes
gay	1941	1978/1976	1969	1969	yes	1970	yes
strike	1810	1978/1911	1899	1901	yes	1904	yes
bit	1948	1978/1976	1966	1958	yes	1958	yes
compile	1952	1978/1976	1972	1965	yes	1965	yes
paste	1975	1995/2004	1982	1981	yes	1981	yes
surf	1992	1995/2004	1993	1992	yes	1992	yes
boot	1980	1987/2004	1982	1980	yes	1981	yes
rock	1956	1987/1990	1967	1960	yes	1960	yes
stoned	1952	1978/1976	1959	1960	yes	1965	yes

Table 6.6 – D_0^c(OED) gives the date of first citation in the OED, under D_0^i LDCOE (resp. COE) gives the date of first dictionary inclusion in LDOCE (resp. COE), C_0(Tracks) gives a tracks-based corpus-emergence data in the relevant Google 5 gram sub-corpus, GS-Date and EM-Date give sense emergence dates derived from $\boldsymbol{\pi}_t$ parameters inferred by GS and EM, and $GS < 10\%$ and $EM < 10\%$ indicate whether these agree with 10% of the time-span with C_0(Tracks)

Table 6.5 gives some size information for sub-corpora for the targets. The 'Years' column provides the year span considered for the EM and Gibbs sampling experiments, 'Lines' column is the number of data items (5-grams) with target T and 'New sense' column indicates the expected neologistic sense associated with T. The 'Max occs' column is the maximum number of tokens of T these lines could represent and is the sum of all of T's 5-gram frequencies, while 'Min occs' is the minimum number of tokens T these lines could represent $(= max/5)$.

Table 6.6 gives some dating information. For each target D_0^c(OED) gives the date of first citation in the OED. Following the discussion in chapter 4 this is expected to be a lower bound to the true corpus emergence date. D_0^i gives dates of first inclusion of the novel sense inclusion in a sequence of dictionary editions; the LDOCE numbers are from 4 editions of the Longman Dictionary of Contemporary English (1978,1987,1995,2005), the COE are from 4 editions of the Concise Oxford Dictionary (1911,1976,1990,2004). Following the discussion in chapter 4 this is expected to be an upper bound to the true corpus emergence date. C_0(Tracks) gives the tracks-based emergence date for the target in the relevant Google 5-gram corpus and this

C_0(Tracks) will be treated as a reference with which to compare the emergence date that is based on the inferred parameters via EM and GS. One thing to note from Table 6.6 is that the proposed 'tracks'-based dating by which to identify the corpus emerge date, C_0(Tracks), leads to the expected ordering $D_0^c < C_0(\text{Tracks}) < D_0^i$.

In applying EM and Gibbs sampling algorithms to the 5-gram data each time-stamped 5-gram for T is treated as if it's count n representing n unique tokens of T. When a single token of T does contribute to 5 different 5-grams then the words closer to T will play a greater role in determining the overall apparent data probability than words further way — a word at T_{-1} will appear in 4 of the 5-grams, whilst a word at T_{-4} will appear in one of the 5-grams. Thus one thing that the experiments address is whether this simple way of using the 5 gram data is undermined by this or not.

The number of senses, K, is an input parameter to the EM and GS algorithms. It is not a feature of these algorithms to determine any kind of optimal value for this K. As each target is considered in turn an initial experiment is done with a rather conservative $K = 3$ (for comparison Frermann and Lapata [2016] adopt $K = 10$). In several cases a neologistic sense is then detected. In cases where this does not occur, following the intuition that the neologistic sense may be proportionately less prevalent than several other longer standing sense, we conduct further experiments with $K = 4$ or $K = 5$. This issue of the number of senses is discussed a little further in section 6.7.2.

6.3.1 mouse

In figure 6.4 the first 2 plots provide the inferred EM and Gibbs sampling estimates and also 'tracks' plot for the target *mouse*, with the algorithms run with 3 sense variants. In the EM case for each sense k single solid line shows a sequence of estimated $\pi_t[k]$ for different t. In the Gibbs case, for each k the solid line shows again shows a sequence of estimated $\pi_t[k]$ values, although these are *means* over the Gibbs samples. Additionally, in the Gibbs case the dotted lines around the solid line are the 'min' and 'max' of 90% HPD interval (section 3.4.6) obtained from the inferred Gibbs samples. For both EM and Gibbs the blue line for the inferred $\boldsymbol{\pi}_t[k=1]$ values shows a neologistic pattern: according to the `EmergeTime` algorithm of section 4.1.4 in both cases the emergence time is 1982. For this target and others, the 'EM-Date' column and 'GS-Date' columns of Table 6.6 gives this inflection point obtained by applying the emergence time detection method (discussed in section 4.1.4) over the $\boldsymbol{\pi}_t[k]$ values. Although for this target both EM and Gibbs have found $k = 1$ to have a neologistic pattern (ie. the *blue* line in both) this is really a coincidence. It is also visible in the plots that the other two senses have switched.

The 'tracks' plot (see 4.1.3) shows tracks for *click, button, pointer* and *drag* – words that we expect to be especially associated with the neologistic 'computer peripheral' sense of mouse. . The emergence time based on this according to the procedure in section 4.1.4 is also 1982. For this target and for others the columns '$EM < 10\%$' and '$GS < 10\%$' of table 6.6 compares this

tracks-based date with the 'EM-Date' and 'GS-Date' respectively. For *mouse* the three dates coincide. The table also gives 'OED' first citation date and in this case it is more than 15 years earlier.

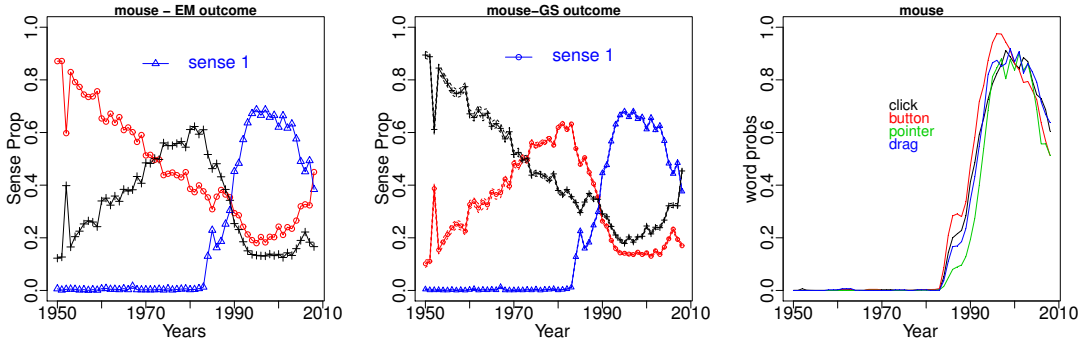


Figure 6.4 – The first and second plots show the EM and Gibbs sampling algorithm's inferred $\boldsymbol{\pi}[k]$ sense parameter outcomes for *mouse* experiment, and the third plot shows the probability 'tracks' for some words that are intuitively associated with the 'computer peripheral' sense of *mouse*. Dating information – EM:1982, GS:1982, OED:1965, Tracks: 1982)

gist words - EM outcome	**gist words - Gibbs sampler outcome**
gist(sense 1) **neologism**: *button, pointer, left, right, release, over, move, down, your, drag, you, hold, to, then, on, when, Release, cursor, use, clicking, click, Move, position, press, Click, while, changes, When, moving, user*	gist(sense 1) **neologism**: *button, pointer, left, click, right, you, over, release, your, down, move, to, drag, _START_, is, hold, use, when, then, Release, or, cursor, clicking, on, ,, Move, can, position, press, it*

Table 6.7 – Top *gist* words for the neologism sense for target *mouse* ranked by comparing word distributions to corpus Probabilities

Sense examples - EM outcomes	**Sense examples - Gibbs sampling outcomes**
L L dragging the T pointer across R R	L L dragging the T pointer across R R
L L drag the T to move R R	L L drag the T to move R R
L L drag the T pointer across R R	L L drag the T pointer across R R
L L roll the T to the R R	sure to release the T R R R R
L L Roll your T over the R R	L just release the T button R R R
L L rolls the T over the R R	L L rolls the T pointer over R R
L L rolls the T pointer over R R	L rollers inside the T . R R R
L L roll your T over any R R	L just hover your T pointer R R R
L rollers inside the T . R R R	L row with the T , R R R
L L roll the T pointer over R R	succession without moving the T R R R R

Table 6.8 – Top neologism sense examples for the target *mouse* extracted from inferred EM and Gibbs sampling estimates for *sense 1*

Table 6.7 provides the top 30 'gist' words for the neologistic sense when ranked by comparing inferred $\boldsymbol{\theta}_{k=1}$ distribution to P_{corp} distribution (discussed in section 4.2.1). For nearly all these 'gist' words (such as *button, pointer, press, move* and *release*) they seem very much associated with the 'pointing device' usage of the target 'mouse'. There are some (such as *when, your, then*) where the association is not obvious but they also do not seem especially associated other senses. For completeness the top 30 'gist' words for the other senses are also provided for reference in 6.29(a).

Data items with $P(S = SENSE\ 1 \mid Y = 1990, \boldsymbol{w})$		Data items with $P(S = SENSE\ 1 \mid Y = 1970, \boldsymbol{w})$	
L L Drag the T down the R R	1	sleeve , threw the T R R R R	1
L L dragging the T over them R R	1	L L L L T with pale throat Burrows	1
L L drag the T pointer through R R	1	L L L L T promptly becomes absorbed in	1
L L Drag the T pointer to R R	1	L crushed and ridiculed T promptly R R R	1
L L Drag the T down and R R	1	pars distalis of the T R R R R	1
L L Drag the T to highlight R R	1	, crushed and ridiculed T R R R R	1
L L drag the T pointer down R R	1	L _START_ While the T is R R R	1
L L drag the T to move R R	1	L L L ridiculed T promptly becomes absorbed R	1
L L drag the T pointer across R R	1	L L and ridiculed T promptly becomes R R	1
L L roll the T to the R R	1	With shrill command the T R R R R	1
row height with the T R R R R	1	our kites do a T R R R R	1
L row with the T , R R R	1	L shrill command the T controls R R R	1
L L roll the T on your R R	1	L to hold a T in R R R	1
L L roll the T over the R R	1	_START_ Every time the T R R R R	1
sure to release the T R R R R	1	L L dragging the T in the R R	1
succession without moving the T R R R R	1	L L L the T on the left R	0.99
L just drag the T pointer R R R	1	L L L the T on the right R	0.99
L just release the T button R R R	1	L L L L T in your left hand	0.99
dialog box with the T R R R R	1	L _START_ As the T moves R R R	0.89
L displayed when the T pointer R R R	1	L _START_ Place the T in R R R	0.86
direction you drag the T R R R R	1	L you want the T to R R R	0.83
L L distance the T has moved R R	1	L L L L T (left) and	0.74
L directly with the T . R R R	1	L _START_ When the T is R R R	0.56
displayed , move the T R R R R	1	L L L L T (right) .	0.55
Enter or click the T R R R R	1	L L When the T is placed R R	0.51
highlighted , release the T R R R R	1		
highlighting it with the T R R R R	1		
highlight it with the T R R R R	1		
pixel coordinates of the T R R R R	1		
L shrill command the T controls R R R	1		

Table 6.9 – Top neologism sense examples for the target *mouse* – extracted from inferred EM estimates for *sense 1*

Additionally for EM and Gibbs, table 6.8 gives examples of cases whose most probable sense is $k = 1$. In particular for the year 2008, it gives a top 10 sense examples out of cases whose most probable sense is $k = 1$, as ranked by their probability to have sense $k = 1$. The sense examples provided in the table is presented the way the data was provided for the algorithms for inference, where T refers to the target, L's and R's are the pad words when the n-gram is short of 4 words to the left and right of the target. The sense examples are consistent with the 'computer pointing device' sense.

Following is a further insight of looking into the sense examples: Based on the 'tracks' plot in figure 6.4, the corpus emergence date C_0 appears to be 1983. Therefore it is expected that there is (literally) no sense examples plot for 'sense 1' (neologistic sense) before 1983. Table 6.9 gives (i) sense $k = 1$ examples from the year 1990 and (ii) sense $k = 1$ examples from the year 1970, seeking a top 30 in each case. Those from 1990 are consistent with the 'computer pointing device' sense. From 1970 the model returned just 24 examples whose most probable sense is $k = 1$. Some are not at all related to the neologistic sense and some possibly are, possibly due to wrong year annotations coming from Google n-grams dataset.

As an indication of the time-taken for the experiments, for the *mouse* dataset with approximately $910k$ data items, the EM run for 1 iteration takes 8.01 seconds, while the Gibbs sampler takes 5.5 seconds for 1 iteration. With these, for *mouse* the total time taken for 60 EM iterations turned out to be approximately 8 minutes and the total time taken for 10000 Gibbs iterations is

approximately 15.2 hours.

6.3.2 gay

In figure 6.5 the first 2 plots provide the inferred EM and Gibbs sampling estimates and also 'tracks' plot for the target *gay*, with the algorithms run with 3 sense variants. The word 'gay' according to OED always used to refer to 'being happy', but at some time during the 20th century it took on an additional 'homosexual' sense, which is by now its predominant sense. For both EM and Gibbs the black line for the inferred $\boldsymbol{\pi}_t[k=2]$ values show a neologistic pattern: according to the `EmergeTime` algorithm of section 4.1.4 the emergence dates are 1970 and 1969 respectively.

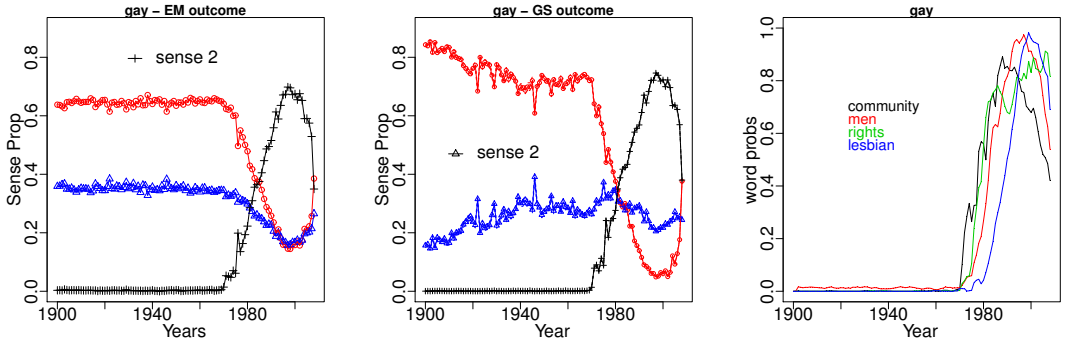


Figure 6.5 – The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *gay* experiment, and the third plot shows the probability 'tracks' for some words that are intuitively associated with the 'homosexual person' sense of *gay*. Dating information – EM:1970, GS:1969, OED:1941, Tracks:1969

gist words - EM outcome	**gist words - Gibbs sampler outcome**
gist(sense 2) **neologism**: *lesbian, lesbians, men, bisexual, rights, movement, /, women, liberation, couples, for, straight, male, studies, community, and, issues, parents, people, among, against, Lesbian, or, communities, (, relationships, families, youth, movements, abortion*	gist(sense 2) **neologism**: *lesbian, men, lesbians, rights, bisexual, community, movement, /, liberation, straight, male, women, couples, people, for, or, studies, parents, issues, who, anti, identity, among, (, marriage, Lesbian, against, communities, relationships, have*

Table 6.10 – Top 30 *gist* words for the target *gay* ranked by comparing word distributions to corpus Probabilities

The third plot in figure 6.5 show tracks for *community, men, rights* and *lesbian* – words that we expect to be associated with the neologistic 'homosexual' sense of *gay*. The emergence time based on this according to the procedure in section 4.1.4 is 1969. For *gay*, the 'EM-Date' and 'GS-Date' are very close to the 'tracks' date. The 'OED' first citation date of 1941 in this case is around 28 years earlier than the tracks-based and the inferred sense emergence dates.

Table 6.10 provides the top 30 'gist' words for the neologistic sense when ranked by comparing inferred $\boldsymbol{\theta}_{k=2}$ distribution to P_{corp} distribution. For completeness the top 30 'gist' words for the other senses of *gay* are also provided for reference in 6.29(c). For nearly all these 'gist'

Sense examples - EM outcomes	Sense examples - Gibbs sampling outcomes
HIV seropositive and seronegative *T* R R R R	L L L L *T* / bisexual / transgender
perspectives on lesbian , *T* R R R R	L L L / *T* / bisexual studies R
history : Lesbians and *T* R R R R	L L L L *T* / lesbian rights movement
L HIV infection in *T* men R R R	L L L L *T* / lesbian liberation movement
perceived workplace discrimination against *T* R R R R	L L L L *T* rights and feminist movements
Permanent partners : Building *T* R R R R	L L L L *T* liberation and feminist movements
L high - risk *T* men R R R	L L L L *T* / lesbian university students
L gender studies , *T* and R R R	les / bi / *T* R R R R
Gender roles among Latino *T* R R R R	L L L L *T* / lesbian / bi
L L L General *T* and lesbian travel R	L L feminist and *T* liberation movements R R

Table 6.11 – Top neologism sense examples for the target *gay* extracted from inferred EM and Gibbs sampling estimates for *sense 2*

words (such as *lesbian, men, rights, movement* and *relationships*) they seem especially associated with the 'homosexual' usage of the target *gay*. Further these words are unanimously found from both the inferred outcomes. Also none seems conspicuously identified with a different sense.

6.3.3 strike

The word *strike* has been in existence for a long time, and according to OED it had multiple senses since 12th century, such as 'hit' and 'find a deal'. After industries and unions were formed the word *strike* acquired a new sense relating to 'industrial action' – the OED date for this is 1822. So a long time-span between 1800 and 2008 has been considered for this target – due to this long time-span it turns out that there is a lot of data items (approximately 5052*k*, provided in table 6.5). All these data were considered for the experiments.

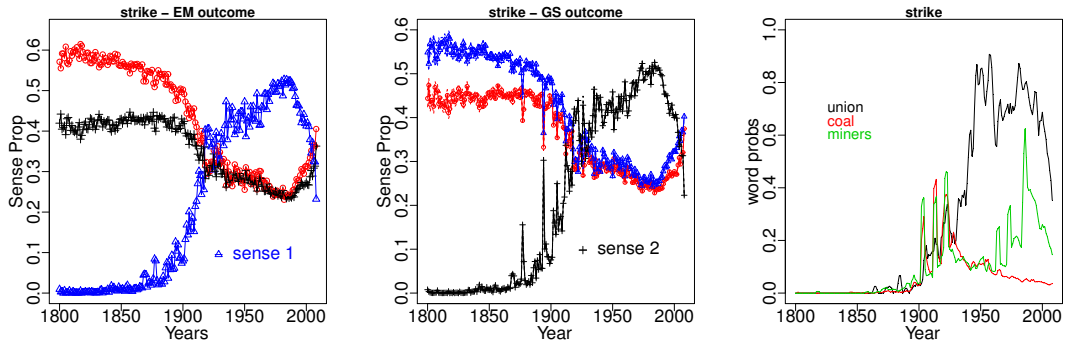


Figure 6.6 – The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *strike* experiment, and the third plot shows the probability 'tracks' for some words that are intuitively associated with the 'industrial action' sense of *strike*. Dating information: EM:1904, GS:1901, OED:1822, Tracks: 1899

In figure 6.6 the first 2 plots provide the inferred EM and Gibbs sampling estimates and also 'tracks-plot' for the target *strike*, with the algorithms run with 3 sense variants ($K = 3$). For the EM case, the blue line for $\boldsymbol{\pi}_t[k = 1]$ values shows a neologistic pattern according to `EmergeTime` algorithm of section 4.1.4, with emergence time 1904, while for the Gibbs case,

gist words - EM outcome	gist words - Gibbs sampler outcome
gist(sense 1) **neologism**: -, *general, of, went, hunger, on, ', price, in, slip, by, called, The, workers, during, ., miners, _END_, day, was, coal, after, no, the, first, emptive, capability, a, sit, lock*	gist(sense 2) **neologism**: -, *general, of, went, ', hunger, on, in, slip, by, price, miners, called, The, workers, during, day, ., coal, _END_, was, after, no, the, first, great, emptive, capability, a, sit*

Table 6.12 – Top 30 *gist* words for the target *strike* ranked by comparing word distributions to corpus Probabilities

Sense examples - EM outcomes	Sense examples - Gibbs sampling outcomes
L shopmen ' s *T* of R R R	L L Boston police *T* of 1919 R R
L L shearers ' *T* . _END_ R R	L L anthracite coal *T* of 1902 R R
short distances along the *T* R R R R	L L L L *T* by anthracite coal miners
short - lived general *T* R R R R	L L anthracite coal *T* in 1902 R R
L shop stopped to *T* the R R R	L L slip , *T* - slip R R
L L shopmen 's *T* of 1922 R R	L L San Francisco *T* of 1934 R R
L Shortly after the *T* , R R R	L pre - emptive *T* against R R R
L L shopmen 's *T* . _END_ R R	nation - wide railway *T* R R R R
L Shortly after the *T* began R R R	L L L railway *T* of 191 1 R
short - lived hunger *T* R R R R	L L L L *T* of bituminous coal miners

Table 6.13 – Top neologism sense examples for the target *strike* extracted from inferred EM and Gibbs sampling estimates for *sense 1* and *sense 2* respectively

it is the black line for $\boldsymbol{\pi}_t[k=2]$ values which shows the neologistic pattern, with emergence time 1901. Here, EM and Gibbs have found different sense numbers $k=1$ and $k=2$ with neologistic patterns. Such allocation of sense numbers by both the algorithms are random.

The 'tracks-plot' shows tracks for *union, coal, miners* – words which we expect to be especially associated with the neologistic 'industrial action' sense of *strike*. The emergence time based on this according to the procedure in section 4.1.4 is 1899. From the dating information provided for *strike* in table 6.6, the 'tracks' date is is still closer to the 'EM-Date' and 'GS-Date' and all these dates are later than the 'OED' first citation date.

For EM and Gibbs, table 6.12 provides the top 30 'gist' words for the neologistic sense when ranked by comparing inferred $\boldsymbol{\theta}_{k=1}$ and $\boldsymbol{\theta}_{k=2}$ distributions to P_{corp} distribution. (For the 'gist' words for the other senses see 6.29(g)). For most of these 'gist' words (such as *hunger, miners, workers* and *general*) they seem especially associated with the 'industrial action' usage of the target *strike*. Further these words are unanimously found from both the inferred outcomes. There are some (such as *slip, emptive, capability*) which seem unlikely in the context of usage with *strike* in its 'industrial action' sense. 'strike-slip' (or 'strike slip') is a technical term from geology. 'pre-emptive strike' and 'strike capability' are phrases from discussions of military matters. Figure 6.7 shows the tracks plots for *slip, emptive, capability*). They all show a steep increase in the 2nd half of the 20th century. Even though the apparently neologistic sense components ($k=1$ for EM and $k=2$ for Gibbs) have a climbing trend starting earlier than that, as the other components do not have a climbing trend, this is arguably why the model has accommodated these aspects of the context of *strike* within the neologistic sense component. For completeness the top 30 'gist' words for the other senses are also provided for reference in 6.29(g). Additionally, for EM and Gibbs table 6.13 gives examples of cases whose most probable sense is $k=1$ and $k=2$. In particular for the year 1950, it gives a top 10 sense

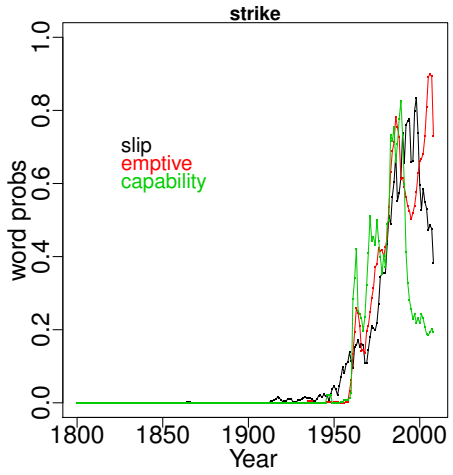


Figure 6.7 – Tracks plot for words *slip, emptive, capability*

out of cases whose most probable sense is $k = 1$ and $k = 2$, as ranked by their probability to have senses $k = 1$ and $k = 2$ respectively.

It can be observed from the Gibbs sampling estimates the HPD interval is very narrow for the neologism sense compared to the other senses, assuring that the model is extremely confident about the inference estimates obtained for the neologism sense.

6.3.4 bit

In figure 6.8 the first two plots provide the inferred EM and Gibbs sampling estimates for the target *bit* and also a 'tracks-plot' with the algorithms run with 3 sense variants $K = 3$.

For both EM and Gibbs the black line for the inferred $\pi_t[k = 2]$ values show a neologistic pattern: according to the `EmergeTime` algorithm of section 4.1.4 in both cases the emergence time is 1958. On looking further into the Gibbs sampling inferred plot, it can be observed that the HPD interval for all the senses are narrow and are very close to the mean of the estimates, which indicates the higher confidence of the model.

The 'tracks'[5] plot show tracks for *8, 32, 64, memory* – words which we expect especially associated with the neologistic 'a basic unit of information' sense of *bit*. The emergence time based on this according to the procedure in section 4.1.4 is 1966, which 8 years later than the EM and Gibbs inferred emergence dates. This may be because there are words other than the ones used for 'tracks' that are associated with the emerging sense and has had an impact in the inference outcomes. However all these dates are later than the 'OED' first citation date.

Table 6.14 provides the top 30 'gist' words for the neologistic sense when ranked by comparing inferred $\boldsymbol{\theta}_{k=2}$ distribution to P_{corp} distribution. (For the 'gist' words for the other senses see 6.29(d)). For nearly all these 'gist' words (such as *16, 32, 8, rate* and *significant*) they seem especially associated with the 'a basic unit of information' usage of the target *bit*. There are

[5]The tracks plot for target *bit* has labels with 'X' padded to their left – this is done due to the technical constraints with the plotting function in R.

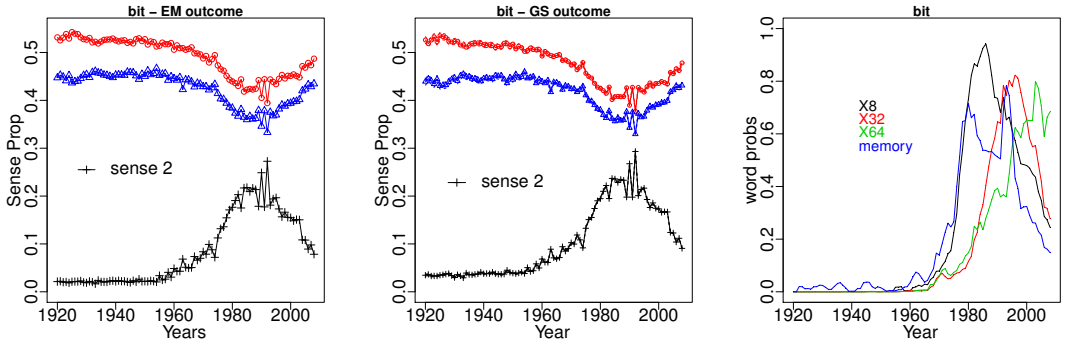


Figure 6.8 – The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *bit* experiment, and the third plot shows the probability 'tracks' for some words that are intuitively associated with the 'basic unit of information' sense of *bit*. Dating information – EM:1958, GS:1958, OED:1948, Tracks: 1966

gist words - EM outcome	**gist words - Gibbs sampler outcome**
gist(sense 2) **neologism**: *-, 16, 32, 8, bit, by, rate, (, significant, –, 64, data,), 4, 1, an, address, error, The, binary, number, /, word, most, bus, two, set, 24, or, register*	gist(sense 2) **neologism**: *-, 16, 32, 8, as, bit, by, rate, (, significant, every, –, data, 64,), an, 4, 1, The, address, error, binary, most, number, word, /, set, the, bus, two*

Table 6.14 – Top 30 *gist* words for the target *bit* ranked by comparing word distributions to corpus Probabilities

some (such as *the, by, /*) where the association is not obvious but they also do not seem especially associated with other senses. Additionally for EM and Gibbs, table 6.15 shows examples of cases whose most probable sense is $k = 2$. In particular for the year 2008, it gives a top 10 sense out of cases whose most probable sense is $k = 2$, as ranked by their probability to have sense $k = 2$.

6.3.5 compile

For this experiment, a sub-corpus just for *compiling* was first created, but when it was seen that this dataset had just 75644 items it was supplemented also with a sub-corpus for *compile*. The datasets pertaining to these words were considered to be a single dataset for the inference

Sense examples - EM outcomes	**Sense examples - Gibbs sampling outcomes**
perform a 64 - *T* R R R R	L Unsigned 32 - *T* integer R R R
pixels with 8 - *T* R R R R	L unsigned 32 - *T* integer R R R
pieced together bit by *T* R R R R	L Unsigned 64 - *T* integer R R R
pixel , 8 - *T* R R R R	L Unsigned 16 - *T* integer R R R
piece and bit by *T* R R R R	L unsigned 16 - *T* integer R R R
piece , bit by *T* R R R R	L unsigned 64 - *T* integer R R R
piece together bit by *T* R R R R	L L L L *T* analog / digital converter
L full 16 - *T* data R R R	L 32 32 - *T* registers R R R
L full 24 - *T* color R R R	L 16 32 - *T* registers R R R
full adder for each *T* R R R R	L unsigned 32 - *T* integers R R R

Table 6.15 – Top neologism sense examples for the target *bit* extracted from inferred EM and Gibbs sampling estimates for *sense 2*

procedures. This process of making a sub-corpus covering more than one alternate form of a base word is a simple work-around for the fact that the 5-gram data is not itself reduced to base forms. In the discussion we will for simplicity just refer to the 'target' compile.

compile has long standing senses such as 'to gather' or 'to put together', but after the introduction of computer and programming languages, it acquired a new usage 'transform to machine code'. In figure 6.9 the first 2 plots provide the inferred EM and Gibbs sampling estimates and also 'tracks' plot for the target *compile*, with the algorithms run with 3 sense variants ($K = 3$). For both EM and Gibbs the *black* line for the inferred $\boldsymbol{\pi}_t[k = 2]$ values show a neologistic pattern: according to the `EmergeTime` algorithm of section 4.1.4 in both cases the emergence time is 1965.

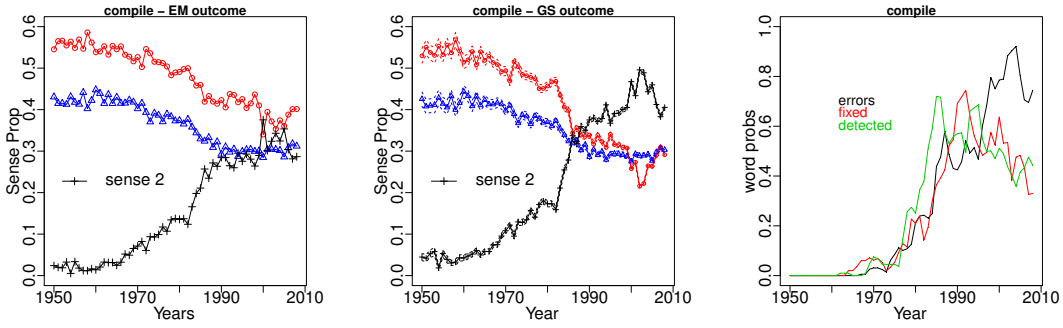


Figure 6.9 – The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *compile/compiling* experiment, and the third plot shows the probability 'tracks' for some words that are intuitively associated with the 'transform to machine code' sense of *compile/compiling*. Dating information – EM:1965, GS:1965, OED:1952, Tracks: 1972

gist words - EM outcome	**gist words - Gibbs sampler outcome**
gist(sense 2) neologism: *time, at, -, error, known, not, will, run, determined, link, ,, ., edit, _END_, –, errors, disseminate, or, editing, archive, type, rather, done, and, than, checking, occurs, because, detected, At*	gist(sense 2) neologism: *time, at, -, run, error, known, ,, link, program, and, ., _END_, execute, code, source, or, determined, edit, not, application, –, will, disseminate, rather, errors, than, your, archive, type, it*

Table 6.16 – Top 30 *gist* words for the target *compile* ranked by comparing word distributions to corpus Probabilities

Sense examples - EM outcomes	**Sense examples - Gibbs sampling outcomes**
L L enforced at T time . R R	L L L L T - link - run
L L enter , T , and R R	L L L L T time rather than run
L L Enter , T , and R R	L L L - T - run cycle R
L L generate a T - time R R	L L L L T - time error occurs
L L generated at T time . R R	L L L either T time or run R
L L L Periodically T – (A R	L L edit - T - run R R
L L performed at T time and R R	L L L L T time or at run
L L performed at T time , R R	L L L L T time or run time
L L performed at T time . R R	L L L - T - link - R
L L performed at T - time R R	L L L at T time or run R

Table 6.17 – Top neologism sense examples for the target *compile* extracted from inferred EM and Gibbs sampling estimates for *sense 2*

The third plot in figure 6.9 show tracks for *errors, fixed, detected* – words that we expect to be associated with the neologistic 'transform to machine code' sense of *compile*. The emergence time based on this according to the procedure in section 4.1.4 is 1972. For *compile*, the 'EM-Date' and 'GS-Date' are not too far from the 'tracks' date. The 'OED' first citation date is 1952 and in this case it is 13 years earlier.

Table 6.16 provides the top 30 'gist' words for the neologistic sense when ranked by comparing inferred $\boldsymbol{\theta}_{k=2}$ distribution to P_{corp} distribution (for the 'gist' words for the other senses see 6.29(f)). For nearly all these 'gist' words (such as *run, error, link, type*) they seem especially associated with the 'transform to machine code' usage of the target *compile*. Looking at the gist words via GS, the list contains more things that stand out as related to computer code than the EM version (eg. code, source, program). If you look at the other senses, the EM version has some of these under other senses. Also comparing the $\pi_t(k=2)$ plots, the GS version climbs to a higher value than the EM plot. One could argue that the GS version has done a better job at clearly isolating the computer-related sense in this case.

Additionally, for EM and Gibbs table 6.17 gives examples of cases whose most probable sense is $k=2$. In particular for the year 1990, it gives a top 10 sense out of cases whose most probable sense is $k=2$, as ranked by their probability to have sense $k=2$. Combining two datasets for *compile* and *compiling* was just a choice made during experiments and had no intention to artificially manipulate the dataset to get a positive outcome.

6.3.6 paste

In figure 6.10 the first 2 plots provide the inferred EM and Gibbs sampling estimates and also 'tracks' plot for the target *paste*, with the algorithms run with 3 sense variants. The word 'paste' always used to refer to 'a sticky semi-solid substance', but after the introduction of computer operating systems, it took on an additional 'duplicate text/images in computer edit' sense. For EM and Gibbs, the 'black' line for the inferred $\boldsymbol{\pi}_t[k=2]$ and 'blue' line for the inferred $\boldsymbol{\pi}_t[k=1]$ shows a neologistic pattern: according to the `EmergeTime` algorithm of section 4.1.4 in both cases it is 1981.

gist words - EM outcome	gist words - Gibbs sampler outcome
gist(sense 2) **neologism**: *cut, copy, you, can, want, ", -, –, You, V, Copy, +, and, Cut, ', scissors, Ctrl, text, Clipboard, method, Paste, then, also, where, operations, out, eugenol, •, oxide, job*	gist(sense 1) **neologism**: *cut, copy, you, can, want, -, ", –, Add, and, You, Copy, scissors, tomatoes, ', Cut, _START_, then, ,, text, ounce, Clipboard, method, When, Paste, Stir, out, also, where, •*

Table 6.18 – Top 30 *gist* words for the target *paste* ranked by comparing word distributions to corpus Probabilities

The 'tracks' plot show tracks for *cut, copy, text, image, clipboard* – words that we expect to be associated with the neologistic 'duplicate text/images in computer edit' sense of *paste*. The emergence time based on this according to the procedure in section 4.1.4 is 1982. For *paste*, the 'EM-Date' and 'GS-Date' are very close to the 'tracks' date. The 'OED' first citation date of 1975 in this case is just 7 years earlier than the 'tracks'-based and the inferred sense emergence

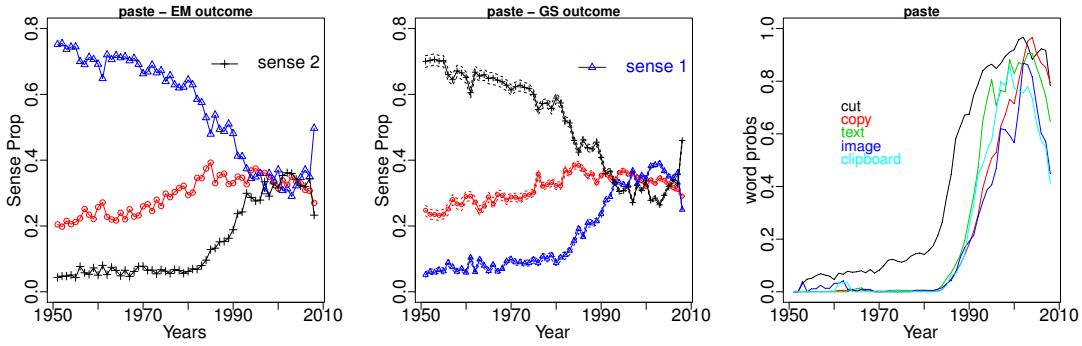


Figure 6.10 – The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *paste* experiment, and the third plot shows the probability 'tracks' for some words that are intuitively associated with the 'homosexual person' sense of *paste*. Dating information – EM:1981, GS:1981, OED:1975, Tracks:1982

Sense examples - EM outcomes	Sense examples - Gibbs sampling outcomes
L just copy and *T* the R R R	Edit ¿ Paste to *T* R R R R
L just cut and *T* . R R R	Paste Special command to *T* R R R R
L just cut and *T* it R R R	where you wish to *T* R R R R
L just copy and *T* it R R R	You may want to *T* R R R R
L just cut and *T* the R R R	Edit , Paste to *T* R R R R
L L layout and *T* - up R R	You can cut and *T* R R R R
L L layout , *T* - up R R	You could cut and *T* R R R R
L enables you to *T* the R R R	_START_ You can even *T* R R R R
perpetual , pistareen , *T* R R R R	You can copy and *T* R R R R
L L pistareen , *T* - pot R R	You can cut , *T* R R R R

Table 6.19 – Top neologism sense examples for the target *paste* extracted from inferred EM and Gibbs sampling estimates for *sense 2* and *sense 1*

dates.

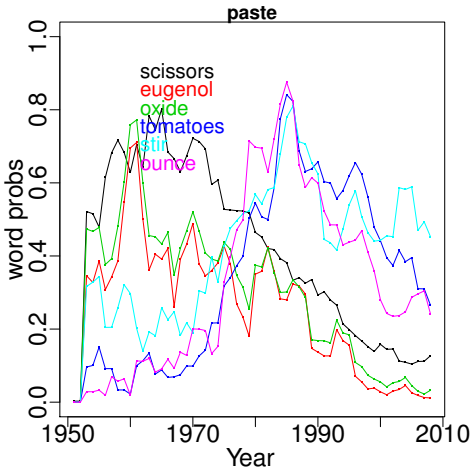


Figure 6.11 – Tracks plot for words *scissors, eugenol, oxide* in comparison with the words *cut, copy, text, image, clipboard*

For the EM and Gibbs sampling outcomes, table 6.18 provides the top 30 'gist' words for

the neologistic sense when ranked by comparing inferred $\boldsymbol{\theta}_{k=2}$ and $\boldsymbol{\theta}_{k=1}$ distributions to P_{corp} distribution (for the 'gist' words for the other senses see table 6.29(j)). For many of these 'gist' words (such as *cut, copy, Ctrl, text, clipboard*) they seem especially associated with the 'duplicate text/images in computer edit' usage of the target *paste*. Further these words are unanimously found from both the inferred outcomes. However, the words *scissors, eugenol, oxide, tomatoes, stir, ounce* stands out as rather unexpected in the context of a 'duplicate text/images in computer edit' usage of *paste*. A 'tracks' plot (in figure 6.11) for *scissors, eugenol, oxide* associated with the neologistic sense of EM outcome are clearly against the tracks for the words associated with the 'duplicate text/images in computer edit' sense, so there is no particular reason to be attributed for the inclusion of such words to the neologistic sense – these can just be considered as 'false' positives. However, tracks for the words *tomatoes, stir, ounce* associated with the neologistic sense of GS outcome seem to start close to zero and go up around 1970. Even though this trend has a somewhat earlier starting point than that for the neologistic sense of *paste*, the absence of other inferred sense components with an increasing trend perhaps explain why the model has accommodated these aspects of the context of *paste* within the component predominantly associated with the 'duplicate text/images in computer edit' usage.

Additionally, for EM and Gibbs, table 6.19 gives examples of cases whose most probable senses are $k = 2$ and $k = 1$. For EM and Gibbs, in particular for the year 1995 it gives a top 10 sense examples out of cases whose most probable senses are $k = 2$ and $k = 1$, as ranked by their probability to have senses $k = 2$ and $k = 1$.

6.3.7 surf

Besides long established senses (relating to waves and to a particular water sport), after the evolution of the internet the word *surf* acquired a new usage 'exploring the internet': the 'OED' first citation date for this is 1992. For this experiment, sub-corpora for several forms of the word *surf* for the period 1950 – 2008 were extracted and then combined; the forms were *surf*, *surfed*, and *surfing*. This was done for the same reason as for the experiments relating to *compile*. Initially a sub-corpus just for *surfing* was created, but it was found that it contained only 15k items. Merging the separate sub-corpora is a simple work-around for the fact that the 5-gram data is not itself reduced to base forms.

For all the targets discussed earlier, the experiments were conducted with the number of senses set to $K = 3$, and it turned out that the inference procedures discovered the neologistic sense . It may be that a given neologistic sense is too minor relative to other senses to be detected with $K = 3$ and it seems target *surf* is an example of such a case.

In figure 6.12, the left and right-hand side plots show the EM inferred outcomes for the target *surf/surfing/surfed* with 3 and 4 sense settings. In neither case was one of the senses assessed to be a neologism – in the 4-sense case the *green* line in the plot for 'sense 3' visually resembles a neologistic sense though displaced upwards. So, a further experiment with $K = 5$

was conducted.

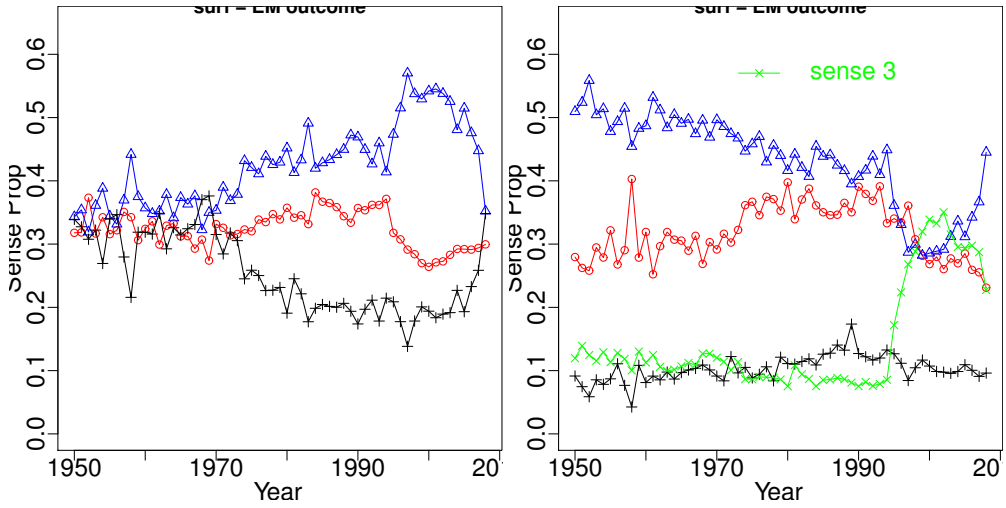


Figure 6.12 – EM inferred plot – *surf/surfing/surfed* – 3 and 4 sense settings

In figure 6.13 the first 2 plots show the the inferred EM and Gibbs estimates, with the algorithms run with 5 sense variants ($K = 5$), whilst the third plot gives a 'tracks' plot for the target *surf*. For the EM case, the *green* line for the inferred $\pi_t[k = 3]$ values shows a neologistic pattern and for the Gibbs case, the *purple* line for the inferred $\pi_t[k = 4]$ values shows a neologistic pattern: according to the `EmergeTime` algorithm of section 4.1.4 the emergence date in both cases is 1992.

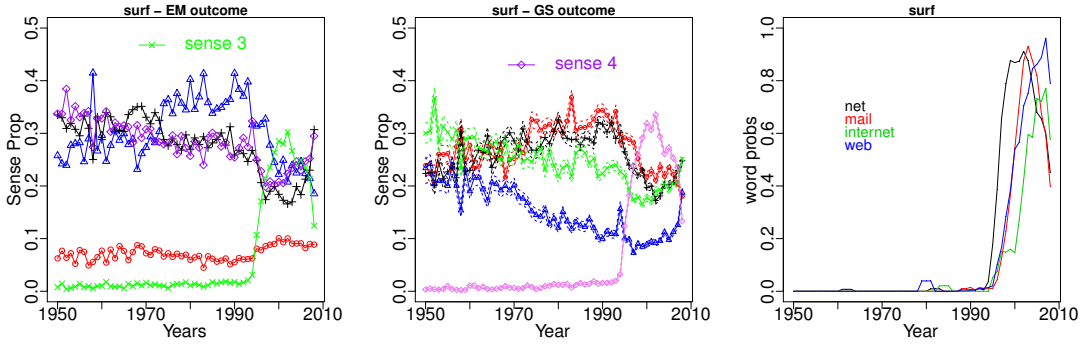


Figure 6.13 – The first and second plots show the EM and Gibbs sampling algorithm's inferred $\boldsymbol{\pi}[k]$ sense parameter outcomes for *surf/surfing/surfed* experiment, and the third plot shows the probability 'tracks' for some words that are intuitively associated with the 'exploring internet' sense of *surf/surfing/surfed*. Dating information – EM:1992, GS:1992, OED:1992, Tracks: 1993

The 'tracks' plot show tracks for *net, mail, internet, web* – words that we expect to be associated with the neologistic 'exploring inter-sense' sense of *surf/surfing/surfed*. The emergence time based on this according to the procedure in section 4.1.4 is 1993. So for *surf/surfing/surfed*, the 'EM-Date', 'GS-Date' and 'tracks' date are all very close. In this case the 'OED' first citation date is 1992 and so is also very close to other emergence dating. The 'OED' first

gist words - EM outcome	gist words - Gibbs sampler outcome
gist(sense 3) **neologism**: *_END_, Internet, ., Net, Web, net, ", or, web, Wide, World, ', for, mail, and, turf, ?, while, internet, ,, time, L, games, your, looking, -, go, e, beach, information*	gist(sense 4) **neologism**: *Web, Internet, you, ,, Net, or, net, to, can, ", web, how, _START_, ', Wide, World, learn, while, time, games, 're, are, ?, when, want, not, turf, If, mail, You*

Table 6.20 – Top 30 *gist* words for the target *surf* ranked by comparing word distributions to corpus Probabilities

Sense examples - EM outcomes	Sense examples - Gibbs sampling outcomes
L L L L *T* the Internet , play	L L you 're *T* the Internet R R
L L L L *T* the Web and check	L L L you *T* the World Wide R
L L L L *T* the Web and send	L World Wide Web *T* . R R R
L L L L *T* the Web , send	L L you 're *T* the Net R R
L L L L *T* the Internet , send	L L L L *T* " the World Wide
L L L L *T* the Web to find	L L L time *T* the World Wide R
L L L L *T* the Internet in search	L L Do you *T* ? " R R
L L L L *T* the Internet , read	L L L can *T* the World Wide R
L L L L *T* , and every accessible	L L L L *T* ' n ' turf
L L L L *T* the Web , play	L L L L *T* the World Wide Web

Table 6.21 – Top neologism sense examples for the target *surf* extracted from inferred EM and Gibbs sampling estimates for *sense 3* and *sense 4* respectively.

citation date is 1992 and in this case it is very close to the corpus emergence date.

For both cases, table 6.20 provides the top 30 'gist' words for the neologistic sense when ranked by comparing inferred $\boldsymbol{\theta}_{k=3}$ and $\boldsymbol{\theta}_{k=4}$ distributions to P_{corp} distribution (for the 'gist' words for the other senses see 6.29(b)). For most of these 'gist' words (such as *internet, world, wide, web* and *net*) they seem especially associated with the 'exploring internet' usage of the target *surf/surfing/surfed*. Further these words are unanimously found from both the inferred outcomes. The word *turf* stands out as rather unexpected in the context of a 'water-sport' usage of *surf*. The expression *surf 'n' turf* apparently refers to 'a meal combining fish with meat'. A tracks plot (in figure 6.14) for *turf* and *n* shows them to have sharp increase in probability around 1980 and this probably explains why the model has accommodated this aspect of the context of *surf* within the component that predominantly relates to the 'exploring internet' sense. Looking at the 'gist' words for the other senses in table 6.29(b)) it can be seen that some senses are very closely related (say sense 1, sense 2 and sense 3 are seemingly related to 'water sport' sense). Such relationships can be confirmed by computing a KL-divergence distance between the inferred word distributions $\boldsymbol{\theta}_k$ for all senses – work with respect to this is discussed further in section 6.7.2.

Additionally, for EM and Gibbs table 6.21 gives examples of cases whose most probable senses are $k = 3$ and $k = 4$. For EM and Gibbs, in particular for the year 2008 it gives a top 10 sense examples out of cases whose most probable senses are $k = 3$ and $k = 4$, as ranked by their probabilities to have senses $k = 3$ and $k = 4$.

6.3.8 boot

For this experiment, sub-corpora for *boot, boots, booted* and *booting* were extracted an combined. The same remarks apply as in the case of *compile* and *surf*: due to relatively small

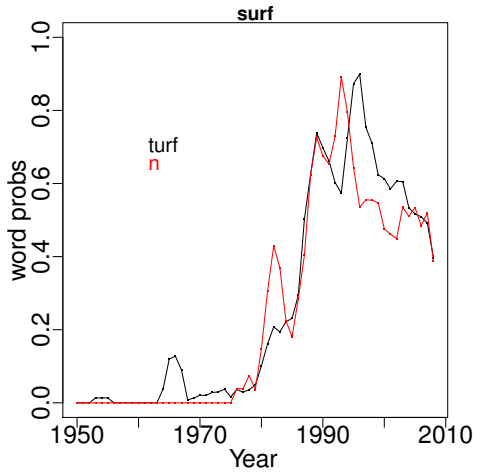


Figure 6.14 – 'Tracks' plot showing tracks for words *turf, n*

amounts of data for the separate forms, the data for several variants are combined as a work-around for the fact that the 5-gram data is not itself reduced to base forms.

In figure 6.15 the first 2 plots provide the inferred EM and Gibbs sampling estimates and also 'tracks' plot for the target *boot/boots/booting/booted*, with the algorithms run with 5 sense variants ($K = 5$). It required a 5 sense setting ($K = 5$) for both inference procedures to discover the neologistic sense representing 'computer start up' usage from the *boot/boots/booting/booted* dataset. For the EM case, the *green* line for the inferred $\boldsymbol{\pi}_t[k = 3]$ values shows a neologistic pattern and for the Gibbs case, the *black* line for the inferred $\boldsymbol{\pi}_t[k = 2]$ values shows a neologistic pattern: according to the `EmergeTime` algorithm of section 4.1.4 the emergence date are 1981 and 1980 respectively.

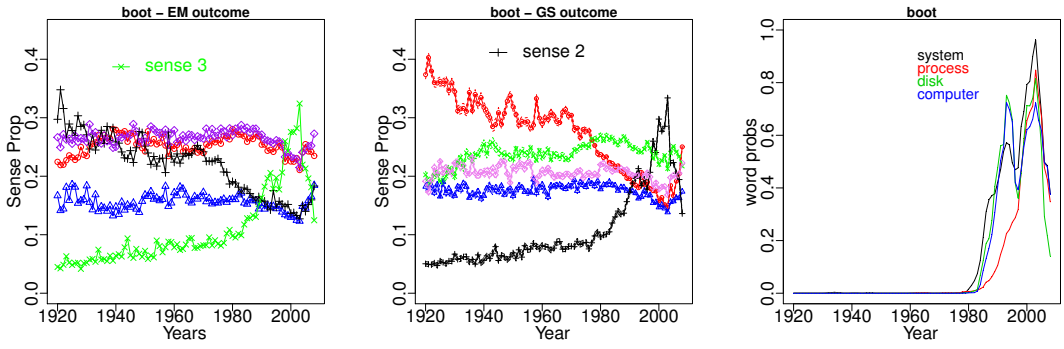


Figure 6.15 – The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *boot/boots/booted/booting* experiment, and the third plot shows the probability 'tracks' for some words that are intuitively associated with the 'computer start up' sense of *boot/boots/booted/booting*. Dating information – EM:1981, GS:1980, OED:1980, Tracks: 1982

The 'tracks' plot show tracks for *system, process, disk, computer* – words that we expect to be associated with the neologistic 'computer start-up' sense of *boot/boots/booting/booted*.

gist words - EM outcome	gist words - Gibbs sampler outcome
gist(sense 3) neologism: *system, you, time, process, camp, is, computer, ", can, to, from, received, your, /, disk, (,), sector, not, record, dual, be, when, at, during, will, up, master, as, the*	gist(sense 2) neologism: *system, you, is, time, computer, process, from, car, the, can, /, not, be, up, sector, disk, it, (, at, _START_, your, record, received, when, camp, other, to, will,), dual*

Table 6.22 – Top 30 *gist* words for the target *boot* ranked by comparing word distributions to corpus Probabilities

Sense examples - EM outcomes	Sense examples - Gibbs sampling outcomes
L directory of your *T* disk R R R	L system during the *T* process R R R
disk is used to *T* R R R R	L automatically during the *T* process R R R
L L disk 's *T* sector . R R	L operating system at *T* time R R R
L disk 's master *T* record R R R	L L When you *T* your computer R R
did not receive any *T* R R R R	L L When you *T* the system R R
L directory of the *T* disk R R R	L displayed during the *T* process R R R
L directory of the *T* device R R R	L L When you *T* the computer R R
distribution is treated as *T* R R R R	L operating system is *T* . R R R
L L disk , *T* from the R R	L L L master *T* record (MBR R
L L L disk *T* sector . _END_ R	L L L L *T* from the floppy disk

Table 6.23 – Top neologism sense examples for the target *boot* extracted from inferred EM and Gibbs sampling estimates for *sense 3* and *sense 2* respectively

The emergence time based on this according to the procedure in section 4.1.4 is 1982. For *boot/boots/booting/booted*, the 'EM-Date' and 'GS-Date' are close to the 'tracks' date and lies within $EM < 10\%$ and $GS < 10\%$ (provided in table 6.6. The 'OED' first citation date is 1980 and in this case it is very close to the corpus emergence date.

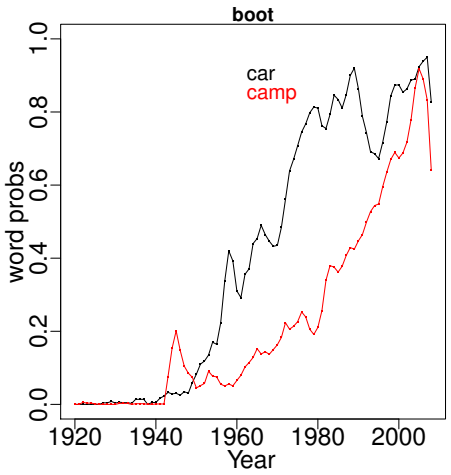


Figure 6.16 – Tracks plot for words *car, camp*

For both cases, table 6.22 provides the top 30 'gist' words for the neologistic sense when ranked by comparing inferred $\boldsymbol{\theta}_{k=3}$ and $\boldsymbol{\theta}_{k=2}$ distributions to P_{corp} distribution (for the 'gist' words for the other senses see 6.29(e)). For many of these 'gist' words (such as *system, process, time, computer, sector*) they seem especially associated with the 'computer start-up' usage of the target *boot/boots/booted/booting*. Further these words are unanimously found from both the inferred outcomes. There are words (such as *car* and *camp*) which stand out as not particularly

likely in the context of a use of *boot* in its computer-related sense. These likely stem from the expressions *car boot* and *boot camp*. A 'tracks' plot for *car* and *camp* is shown in figure 6.16 and it shows them to have increasing probability of occurrence in the context of *boot* in the second half of the 20th century, and very little probability before that. Even though this trend has a somewhat earlier starting point than that for the neologistic sense of *boot*, the absence of other inferred sense components with an increasing trend perhaps explain why the model has accommodated these aspects of the context of *boot* within the component predominantly associated with the 'computer start-up' usage.

Additionally, for EM and Gibbs table 6.23 gives examples of cases whose most probable senses are $k = 3$ and $k = 2$. For EM and Gibbs, in particular for the year 1990 it gives a top 10 sense out of cases whose most probable senses are $k = 3$ and $k = 2$, as ranked by their probability to have senses $k = 3$ and $k = 2$.

6.3.9 rock

Besides several long standing senses *rock* has according to OED a new usage referring to a 'genre of music' after the evolution of such a music form: the OED first citation is from 1956. Therefore the target *rock* was considered for a longer period ranging between 1920 and 2008 to identify the neologism sense from the dataset.

The experiments were executed starting with a 3 sense setting ($K = 3$), then with a 4-sense setting ($K = 4$) without the detection of the neologism sense. At the setting $K = 5$ a neologism sense was detected. In figure 6.17 the first 2 plots provide the inferred EM and Gibbs sampling estimates and also 'tracks' plot for the target *rock*, with the algorithms run with 5 sense variants ($K = 5$). For the EM case, the *red* line for the inferred $\boldsymbol{\pi}_t[k = 0]$ values shows a neologistic pattern and for the Gibbs case, the *blue* line for the inferred $\boldsymbol{\pi}_t[k = 1]$ values shows a neologistic pattern: in both cases according to the `EmergeTime` algorithm of section 4.1.4 the emergence date is 1960.

The 'tracks' plot show tracks for *music, concert, pop, band* – words that are expected to be associated with the neologistic 'music genre' sense of *rock*. The emergence time based on this according to the procedure in section 4.1.4 is 1967. For *rock*, the 'EM-Date' and 'GS-Date' are close to the 'tracks' date. The 'OED' first citation date is 1956 and in this case it is close to the EM and GS emergence dates.

gist words - EM outcome	**gist words - Gibbs sampler outcome**
gist(sense 0) **neologism**: *', roll, n, -, forming, music, forth, and, climbing, minerals, cut, jazz, back, –, hard, pop, place, strewn, blues, salt, drugs, tombs, bottom, caves, shelters, &, bound, prices, crystal, (*	gist(sense 1) **neologism**: *', roll, n, -, forming, and, music, forth, cut, minerals, climbing, jazz, –, hard, back, salt, bottom, place, strewn, pop, (, whole, blues, tombs, bound,), shelters, &, caves, crystal*

Table 6.24 – Top 30 *gist* words for the target *rock* ranked by comparing word distributions to corpus Probabilities

For both cases, table 6.24 provides the top 30 'gist' words for the neologistic sense when ranked by comparing inferred $\boldsymbol{\theta}_{k=0}$ and $\boldsymbol{\theta}_{k=1}$ distributions to P_{corp} distribution. For some of

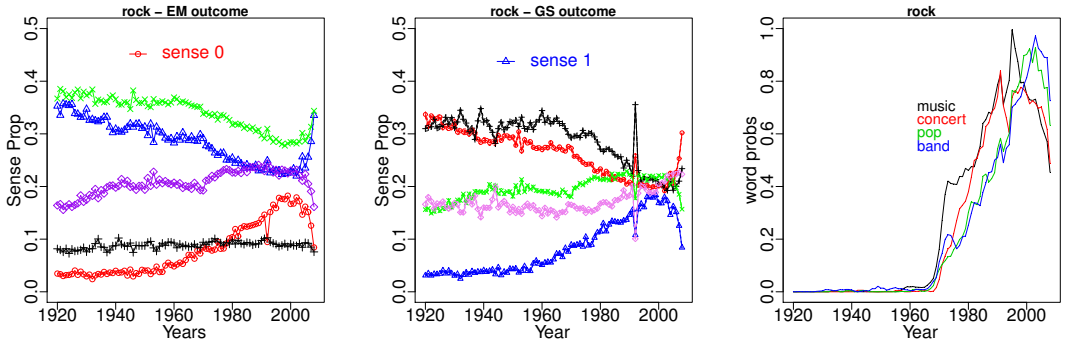


Figure 6.17 – The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *rock* experiment, and the third plot shows the probability 'tracks' for some words that are intuitively associated with the 'genre of music' sense of *rock*. Dating information – EM:1960, GS:1960, OED:1956, Tracks: 1967

Sense examples - EM outcomes	Sense examples - Gibbs sampling outcomes
L L out of *T* ' n R R	L L L L *T* ' n ' roll
L L L English *T* ' n ' R	L L L early *T* ' n roll R
L L history of *T* ' n R R	L L L L *T* Rb - Sr isochron
L L history of *T* V roll R R	L L L L *T* Rb – Sr isochron
L L L hit *T* ' n ' R	L L L L *T* Rb / Sr isochron
L L L his *T* ' n ' R	L L L L *T* ' n roll era
L L L his *T* - hard muscles R	L L L L *T* ' n ' rollers
hiking , camping , *T* R R R R	L L L L *T* ' n roll stars
L L generation of *T* ' n R R	L L L L *T* n ' roll music
L L genre of *T* ' n R R	L L L L *T* ' n roll music

Table 6.25 – Top neologism sense examples for the target *rock* extracted from inferred EM and Gibbs sampling estimates for *sense 2*

these 'gist' words (such as *music, roll, blues, pop, n*) they seem especially associated with the 'music genre' usage of the target *rock*. Further these words are unanimously found from both the inferred outcomes. However, there are some words (such as *minerals, crystal, caves*) where the association is not obvious but they are related to a different sense (say for EM case 'sense 3' and 'sense 4' seem related to 'stone' related sense). For completeness, the top 30 'gist' words for the other senses of *rock* are also provided for reference in 6.29(h).

Additionally, for EM and Gibbs table 6.25 gives examples of cases whose most probable senses are $k = 0$ and $k = 1$. For EM and Gibbs, in particular for the year 1980 it gives a top 10 sense out of cases whose most probable senses are $k = 0$ and $k = 1$, as ranked by their probability to have senses $k = 0$ and $k = 1$. It can be seen that some sense examples from the EM outcomes are not consistent with the neologistic sense 'genre of music', rather are related to 'huge stone' sense.

6.3.10 stoned

For the target *stoned*, figure 6.18 shows the outcome of an initial experiment, using the Google 5-gram books dataset (with $92k$ data items) and a 5-sense setting. The expected neologistic

sense was not identified.

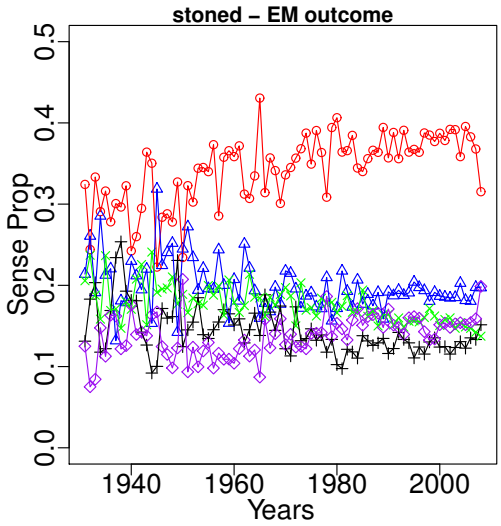


Figure 6.18 – EM inferred plot – *stoned* – 5 sense setting on Google 5-gram books dataset

The target *stoned* in the neologistic usage according to OED refers to 'drunk, extremely intoxicated' and is a slang or informal term. It was conjectured that the relative frequency of this usage would be higher in the *fiction* subset of the n-gram dataset than it is in the complete data set, and that perhaps though undetectable in the entire data set it might prove detectable in the fiction data-set. To test this a subsequent experiment was run using the fiction subset ($12k$ data items)

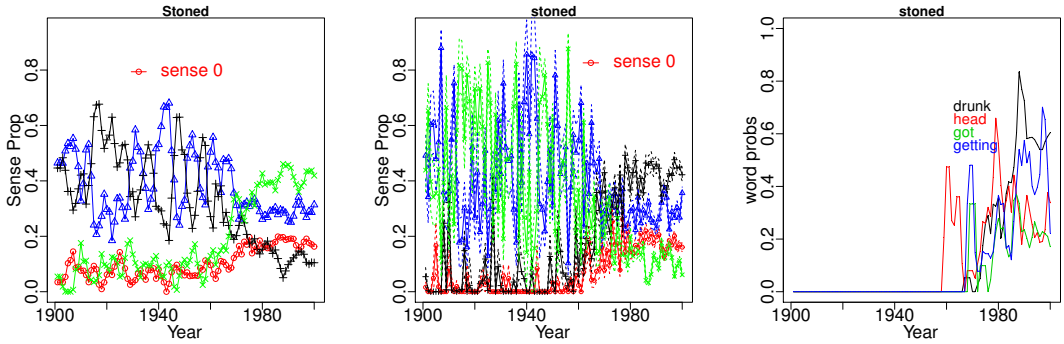


Figure 6.19 – The first and second plots show the EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *stoned* experiment, and the third plot shows the probability 'tracks' for some words that are intuitively associated with the 'under the influence of drug' sense of *stoned*. Dating information – EM:1965, GS:1960, OED:1952, Tracks: 1959

Figure 6.19 shows the outcomes of this experiment. The first 2 plots provide the inferred EM and Gibbs sampling estimates and also a 'tracks' plot for the target *stoned*, with the algorithms run with 4 sense variants ($K = 4$). In the EM case, the inferred $\boldsymbol{\pi}_t[k = 0]$ (red line) is detected as showing a neologistic patterns, with an emergence time at 1968. In the GS case,

gist words - EM outcome	gist words - Gibbs sampler outcome
gist(sense 0) **neologism**: *mind, out, his, of, time, her, minds, their, my, gourd, all, head, most, your, eyes, –, skull, mutilated, people, heads, morning, bairns, asunder, whole, at, mile, corpse, lists, the, with*	gist(sense 0) **neologism**: *out, mind, of, his, my, time, their, minds, her, gourd, head, your, most, all, patches, skull, road, newly, heads, 's, our, 're, lists, skulls, middle, quadrangle, 'm, the, eyes, at*
gist(sense 3) **neologism**: *drunk, _START_, or, I, 're, You, ", you, 'm, ?, so, Are, get, He, he, when, She, not, 's, was, said, she, both, once, getting, thrice, something, _END_, got, .*	gist(sense 2) **neologism**: *you, 're, You, ?, 'm, drunk, ", Are, _START_, 's, She, so, He, not, too, or, I, little, get, said, when, she, _END_, both, never, something, he, getting, ., kitten*

Table 6.26 – Top 30 *gist* words for the target *stoned* ranked by comparing word distributions to corpus Probabilities

Sense examples - EM outcomes *(sense 0)*	Sense examples - Gibbs outcomes *(sense 0)*
L L L L *T* out of her head	L L L L *T* out of his mind
L L L L *T* out of his gourd	L L L L *T* out of my mind
L L L L *T* out of my skull	L L L L *T* out of your mind
L L L L *T* all the time ,	L L L L *T* out of her mind
L L L L *T* out of my mind	L L L L *T* out of their minds
L L L L *T* out of his head	L L L L *T* out of his gourd
L L L L *T* , if the people	L L L L *T* out of his head
L L L L *T* out of their heads	L L L L *T* out of my head
L L L L *T* out of their minds	L L L L *T* most of the time
L L L L *T* out of my head	L L L L *T* out of our minds

Table 6.27 – Top neologism sense examples for the target *stoned* extracted from inferred EM and Gibbs sampling estimates for *sense 0*

Sense examples - EM outcomes *(sense 3)*	Sense examples - Gibbs outcomes *(sense 2)*
L she was so *T* she R R R	_START_ " You 're *T* R R R R
she was drunk or *T* R R R R	_START_ " She 's *T* R R R R
L L drunk and *T* . _END_ R R	L _START_ I 'm *T* . R R R
L L drunk , *T* , or R R	L " You 're *T* . R R R
L L drunk and *T* , and R R	was not drunk or *T* R R R R
L L drunk or *T* or both R R	L L Are you *T* ? _END_ R R
L drunk , or *T* , R R R	L L drunk or *T* or both R R
L L drunk or *T* . _END_ R R	L " You 're *T* , R R R
L either drunk or *T* . R R R	L I 'm not *T* . R R R
L L L or *T* , or both R	L when you 're *T* . R R R

Table 6.28 – Top neologism sense examples for the target *stoned* extracted from inferred EM and Gibbs sampling estimates for *sense 3* and *sense 2*.

two inferred values are detected as showing a neologistic pattern: $\boldsymbol{\pi}_t[k=0]$ (red line) with emergence time 1960 and $\boldsymbol{\pi}_t[k=2]$ (black line) with emergence time 1958. The plots in both the cases are jagged. The jaggedness in the plots can very well be attributed to the smaller number of data items available from the fiction dataset – this is further discussed in section 6.7.1 under the heading of *ablation* tests .

The 'tracks' plot show tracks for *drunk, head, got, getting* – words that are intuitively associated with the neologistic 'under the influence of drug' sense of *stoned*, and the tracks-based emergence date using these words is 1959. This is close to the emergence dates found by EM and Gibbs sampling. The 'OED' first citation date is 1952 and in this case 7 years earlier.

Table 6.26 provides in the first row the top 30 'gist' words for sense 0 as obtained via EM and GS, when ranked by comparing inferred $\boldsymbol{\theta}_{k=0}$ distribution to P_{corp} distribution. The second row gives the 'gist' words for sense 2 as obtained by GS. It also gives the 'gist' words for sense 3 as obtained by EM, which seems of the other senses the closest to also showing

a neologistic pattern (For completeness, the top 30 'gist' words for the other senses of *stoned* are also provided for reference in table 6.29(i)). For some of these (such as *mind, out, head, his* and *her*) they are intuitively especially associated with the 'under the influence of drug' usage of the target *stoned*, though there are others which are unexpected (such as *mutiliated, corpse, list, quadrangle* and *kitten*). On the evidence of the 'gist' words, the sense 0 component inferred by GS seems similar to the sense 0 component inferred by EM, and also the sense 2 GS component seems similar to the sense 3 EM component.

Additionally, for EM and Gibbs cases, tables 6.27, 6.28 show examples from the same pairs of senses. For the senses which were identified as showing a neologistic pattern the examples seem mostly consistent with the neologistic sense 'under the influence of drug'. Also the examples reinforce the impression of the similarity of the sense 0 GS and sense 0 EM components, and of the sense 2 GS component and sense 3 EM component.

Overall the outcomes for this experiment on the fiction subset do suggest that the algorithms are able to detect the anticipated neologistic sense in this subset, though with less clarity than was the case with the other targets. For example, the GS procedure seems to have identified two components both arguable corresponding to the anticipated sense, and the inferred time-lines seem considerably less smooth.

6.4 Non-neologism targets

The experiments discussed in the previous section 6.3 concerned neologism targets, so words concerning which it was known that during a particular time period a new sense emerged. It was argued that the experiments provide good evidence that the 'diachronic model' was able to identify these instances of sense emergence. This section will complement this with experiments on *non*-neologism targets, so words which, for a certain time period it is known that *no* new sense emerged. For these words the desired behaviour is that the model does *not* identify a sense emerging within the time period looked at.

The targets which were used for this purpose are given in table 6.30. The 'Years' column in the table provides the year span that was considered for the EM and Gibbs sampling experiments and 'Lines' column is the number of data items (5-grams) with target T.

In the first instance these targets were chosen based on native speaker intuition. *ostensible* and *cinema* were chosen in the expectation that these have just one sense, and that remained so throughout the indicated time period. The targets *present, promotion, theatre, play, plant, spirit* where chosen in the expectation that these have multiple senses, and that none of these senses emerged in the indicated time period.

As with the neologism targets, ground-truth concerning *non*-neologism is a somewhat difficult issue: there is no gold-standard reference set from which to select *non*-neologism targets. For the above-mentioned targets we made use of the OED to try to confirm absence of sense emergence for the time periods chosen. Concerning *ostensible* this confirms a single sense '*Declared, avowed, professed; presented (esp. untruthfully or misleadingly) as actual*'. For

gist words - EM outcome	gist words - Gibbs sampler outcome
(a) mouse - outcomes	
gist(sense 0): *cat, rat, a, ", as, keyboard, anti, game, -, ,, like, such, rabbit, In, little, or, and, not, have, clicks, IgG, was, field, but, than, quiet, ', few, white, –*	gist(sense 0): *in, cells, of, anti, embryo, mammary, model, embryos, -, brain, human, (, cell, house, tumor, /, development, virus, gene, from, _END_, :, bone, IgG, Mus, skin, marrow, embryonic, ., adult*
gist(sense 2): *cells, in, embryo, mammary, embryos, brain, cell, of, tumor, model, development, virus, /, house, gene, bone, :, Mus, _END_, marrow, skin, (, from, embryonic, adult, ., during, early, musculus, 2*	gist(sense 2): *cat, rat, a, ", keyboard, as, game, -, in, like, human, such, ,, In, of, little, model, for, rabbit, than, field, A, have, ', but, white, quiet, and, clicks, not*
(b) surf - outcomes	
gist(sense 0): *_START_, was, The, is, hear, could, high, heavy, that, be, so, zone, a, I, no, up, too, not, there, in, If, when, running, In, When, but, 's, There, R, where*	gist(sense 0): *_END_, zone, in, ., into, through, by, ", ;, for, clam, from, wind, Spisula, beyond, above, body, swimming, (, fishing, sun, outside, channel, world, solidissima, away, gentle, toward, inner, R*
gist(sense 1): *zone, in, of, into, through, pounding, heavy, line, is, ;, by, a, crashing, breaking, at, beyond, ace, up, from, outside, waves, _END_, R, below, high, above, rolling, ocean, wind, .*	gist(sense 1): *on, shore, beach, upon, ,, against, mail, -, that, breaking, beat, rocks, sea, beating, coast, e, white, reef, which, along, beaten, was, ace, its, The, a, turf, great, breaks, their*
gist(sense 2): *of, on, in, breaking, zone, a, swimming, which, through, upon, against, wind, shore, was, beach, heavy, into, white, where, with, it, along, rocks, diving, pounding, as, sea, -, beat, clam*	gist(sense 2): *right, ,, _START_, was, as, The, but, which, where, it, when, and, swimming, diving, is, so, water, skiing, with, I, sailing, then, that, there, or, scuba, they, too, not, sun, transport*
gist(sense 4): *of, sound, in, into, out, roar, through, pounding, edge, R, thunder, line, The, white, breaking, noise, sun, down, waded, by, boom, a, distant, crash, best, heavy, crashing, up, like, deep*	gist(sense 3): *sound, roar, out, ,, edge, The, thunder, of, could, hear, on, noise, to, waded, boom, crash, be, white, sounds, listening, I, he, plunged, _START_, down, beach, R, run, can, distant*
(c) gay - outcomes	
gist(sense 0): *", _END_, ;, ?, happy, life, world, man, ., but, lively, bar, so, light, flowers, a, ', bright, very, young, !, cheerful, as, with, colors, good, his, little, scene, full*	gist(sense 0): *the, of, with, and, world, light, grave, from, bright, happy, in, life, lively, flowers, young, their, ;, ,, colors, cheerful, full, hearted, his, company, little, laugh, by, brilliant, ladies, And*
gist(sense 1): *_START_, was, I, he, be, It, that, He, not, you, out, to, a, grave, 'm, is, R, She, The, they, had, because, were, been, if, 're, it, very, know, openly*	gist(sense 1): *", I, not, he, you, ', be, 'm, man, ?, lesbian, 're, who, being, _START_, was, they, It, that, 's, are, He, or, is, openly, it, as, am, have, she*
(d) bit - outcomes	
gist(sense 0): *_END_, lip, ., her, than, his, as, L, time, information, much, money, ?, but, luck, more, too, on, ", of, every, my, lower, fun, then, from, paper, work, into, complicated*	gist(sense 0): *_END_, ., lip, her, than, his, time, of, more, information, ", but, on, too, money, ?, in, my, from, luck, and, then, ,, at, into, lower, work, fun, paper, off*
gist(sense 1): *not, was, 's, It, be, have, just, I, had, been, _START_, like, 'm, quite, may, feel, do, there, R, it, you, up, little, were, He, to, There, 're, felt, Not*	gist(sense 1): *was, not, 's, be, It, I, have, just, _START_, had, like, quite, been, it, little, to, 'm, may, do, you, feel, up, a, there, is, He, were, me, Not, 're*
(e) boot - outcomes	
gist(sense 0): *_END_, ., -, high, ;, league, seven, heeled, knee, leather, toed, black, lace, their, straps, into, ", riding, her, own, go, nailed, tucked, ', with, under, steel, length, heavy, polished*	gist(sense 0): *shoes, shoe, ,, and, he, which, were, but, spurred, gloves, clothing, industry, as, spurs, hats, had, said, so, then, that, all, they, a, she, hat, or, I, made, was, clothes*
gist(sense 1): *he, were, which, was, car, on, but, had, that, other, I, it, then, feet, His, the, been, so, said, foot, floor, as, out, they, she, them, made, went, one, ground*	gist(sense 1): *pair, of, toe, heel, a, sound, soles, out, the, new, toes, sole, pairs, one, old, heels, down, 's, man, Italian, tops, good, at, knife, top, an, two, The, looked, tip*
gist(sense 2): *shoe, shoes, ,, and, jeans, gloves, hat, cowboy, spurred, breeches, hats, black, leather, clothing, industry, high, spurs, clothes, shirt, trousers, coat, top, or, rubber, -, a, pants, riding, jack, blue*	gist(sense 3): *_END_, ., ", ;, league, seven, ?, !, ', go, feet, their, camp, 's, her, under, my, black, -, hiking, to, with, over, leather, his, riding, heavy, combat, straps, own*
gist(sense 4): *pair, off, toe, of, pulled, put, took, take, his, He, heel, my, sound, a, her, pull, soles, down, on, their, toes, snow, sole, one, new, out, pulling, big, the, I*	gist(sense 4): *off, high, pulled, put, took, -, take, heeled, He, knee, pull, my, on, his, jeans, their, with, leather, patent, big, toed, pulling, cowboy, lace, top, coat, I, black, tucked, her*

Table 6.29 – For (a-e) provides the top 30 *gist* words of non-neologism senses for the targets *mouse, surf, gay, bit, boot* – further continued in next page for other targets

cinema two related senses are distinguished, one for the building where films are shown, and one for cinema films considered collectively, especially as an art form. Based on their first

gist words - EM outcome	gist words - Gibbs sampler outcome
(f) compile - outcomes	
gist(sense 0): *list, this, complete, In, run, a, history, to, of, book, index, program, your, execute, all, To, bibliography, comprehensive, an, inventory, information, lists, them, source, link, dictionary, such, application, ., _END_*	gist(sense 0): *list, book, of, index, complete, history, a, bibliography, this, on, report, all, comprehensive, lists, an, information, In, data, record, volume, table, dictionary, inventory, such, catalogue, collection, exhaustive, statistics, own, work*
gist(sense 1): *used, you, able, to, possible, task, process, order, difficult, need, attempt, made, will, responsible, impossible, If, would, try, necessary, assistance, When, help, In, it, not, want, You, purpose, easy, began*	gist(sense 1): *used, been, was, able, The, have, possible, be, process, task, is, were, order, _START_, you, need, he, in, to, difficult, made, attempt, first, has, I, for, responsible, would, purpose, try*
(g) strike - outcomes	
gist(sense 0): *blow, balance, between, into, at, terror, with, as, his, him, up, out, their, root, _END_, heart, chord, reader, back, bargain, her, ?, them, conversation, own, us, very, ;, decisive, me*	gist(sense 0): *not, did, right, does, he, it, I, to, _START_, be, you, do, that, will, would, they, motion, It, ready, if, me, have, is, we, can, about, able, could, may, seemed*
gist(sense 2): *not, did, right, does, he, it, I, _START_, to, be, you, that, do, will, would, they, motion, It, ready, if, me, have, is, we, can, about, able, could, may, He*	gist(sense 1): *blow, balance, between, into, at, terror, with, as, out, his, him, up, _END_, their, root, heart, chord, reader, back, bargain, ?, conversation, her, own, them, ;, very, us, any, decisive*
(h) rock - outcomes	
gist(sense 1): *which, it, but, he, ,, at, where, they, been, had, that, so, or, has, be, his, L, with, can, have, by, sand, we, middle, is, then, tree, side, one, may*	gist(sense 0): *which, it, ,, but, my, that, he, I, at, will, be, they, been, where, is, build, had, has, so, or, can, by, have, in, side, with, the, his, we, one*
gist(sense 2): *_START_, The, not, I, will, this, is, my, build, grained, A, whole, This, ", boat, was, He, do, It, began, want, fine, Sr, be, If, 's, On, There, upon, In*	gist(sense 2): *of, out, The, _START_, top, cut, mass, the, piece, in, types, bed, an, type, igneous, solid, wall, cleft, masses, ledge, grained, carved, face, hewn, edge, part, great, surface, living, country*
gist(sense 3): *out, down, of, top, rock, R, up, large, sat, from, on, flat, sitting, piece, solid, between, against, carved, cleft, great, ledge, firm, an, cut, foot, a, built, part, head, living*	gist(sense 3): *_END_, ., ;, ", ?, garden, boat, art,), formations, face, solid, wall, !, music, into, bottom, star, for, band, surrounding, walls, gardens, mass, sedimentary, surface, volcanic, concert, bare, or*
gist(sense 4): *_END_, ., ;, ", ?, art, garden,), boat, (, formations, !, star, types, surrounding, music, band, salt, face, wall, concert, bottom, solid, L, volcanic, sedimentary, gardens, walls, for, mass*	gist(sense 4): *on, rock, to, as, a, like, down, upon, up, _START_, sat, against, not, between, was, flat, large, sitting, from, I, He, began, this, struck, behind, firm, his, head, built, big*
(i) stoned - outcomes	
gist(sense 1): *death, to, be, by, being, must, they, have, would, notice, been, care, hardly, until, a, eyeballs, deserve, in, ;, ., for, were, gills, will, Stephen, risk, as, little, should, _END_*	gist(sense 1): *too, me, deserving, notice, care, eyeballs, deserve, been, we, death, heels, to, ., _END_, have, whole, centre, gills, ', asunder, would, sawn, eyes, move, actually, one, had, Saint, 'd, be*
gist(sense 2): *again, up, him, speakers, woman, boys, stones, upon, patches, newly, with, :, yet, cobble, see, away, maybe, from, one, but, had, is, who, camp, the, streets, an, never, home, another*	gist(sense 3): *again, up, until, see, speakers, yet, maybe, away, camp, with, heels, from, repeatedly, had, eyes, upon, woman, home, –, mile, ordinary, ', hardly, their, helms, Christians, men, what, way, been*
(j) paste - outcomes	
gist(sense 0): *tomato, 1, on, 2, tablespoons, tablespoon, Add, tomatoes, it, them, (,), sauce, be, applied, another, teaspoon, contents, cup, your, salt, ,, curry, sesame, The, in, cook, Stir, puree, wine*	gist(sense 0): *1, on, it, tablespoons, 2, tomato, tablespoon, up, them,), (, be, applied, your, contents, another, teaspoon, into, cup, sesame, salt, used, the, in, text, The, ,, sauce, is, cook*
gist(sense 1): *smooth, make, form, a, with, thick, made, cement, of, _END_, ., porcelain, thin, water, soft, stiff, fine, mixed, flour, hardened, by, hard, ;, up, consistency, forms, baking, little, like, Make*	gist(sense 2): *smooth, a, make, form, thick, with, made, of, porcelain, cement, thin, ., _END_, water, soft, stiff, fine, mixed, hardened, hard, by, consistency, to, flour, forms, little, Make, like, ;, until*

Table 6.29 – continued from previous page – For EM and Gibbs sampling word distributions $\boldsymbol{\theta}_k$, the top 30 *gist* words of non-neologism senses for targets *mouse, surf, gay, bit, boot, compile, strike, rock, stoned, paste* (a-j) ranked by comparing word distributions to corpus Probabilities.

citation dates (1909, 1919), for the period chosen (1950–2008), it seems these senses were in existence prior to the period chosen. For the other targets a similar process was followed considering the senses identified in the OED and their first citation dates. For the targets which were anticipated to be ambiguous, the OED confirms them to be so and with the anticipated

Target	Years	Lines	Target	Years	Lines
ostensible	1800-2008	130k	theatre	1950-2008	1125k
present	1850-2008	56333k	play	1950-2008	13726k
cinema	1950-2008	305k	plant	1900-2008	8175k
promotion	1930-2008	1681k	spirit	1930-2008	11573k

Table 6.30 – Google 5 gram dataset - the table has the targets for non-neologisms

senses having an emergence pre-dating the time periods chosen, often drastically so. For example, *play* was chosen as a non-neologism target, with time period 1950–2008. Among the OED's noted senses for *play* are: '*a theatrical performance*', '*engage in fun*', '*take part in a sport*', '*to perform on a musical instrument*', all with citation date indicating them to have been in use for many centuries. Being a dictionary that aims to be as comprehensive as possible[6], its sub-division of a word into separate senses is probably more fine-grained than any other dictionary. Given this it is not surprising that there are targets from Table 6.30 such that in the OED's long enumeration of senses there do occur senses with first citation dates indicating an emergence within the time period chosen. For example, for *play* it notes a sense '*The control on a tape player, video recorder, etc., used to initiate playing.*' with first citation date of 1978. It seems a safe assumption that this sense is considerably less frequent than the earlier noted senses, though this assumption has to rest on intuition, as the OED gives no sense frequency statistics. Modulo this caveat concerning some likely to be infreqent senses, consultation of the OED seems to confirm that the targets given in Table 6.30 can serve as examples of words which do not exhibit sense emergence in the time periods indicated.

There was an initial testing of the non-neologism targets, with the number of senses set to 2 or 3. This setting was not very systematic but roughly followed an intuition of assigning 3 to those anticipated to be more ambiguous and 2 to those thought to be less ambiguous. As this initial test was somewhat unsystematic a subsequent round of tests was done, this time setting the number of senses to 5 for all targets, which one would expect to increase the possibility of inferring a sense emergence.

Figure 6.20 shows the outcomes from the initial tests. For each target and for each its senses, the time-line via EM and GS of the sense probabilities $\boldsymbol{\pi}_t[k]$ was assessed using the emergence time detection method which was discussed in section 4.1.4. Recall that this returns an empty set of times if no emergence time is detected. For all of the 8 non-neologism targets an empty set is returned for all senses, indicating no inference of a sense emergence.

Figure 6.21 shows the outcomes K was set to 5 for each target. Again for each of the 5 senses, the time-line of the sense probabilities $\boldsymbol{\pi}_t[k]$ was assessed using the emergence time detection method. In this case for 7 out of the 8 non-neologism targets. However, for the target *promotion* this did not happen. For the EM outcomes sense 3 returns an emergence time of 1974, and sense 4 returns an emergence time of 1941, neither of which was an expected result. For the GS outcomes, only sense 4 returns an emergence time of 1948 and is still not an expected result as with the EM outcomes. Tables 6.31 and 6.32 gives the *gist* words for these

[6]Its printed edition runs to 20 volumes

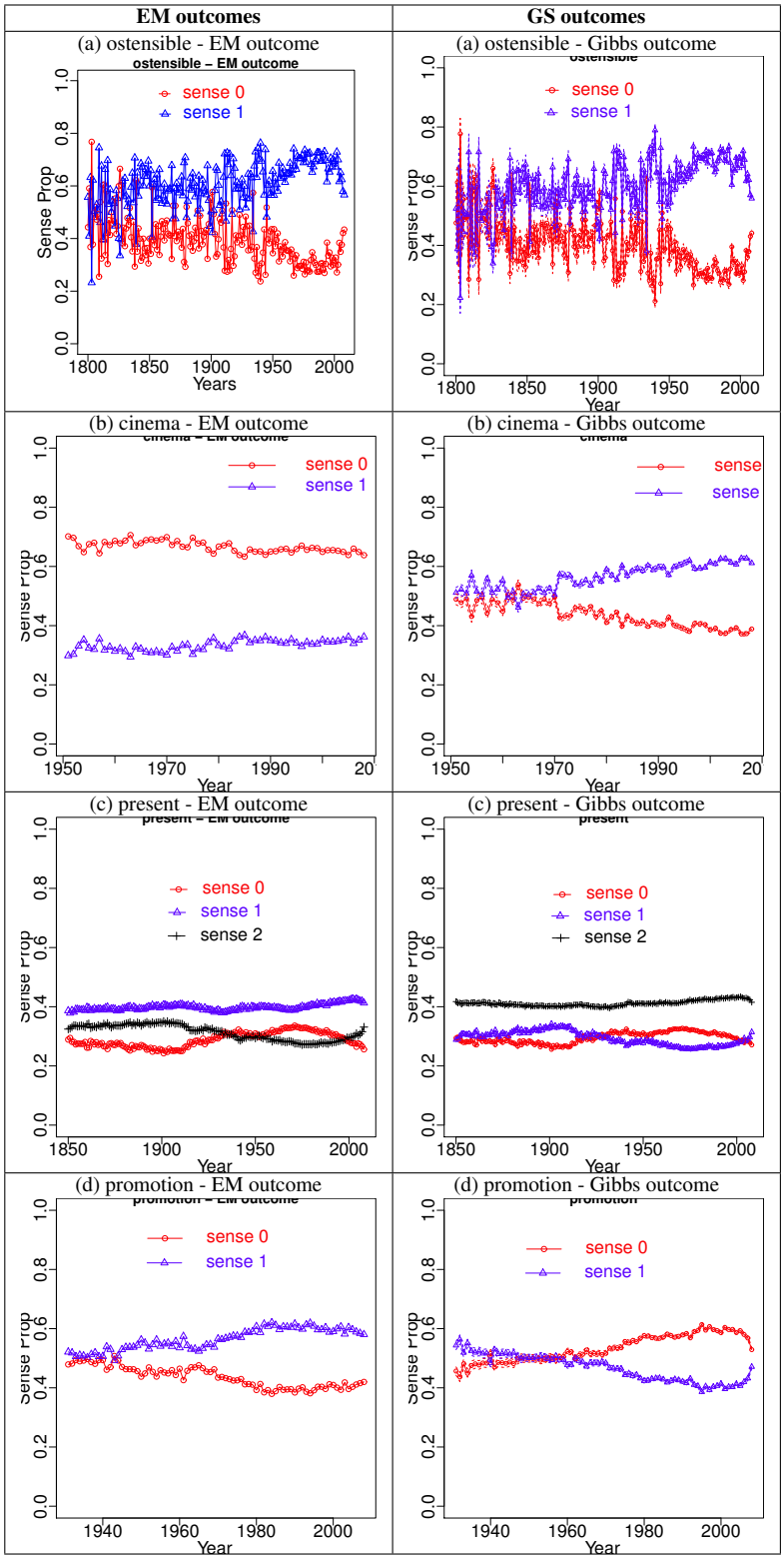


Figure 6.20 – For plots (a-d), the first and second columns show the outcomes of EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *ostensible, cinema, present, promotion* non-neologism targets – (e-h) continued in the next page.

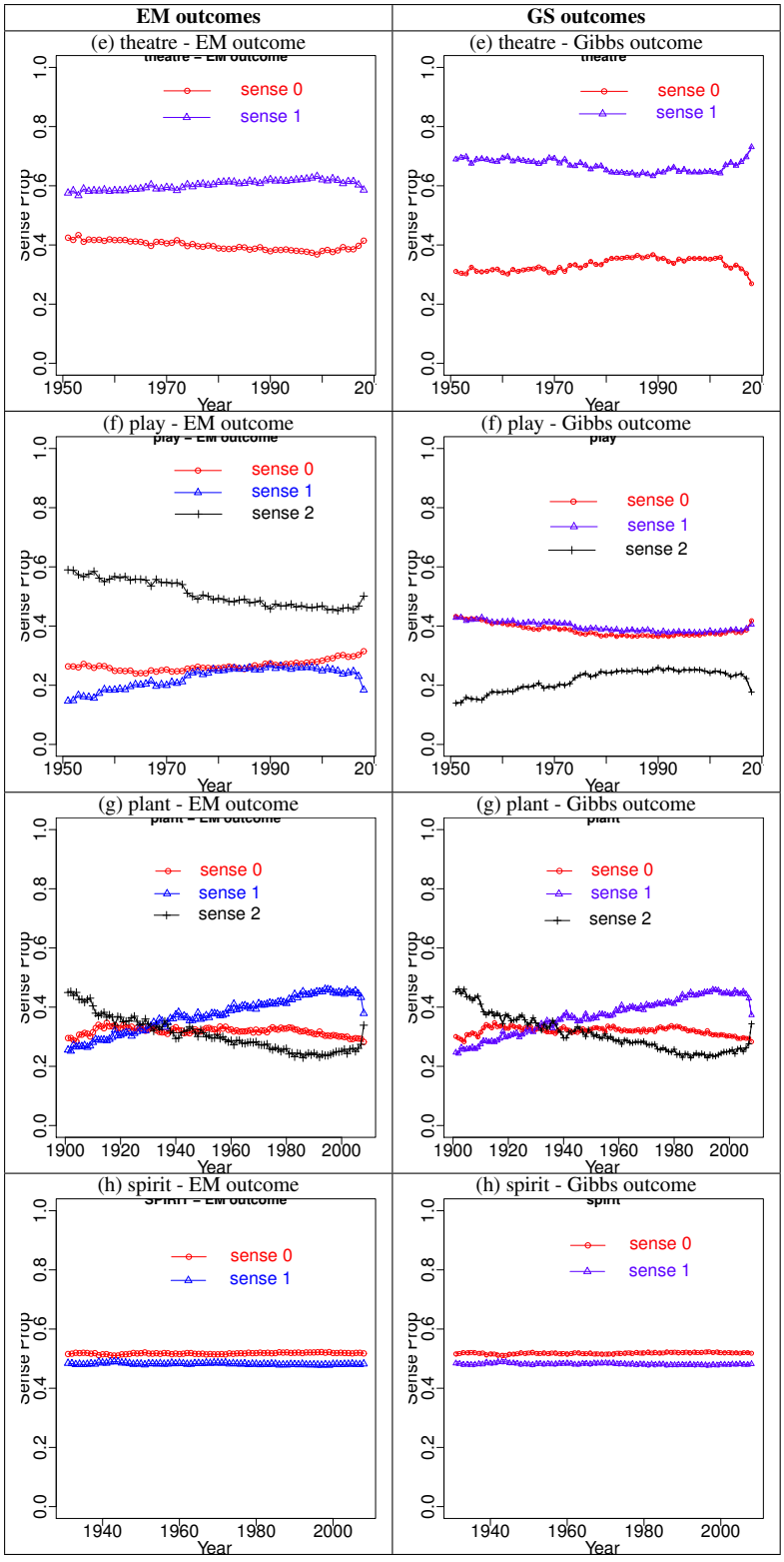


Figure 6.20 – For plots (e-h), the first and second columns show the outcomes of EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *theatre, play, plant, spirit* non-neologism targets.

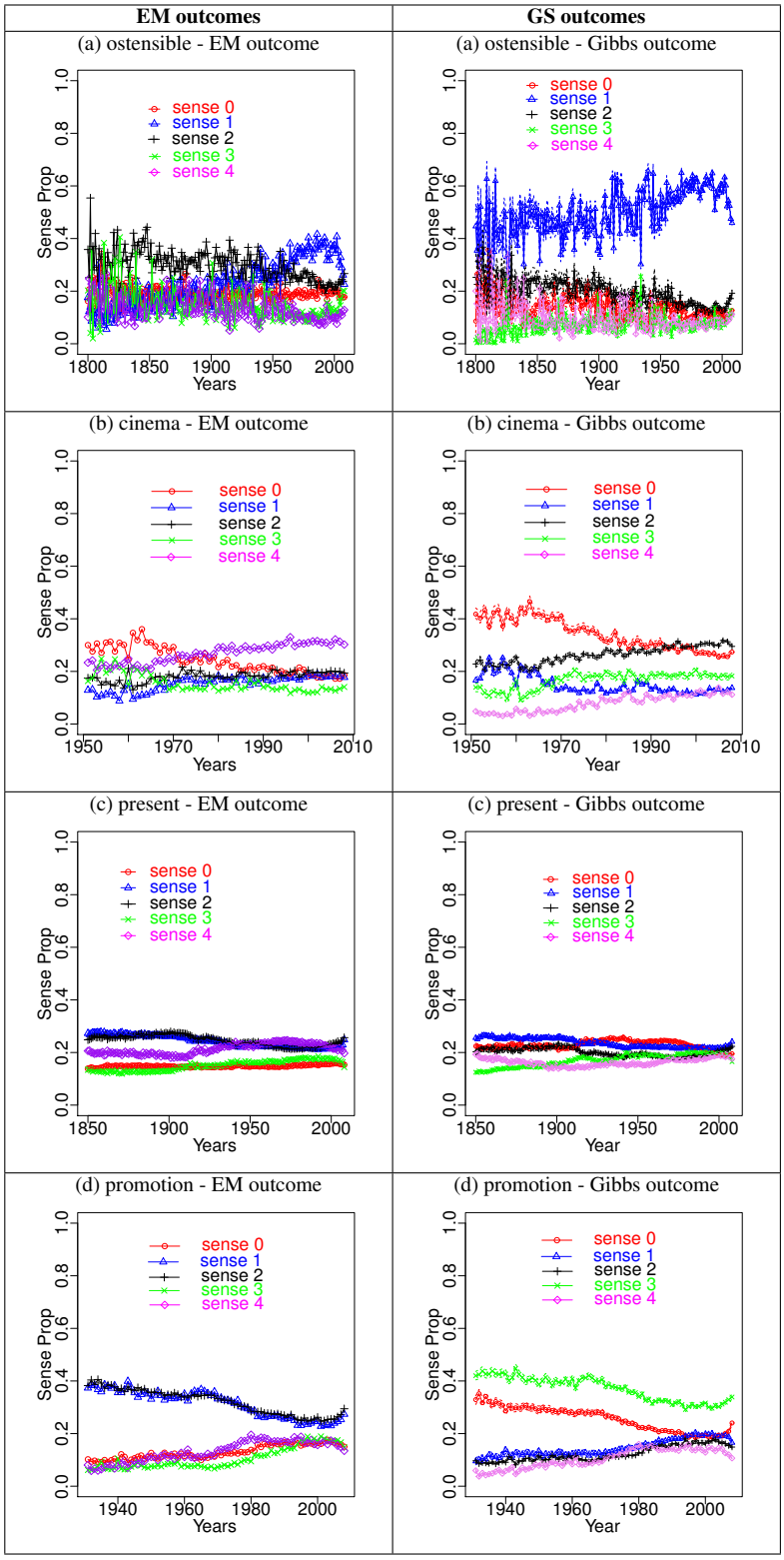


Figure 6.21 – For plots (a-d), the first and second columns show the outcomes of EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *ostensible, cinema, present, promotion* non-neologism targets – (e-h) continued in the next page.

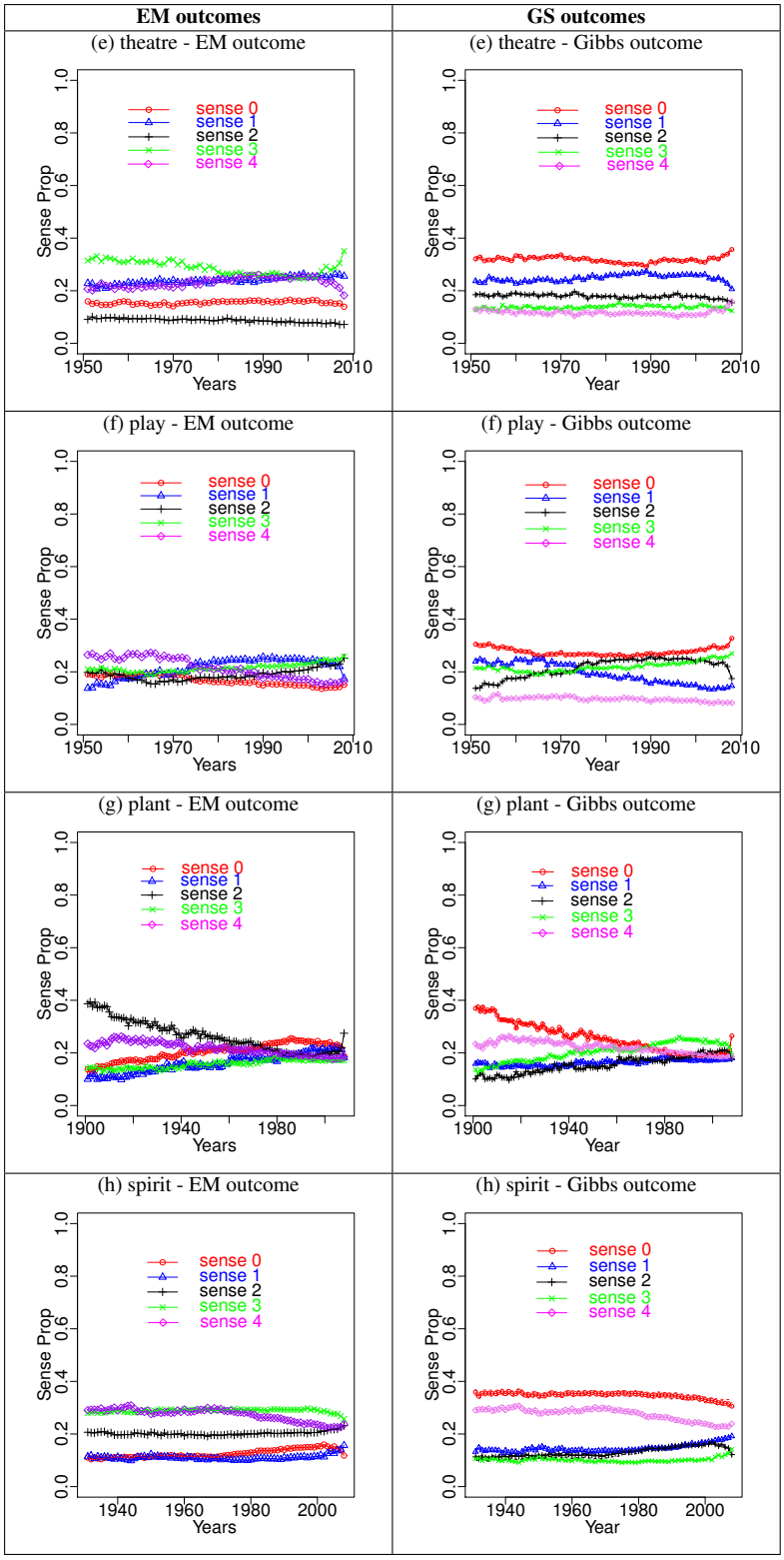


Figure 6.21 – For plots (e-h), the first and second columns show the outcomes of EM and Gibbs sampling algorithm's inferred $\pi_t[k]$ sense parameter outcomes for *theatre, play, plant, spirit* non-neologism targets.

senses.

gist(sense 3): *prevention, disease, protection, health, human, rights, maintenance, education, Health, and, mental, , public, programs, illness, preservation, formation, care, development, sale, planning, new, services, encouragement, National, family, democracy, practice, control, implementation*	gist(sense 4) neologism: *advertising, ,, sales, hiring,), distribution, (, transfer, training, or, over, passed, price, pay, pricing, tenure, place, appointment, job, marketing, recruitment, he, product, publicity, selling, employment, salary, selection, raise, demotion*

Table 6.31 – For the EM outcomes, top 30 *gist* words for the target *promotion* ranked by comparing word distributions to corpus Probabilities are shown

gist(sense 4): *sales, distribution, advertising, transfer, training, or, price, hiring, tenure, pricing, pay, place, job, marketing, publicity, product, selling, creating, salary, substitution, import, raise, demotion, appointment, recruitment, employment, relations, firing*

Table 6.32 – For the GS outcomes, top 30 *gist* words for the target *promotion* ranked by comparing word distributions to corpus Probabilities are shown

The OED identifies one sense of *promotion* as 'action of helping forward' and the gist words for sense 3 (EM) seem mostly consistent with that sense. It also identifies the more specific sense of 'publicizing of a product'. The gist words for sense 4 (EM and Gibbs) are arguably more closely related to that sense than those for sense 3. The OED's first citation date for the 'action of helping forward' sense is 1425, and for the 'the publicizing of a product' sense it is 1914. So concerning the inferred sense 3, the OED has nothing suggesting a *language* change which would be consistent with the trend found for sense 3 with an emergence time of 1974. Perhaps the cause of the trend detected is not language change but changes in opinion or changes in the world, concerning what things are subject to the 'action of helping forward' – in section 6.6 some further cases of this kind are discussed. For sense 4 it is perhaps possible that the first citation date of 1914 for the publicity related sense is substantially earlier than its genuine emergence in the corpus, but we have no ready means to confirm such a speculation.

6.5 Discriminating neologism vs non-neologism targets

As a further test, a scoring mechanism introduced by Lau et al. [2012] called 'Novelty score' is adapted for this thesis work in ranking the neologism targets. The novelty score introduced by Lau et al. [2012] is given by the novelty ratio in equation 6.1.

$$\text{Novelty}_{\text{Ratio}} = \frac{P_f(s)}{P_r(s)} \tag{6.1}$$

In equation 6.1, $P_f(s)$ and $P_r(s)$ are the inferred proportion of usages of a target T corresponding to sense s in the focus corpus and reference corpus, respectively – these corpora come from two different times $\mathcal{E}_2$ and $\mathcal{E}_1$. For a target, the maximum score computed from all the inferred senses is the 'novelty score' assigned to the target. They compute this score for all the targets considered for experiments and rank the targets based on the score in descending order and from this, the actual 'neologism' targets are expected to appear in the top of the table. As

we did, they had both positive and negative targets. Without a time-line their evaluation cannot be a comparison of true and inferred emergence date and instead they count success as a tendency to place positives above negatives when ranked by the 'novelty' score provided in equation 6.1. They obtain thus a ranking on their targets: { **domain**(116.2), **worm**(68.4), **mirror**(38.4), *guess*(16.5), **export**(13.8), *founder*(11.0), *cinema*(9.7), **poster**(7.9), *racism*(2.4), *symptom*(2.1) } (with positive targets in bold and negative in italics).

target	t_{min}	t_{max}	**sense**	**score**
strike	1803	1982	1	$4.69865630943e+113$
gay	1943	1997	2	$9.14862673074e+69$
mouse	1957	1995	1	1377.22470581
surf	1952	2002	3	177.243962439
compile	1954	2000	2	67.4743999785
stoned	1941	1978	3	49.996
bit	1933	1992	2	14.0752091571
boot	1921	2003	3	11.7516786517
paste	1955	2002	2	8.39170393601
rock	1932	1999	0	7.56358762852
ostensible	1815	1986	1	5.14868205866
plant	1902	1994	1	1.82944466751
play	1951	1990	1	1.82368869791
present	1901	1971	0	1.38457323655
promotion	1943	1984	1	1.25953397144
cinema	1963	1981	1	1.24805948421
theatre	1953	1999	1	1.11539934388
spirit	1943	1997	0	1.02329509604

Table 6.33 – Novelty scores for the neologism and non-neologism targets based on the EM inferred $\pi_t[k]$ sense parameter outcomes.

target	t_{min}	t_{max}	**sense**	**score**
stoned	1928	1978	2	103993.451627
strike	1825	1984	2	5442.70895335
gay	1946	1997	2	2791.15008492
mouse	1959	1995	1	1485.97820555
surf	1958	2002	4	156.734009886
compile	1954	2002	2	26.6235467848
bit	1933	1992	2	10.0206427976
paste	1951	2003	1	7.50601173275
rock	1932	1999	1	7.40120508635
boot	1924	2003	2	7.05613461148
ostensible	1803	1940	1	3.53971281131
plant	1902	1994	1	1.89215339136
play	1951	1990	1	1.86010084371
promotion	1932	1995	1	1.40545596937
cinema	1963	2006	1	1.36194139893
theatre	1989	2008	1	1.15514153993
spirit	1943	1999	0	1.0233877599

Table 6.34 – Novelty scores for the neologism and non-neologism targets based on the Gibbs sampling inferred $\pi_t[k]$ sense parameter outcomes.

As they have considered just two times – a focus and reference corpus from later and earlier times respectively, but for the current work a long time period is considered, therefore to compute such novelty scores, a slightly modified version of the Novelty score shown in equation 6.1 is followed. To compute this, consider t_{min} and t_{max} to be the values of time t

where the inferred sense parameter $\pi_t[k]$ is the 'minimum' and 'maximum' values for the sense k. Then the novelty ratio is defined to be:

$$\text{Novelty}_{\text{Ratio}} = \begin{cases} \frac{t_{max}}{t_{min}} \text{ (if } t_{max} \text{ is later than } t_{min}) \\ \frac{t_{min}}{t_{max}} \text{ (if } t_{min} \text{ is later than } t_{max}) \end{cases} \tag{6.2}$$

From the computed $\text{Novelty}_{\text{Ratio}}$ for all senses K of target T, the highest score is considered to be the 'Novelty' score for T – this way, the novel sense is also identified.

Tables 6.33 and 6.34 provides the list of targets considered for the 'actual' neologism and non-neologism EM and Gibbs experiments in which they are ranked in descending order based on the novelty scores computed from the EM and Gibbs inferred $\pi_t[k]$ outcomes. The second and third columns shows the t_{min} and t_{max} ie., the earlier and later dates, considered for computing the $\text{Novelty}_{\text{Ratio}}$. Further, it can be seen from the tables all the 'actual' neologism targets are on the top of the table, while the non-neologisms in the bottom (ranked based on the novelty score in descending order). It can be noticed from the ranked outcomes that the proposed new metric in equation 6.2 works better than [Lau et al., 2012]'s proposal. This way, the current thesis work in identifying the neologistic sense further confirms with other testing standards.

6.6 Neologisms which were undetected

Until now the success of the 'expected' outcomes for a set of neologism and non-neologism targets were discussed in sections 6.3 and 6.4. In this section the unsuccessful attempts to find 'neologism' sense from targets which are expected to have an emerging sense is reported. Table 6.35 show details for the targets[7] *hip, export, mirror, domain, high* which are expected to have a neologism sense, where the 'Exp. new sense' column refers to the expected neologism usage of the corresponding target and the other columns are similar the columns provided in table 6.5.

Target	Years	Lines	Min Occs	Max Occs	Exp. new sense	Vocab size
hip	1851-2008	814k	734k	3673k	trendy (or) stylish	2696
export	1970-2008	1415k	1193k	5967k	convert file format	4323
mirror	1970-2008	1444k	1710k	8554k	store copies of data	4874
domain	1970-2008	1586k	1989k	9949k	suffix of internet address	7239
high	1930-2008	43764k	40551k	202755k	drunk	34937

Table 6.35 – Google 5 gram dataset - the table provides the information for targets that are neologisms

EM experiments were conducted on the targets *hip, export, mirror, domain, high* expecting a neologism sense, the details of the experiments are presented in the following sections.

[7]Explicit dictionary definitions and citation informaton from the online Oxford English Dictionary (OED) for the chosen targets are provided in the appendix A.1

6.6.1 hip

With the OED date for *hip* in 'stylish' sense found to be in 1904, the dataset between years 1851 and 2008 was considered in an attempt to identify the neologistic sense using the EM inference procedure. The EM experiments were conducted on 3, 4 and 5 sense settings. With a 3-sense setting the neologism sense was not discovered, however with 4 and 5 sense settings the EM seems to have discovered an emerging sense – this is provided in figure 6.22.

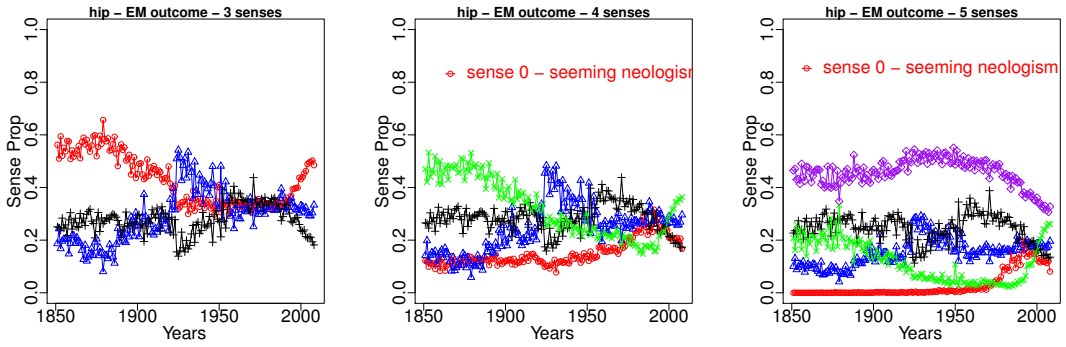


Figure 6.22 – The first, second and third plots show the EM inferred $\pi_t[k]$ sense parameter outcomes for *hip* experiment with 3, 4 and 5 sense settings.

The EM experiments were conducted after seeing the expected sense emergence trend in the Google n-gram viewer. Figure 6.23(a) shows the plots for *very, young, not, really, people, guy* – words that occur with *hip* as a 2-gram where there is an increase in probability around 1960 and a fall-back in 1980. This trend is seen in the EM outcomes, however the top 30 'gist' words for the emerging senses (from 4 sense and 5 sense setting experiments) provided in the table 6.36 do *not* suggest the 'stylish' usage of the word *hip* rather suggests a 'surgery' usage of the word.

gist words - 4 senses	**gist words - 5 senses**
gist(sense 0) **emerging**: *total, replacement, arthroplasty, after, _END_, patients, fracture, surgery, .,), fractures, (, elderly, :, undergoing, following, for, women, replacements, Total, children, revision, spica, prosthesis, thrombosis, in, a, ;, or, roof*	gist(sense 0) **emerging**: *total, replacement, arthroplasty, fracture, after, patients, fractures, surgery, for, :, elderly, risk, a, undergoing, following, replacements, Total, women, with, revision, spica, prosthesis, thrombosis, factors, in, elective, (,), arthroplasties, cast*

Table 6.36 – Top 30 *gist* words for the target *hip* with emerging sense, ranked by comparing word distributions to corpus Probabilities for the EM-outcomes with 4 and 5 sense settings

However figure 6.23(b) showing the 'tracks' for such words *very, young, not, really, people, guy* that we expect were associated with a new sense referring to 'very fashionable' but do not show such an emerging trend as can be seen in the 'Ngram-viewer' plot. From the 'tracks' legend, it can be seen that the words *really, guy* occurs zero times in the dataset, while the words *very, not* do not see any upward trend rather very noisy and the word *young, people* lonely show an upward trend around 1965 and 1975 and does not fallback around 1980 as

it did for 1980. From this it seems that the 'trendy' sense for *hip* is scarcely present in the 5-grams. Therefore Gibbs sampling experiment was not conducted for this target.

One may wonder why there is a change in trend between the Ngram-viewer's 'tracks' and the 'tracks' made from the 5-gram dataset – for Ngram-viewer plot we have considered 2-grams (i.e., 2 words that occur together) so such 2-grams may not occur in 5-grams more than 40 times in the corpus for them to have gained entry into the 5-grams dataset and this could the reason for the words *really, guy* to have zero probabilities in the 'tracks' plot.

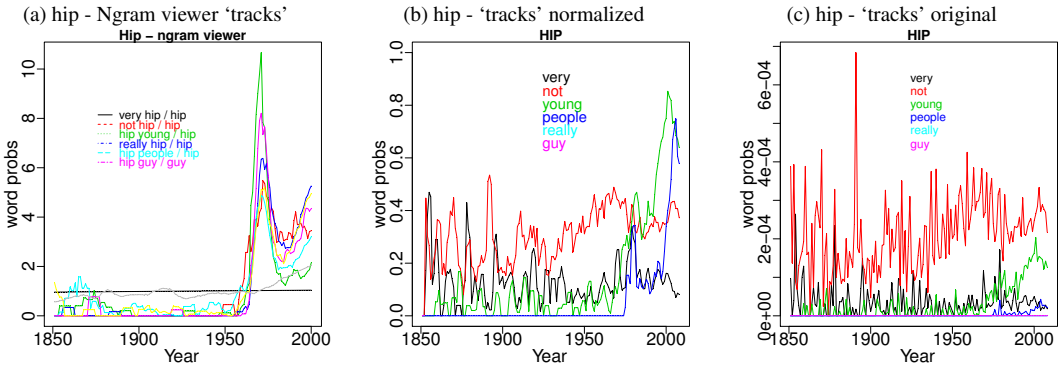


Figure 6.23 – The plot (a) shows the n-gram viewer 'tracks' for 2-grams and plot (b) shows probabilities normalized 'tracks' for some words that we expect to be associated with the 'trendy' sense of *hip* (c) shows the probability (un-normalized) 'tracks' for words that we expect to be associated with the 'trendy' sense of *hip*

Now it is clear that the emerging sense found from the inferred plots is not associated with a language change, rather this is a good example of world change – around 1970, 'hip surgery' and 'hip replacement' procedures were modernized[8] and so usage of such words would have been prevalent, to see such a trend in the 'EM inferred' plots. Additionally the 5-gram "total hip arthroplasty in young" occurs over and over in years between 1978-2008, which is the title of a frequently cited paper. All this relates to the fact that this 5-gram dataset does not have 'trendy' or 'stylish' related sense of the word *hip*.

6.6.2 export

Figure 6.24 shows the EM inferred outcome and 'tracks' plot for the target *export* – (a) shows the inferred sense parameter $\pi_t[k]$ plot for the target *export* with a 5-sense $K = 5$ setting: from the plot it seems the EM did not discover a neologism sense. However a 'tracks' plot was made *data, file, application, program, PDF, HTML, as, format* words that we expect to be associated with 'convert file format' usage of the word *export* – tracks for the normalized and un-normalized probabilities are provided in figures 6.24(b) and (c). The words *pdf, html* when looked in the un-normalized plot, it can be seen that they occur zero times, however from the normalized plots (with smoothing) one can see the tracks for the words *file, PDF, HTML,*

[8] see `https://en.wikipedia.org/wiki/Hip_replacement`

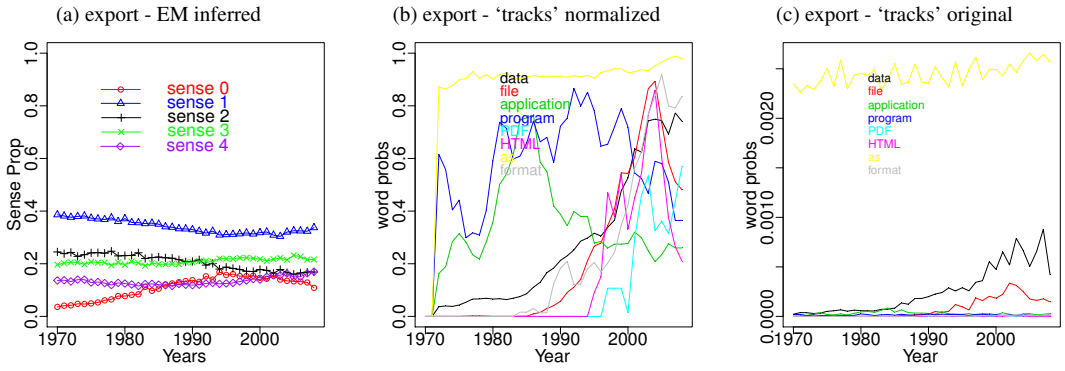


Figure 6.24 – Plots (a) shows the EM inferred $\pi_t[k]$ sense parameter outcomes for *export* experiment with 5 sense settings (b) shows probability 'tracks' for some words that we expect to be associated with 'convert file format' usage of the word *export* that are normalized between 0 and 1 and plot (c) shows the 'non-normalized' version of (b).

format – these have climbing trends emerging between 1985 and 1995. The word *data* has a climbing trend from 1970 onward but that is also likely to occur with other senses. Also, when looked into the un-normalized plot, *file* is the only word that has non-zero probability other than *data*. With these arguments, it is clear that the 5-gram data does not have 'convert file format' related sense of the word *export*.

6.6.3 mirror

Figure 6.25 shows the EM inferred outcome and 'tracks' plots for the target *mirror*. The plot in figure 6.25(a) shows the inferred sense parameter $\pi_t[k]$ plot for the target *mirror* with a 5-sense $K = 5$ setting: from the plot EM seems to have not found the expected emerging sense 'to store a copy of data'. This is not an expected outcome, so the 'tracks' plot for *site, ftp, download, choose, internet, copy, server* – words that we expect to be associated with 'store a copy of data' usage is provided with and without normalization in figures 6.25(b) and (c). For the words *ftp, FTP, download, choose, Internet, internet*, both the normalized and un-normalized 'tracks' plots show zero occurrence, while for the words *site, copy, server* there is non-zero probabilities but they again do not show emergence from a unanimous time. This makes it clear that the 5-gram data set does not have enough 'store a copy of data' related sense of the word *mirror*.

6.6.4 domain

Figure 6.26 shows the EM inferred outcome and 'tracks' plot for the target *domain*. The plot in figure 6.26(a) shows the inferred sense parameter $\pi_t[k]$ plot for the target *domain* with a 5-sense setting: from the plot EM seems to have found two emerging senses $\pi_t[k = 0]$ and $\pi_t[k = 3]$, and for these senses, the top 30 'gist' words such as *binding, terminal, windows, Active, Directory* suggest these senses are related to the 'suffix of internet address' usage of *domain* (these

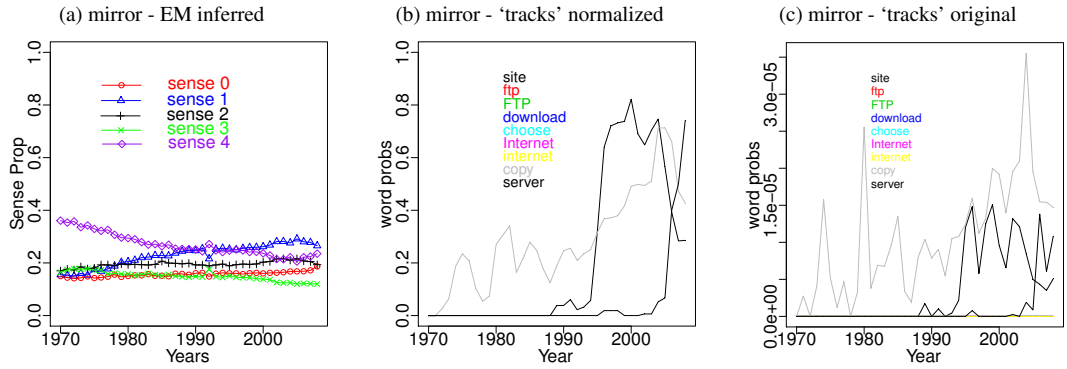


Figure 6.25 – Plots (a) shows the EM inferred $\pi_t[k]$ sense parameter outcomes for *mirror* experiment with 5 sense settings (b) shows probability 'tracks' for some words that we expect to be associated with 'store copies of data' usage of the word *mirror* that are normalized between 0 and 1 and plot (c) shows the 'non-normalized' version of (b).

are provided in table 6.37). However, there are a lot words such as *DNA, transmembrane, knowledge, tyrosine, amino* that do not seem especially associated with the 'suffix of internet address' are also seen in the top 30 'gist' words for both senses. However the tracks in

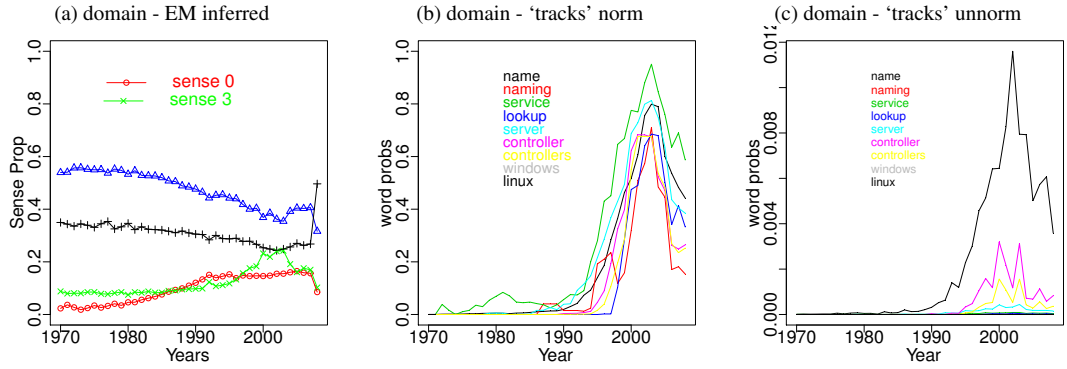


Figure 6.26 – Plots (a) shows the EM inferred $\pi_t[k]$ sense parameter outcomes for *domain* experiment with 5 sense settings (b) shows probability 'tracks' for some words that we expect to be associated with 'suffix of internet address' usage of the word *domain* that are normalized between 0 and 1 and plot (c) shows the 'non-normalized' version of (b).

gist words - sense 0	**gist words - sense 3**
gist(sense 0) **emerging**: *-, binding, terminal, specific, DNA, C, level, N, Windows, (, top, single, ligand, time, –, The, kinase, :,), Directory, Active, NT, protein, transmembrane, knowledge, tyrosine, 2000, amino, activation, like*	gist(sense 3) **emerging**: *name, controller, which, can, as, qualified, it, you, or, names, such, be, (, is, a, ,, fully, one, that, for, but, there, may, IP, set, they, by, your, well, another*

Table 6.37 – Top 30 *gist* words for the target *domain* with emerging sense, ranked by comparing word distributions to corpus Probabilities for the EM-outcome with 4 sense settings – for seemingly emerging senses (0 & 3)

figures 6.26(b) and (c) show the normalized and un-normalized probabilities of *name, naming,*

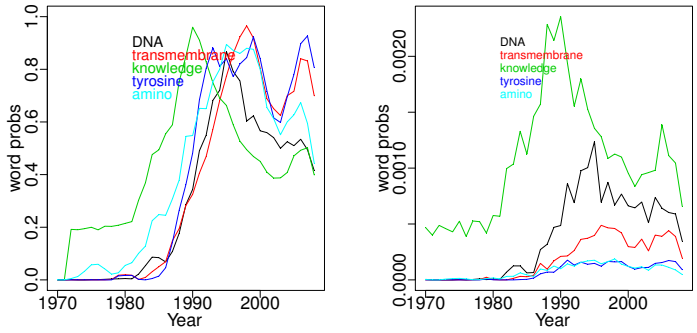


Figure 6.27 – Plots (a) shows probability 'tracks' for some words from 'gist' that are not relevant to 'convert file format' usage of the word *domain* that are normalized between 0 and 1 and plot (c) shows the 'non-normalized' version of (a).

service, lookup, server, controller, controllers, windows, linux words conditioned on *domain*, that we expect to be associated with 'suffix of internet address' usage of the word *domain*. The normalized tracks in (b) shows increase in probabilities between years 1990 and 2007, however the un-normalized tracks in (c) show close to zero occurrence for words *naming, service, lookup, windows, linux*, and very small probabilities for words server, controller, controllers. The word *name* may occur with other usage possibilities as well.

On further looking into the 'tracks' (shown in figure 6.27) for the words *DNA, transmembrane, knowledge, tyrosine, amino* that are not related to the 'suffix of internet address' usage of the target, they show an increase around 1988, the same time the associated words found an increase. However, the un-normalized plot in figure 6.27(b) shows higher probabilities for such un-related words, which makes it clear that the 5-gram data does not have enough 'suffix of internet address' related sense of the word *domain* and the emerging senses seen in the EM inferred plot is the result of the increase in the probabilities of the un-related words seen in 'tracks' of figure 6.27.

6.6.5 high

The word *high* has a new found sense to mean 'in inebriated state' – expecting such a sense to be identified from Google 5-grams dataset, the EM experiments were conducted. Figure 6.28(a) shows the EM inferred outcomes of sense parameters $\pi_t[k]$ with 5 sense $K = 5$ settings, however the top 30 'gist' words provided in table 6.38 do not suggest any word that are associated with the expected 'in inebriated state' usage of the word *high*. The words such as *pressure, speed,*

gist words - 5 senses
gist(sense 2) **emerging**: -, *pressure, speed, blood, risk, tech, energy, heat, quality, frequency, pitched, –, ranking, resolution, medium, performance, technology, rise, voltage, density, over, fat, low, grade, dose, temperature, diet, water, strength, income*

Table 6.38 – Top 30 *gist* words for the target *high* with emerging sense, ranked by comparing word distributions to corpus Probabilities for the EM-outcomes with 5 sense settings

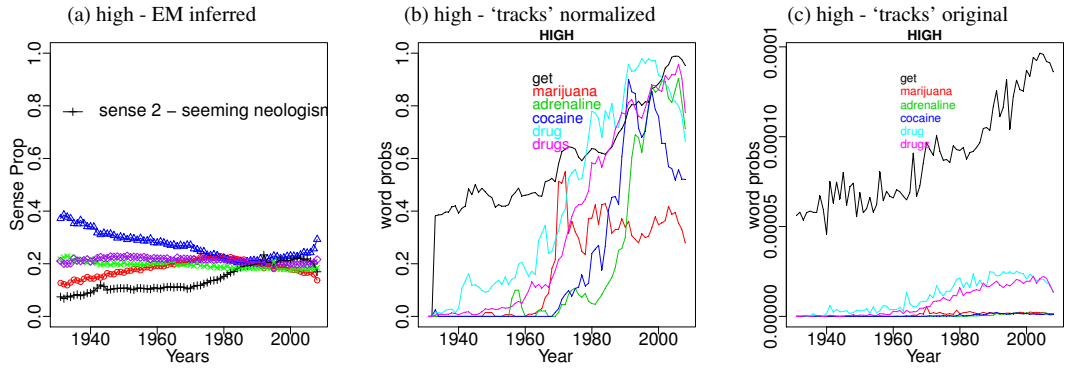


Figure 6.28 – The first and second plots show the EM inferred $\pi_t[k]$ sense parameter outcomes for *high* experiment with 3 and 5 sense settings.

risk, energy, frequency are mostly related to the 'scientific' usage of the word *high* and there are in not relevant to the expected neologism usage 'to be in inebriated condition'. The 'tracks-plot' in figure 6.28(b) for *get, marijuana, adrenaline, cocaine, drug, drugs* – words that expect to be associated with the 'in inebriated state' usage show an increase in usage for the most words *marijuana, adrenaline, cocaine, drug, drugs* between the years 1958 and 1970. However the actual occurrence of these words lie close to zero in the 'tracks' shown in figure 6.28(c) – the plot made out of actual probability values. Additionally, the word *get* seems to have got high occurrences and it is likely to occur with the other senses of the word *high*. This makes it clear that the 5-gram data set does not have enough 'in inebriated state' related sense of the word *high*.

6.7 Further model tests

In section 6.2, there was a discussion on the pseudo-neologism tests conducted to establish the fact that the model has the ability to discriminate senses over time to identify the expected neologism sense. After conducting all the experiments proving the model can identify the expected neologism sense from the actual targets in sections 6.3, 6.4 and 6.6, now it is important to establish the role of the data set in the model's ability to identify the neologistic sense – this is discussed in 'ablation' section 6.7.1. All the neologism experiments were conducted with manual sense assignments for K (number of senses), however this manual intervention can be eliminated by inferring the number of senses required to identify the neologism sense – for this 'Merge tests' were conducted and a discussion on this is provided in section 6.7.2. This suggests just an approach to infer the number of senses. An adaptation of the approach used by Cook et al. [2014], Lau et al. [2012] to rank all the targets in identifying the targets with neologistic ones is provided in section 6.5.

6.7.1 Ablation tests

It is interesting to know the impact of the data (size) on the estimation procedures used on the 'diachronic' model. This is looked at with 'ablation' tests, where the data is made successively smaller. The idea here is to see (i) does the outcome remain roughly the same as the data is reduced (ii) what is the minimum amount of data required for the model to infer the neologism sense? For this kind of test, *mouse* dataset between 1970 and 2008 with $\sim 813k$ data items was considered. Based on run-time considerations (discussed in section 6.3.1), these tests were conducted using EM algorithm. For the test set-up, consider the *mouse* data-set of varying quantities and visualize (as plots) the outcomes of the EM to see the impact of the model has over the change in data quantities. It is expected that as the amount of data is reduced, the smoothness in the sense parameters $\pi_t[k]$ are reduced.

To make a succession of reduced data sets the following was done. A random shuffle of the complete data set ($813k$ data items) between 1970 and 2008 was made. Then for a succession of percentages p, the first $p\%$ of that permutation defines the subset. This was done for the percentages $75\%, 60\%, 45\%, 30\%, 15\%, 5\%, 2\%$ and 0.5%. This should roughly ensure that each year's data gets equally reduced. On each subset the EM estimation was run. Figure 6.29 shows the plots of the ablation test outcomes. In each plot and for each sense k, the single solid line shows the sequence of estimated $\pi_t[k]$ from time t.

From the plots it can be seen that as the data size is reduced between experiments, only subtle changes are seen. It can be noticed that after reducing the input data size to 15%, the smoothness in the neologism sense $\pi_t[k=1]$ looks unchanged, but the other senses seems to have started losing their smoothness. Even with just 2% of the data (ie., $\sim 16k$ data items), the neologism sense can be identified but with some jaggedness. However as the data size is reduced further to 0.5% of the original size with $\sim 4K$ data items, the EM experiment plot provided in figure 6.29(h) did not find the neologism sense. This test establishes a point that we need at least $\sim 16k$ data items to see a neologism sense for a 3 sense variant for 39 years with an average of ~ 410 items per year. But the number of data-items in each year is different with a minimum number of 160 data-items in year 1971 and a maximum number of 643 items identified in year 2004.

6.7.2 Merge tests

The number of senses required to discover a neologism sense may vary depending on the data and its size. Further, it is also noteworthy that the model may discover a sense that a human may not expect such a discovery. Consider the EM outcomes for the target *boot* (section 6.3.8), where the model discovered the neologism sense with 5 sense setting ($K = 5$) (the 'gist' words for the neologism sense $k = 3$ is provided in table 6.22 and the gist words for other senses are provided in table 6.29(e)). In this, the model seems to have inferred a leather sense $k = 0$ and shoe sense $k = 2$ of *boot*, which a human would interpret them to be one sense as $k = 0$ is a mild variant of the $k = 2$ – for a human this is unexpected. In an attempt to infer the number of

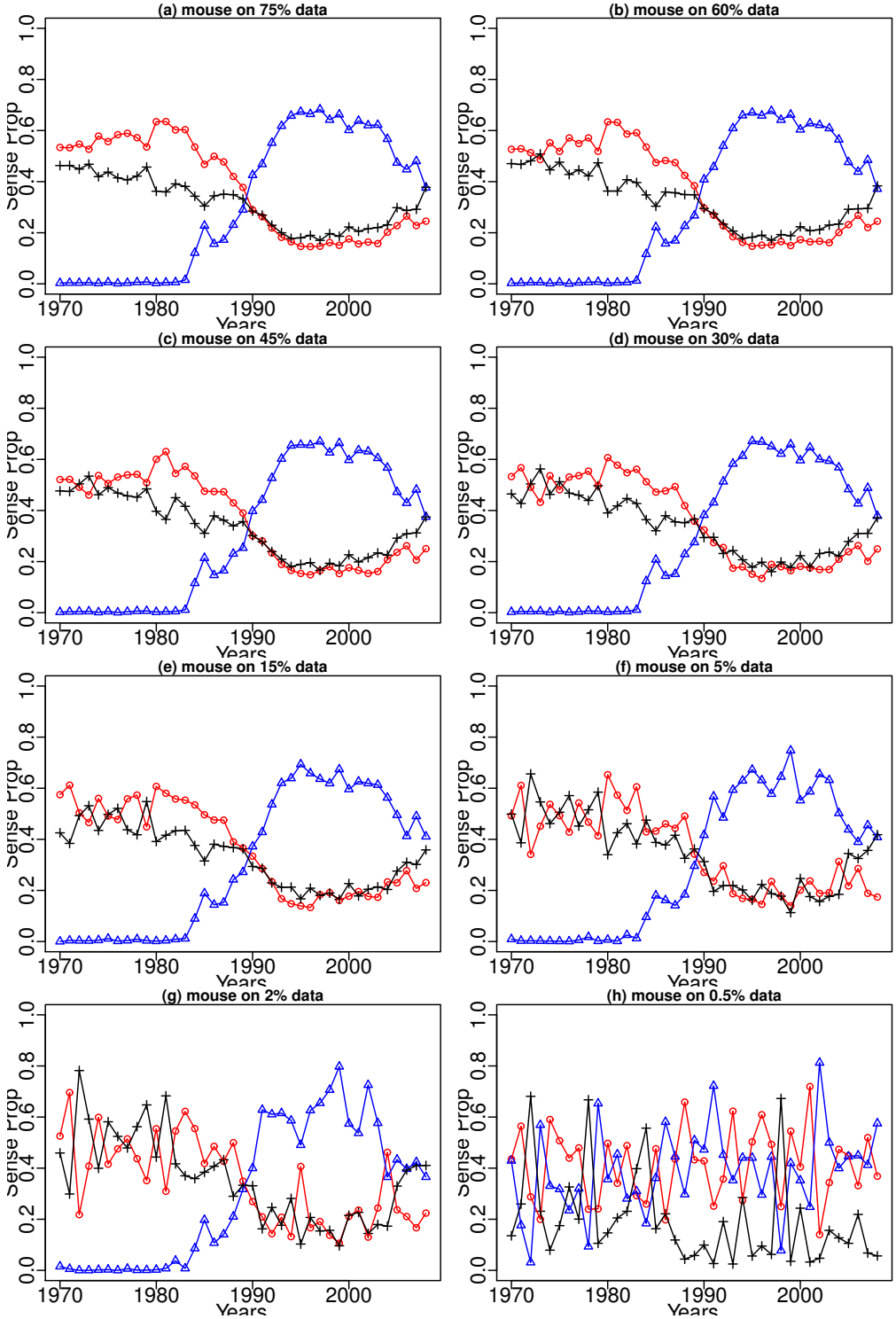


Figure 6.29 – Ablation test outcomes on *mouse* dataset – plots from (a-h) shows the EM outcomes from $75\%, 60\%, 45\%, 30\%, 15\%, 5\%, 2\%$ and 0.5% of the complete dataset between the years 1970 and 2008.

senses required by EM, we conduct this merge experiment where such sense distributions may be merged to still get the neologism sense at $K-1$ senses.

EM algorithm being sensitive to initializations (Biernacki et al. [2003], Yang et al. [2012]), there is a possibility that we can merge two inferred sense distributions $\boldsymbol{\pi}_t[k]$ of sense k and inferred word distributions $\boldsymbol{\theta}_k$ of sense k based on the distances between different senses of inferred $\boldsymbol{\theta}_k$ distribution and the merged sense $\boldsymbol{\pi}_t[k]$ and $\boldsymbol{\theta}_k$ word distributions can be used to initialize the respective parameters for a new EM run. To get the distances among all the senses K of $\boldsymbol{\theta}_k$ distributions, 'Kullback Leibler divergence' (KL-div) distance metric was used. From the KL-div distances, the closest two sense distributions are merged. Merging the sense distributions is done the following way: Consider the total number of senses to be $K=3$ from $\{k_0,k_1,k_2\}$ with k_0 a neologism sense and we want to merge k_1 and k_2 of $\pi_t[k]$ distributions, now the new $\pi_t[k_1]$ will be,

$$\text{new } \boldsymbol{\pi}_t[k_1] = \boldsymbol{\pi}_t[k_1] + \boldsymbol{\pi}_t[k_2]$$
$$\text{new } \boldsymbol{\pi}_t[k_0] = \boldsymbol{\pi}_t[k_0]$$

Similarly, $\boldsymbol{\theta}_k$ parameters can be merged as below

$$\text{new } \boldsymbol{\theta}_{k_1} = 0.5 \times (\boldsymbol{\theta}_{k_1} + \boldsymbol{\theta}_{k_2})$$
$$\text{new } \boldsymbol{\theta}_{k_0} = 0.5 \times (\boldsymbol{\theta}_{k_0} + P_{corp})$$

It would be interesting to see that when EM is run with these initializations deriving from a merge of two close senses, the previously detected neologism sense is preserved.

For this kind of testing, the inferred outcomes *surf* (section 6.3.7) has been considered. From the inferred EM outcomes the KL-div distance between the inferred $\boldsymbol{\theta}_k$ distributions of *surf* are computed and provided in table 6.39. Based on the KL-div distances, 'hierarchical' clustering[9] was used to find the closest sense distributions as the numbers provided in the table may not just be apparent to human interpretation.

	k_0	k_1	k_2	k_3	k_4
k_0	0	5.81866	5.01205	5.46880	5.61781
k_1	5.81866	0	2.75820	5.95901	4.28056
k_2	5.01205	2.75820	0	6.70997	3.00419
k_3	5.46880	5.95901	6.70997	0	8.12131
k_4	5.61781	4.28056	3.00419	8.12131	0

Table 6.39 – KL-divergence (symmetric) distance between inferred $\boldsymbol{\theta}_k$ distributions for *surf*

The hierarchical cluster plot for *surf* is presented in figure 6.30. From the plots it can be visually seen from the RHS plot that (i) 's1' referring to k_1 and 's2' referring to k_2 are closer, similarly (ii) senses k_1, k_2 are closer to k_4. As discussed earlier, it is apparent to human intuition that k_1 and k_2 are close as they both seem to be related to *water sport* related sense. So we first merge the senses in (i) and attempt an EM run to see the new outcomes. Based on merging

[9]For *hierarchical* clustering, R's 'cluster' package and 'agnes' method was used.

the senses for the current parameter estimates, the new sense numbers are allocated as per the table 6.40, where the sense numbers allocated after the first level merge are provided – here k_0 is repeated for the merged senses, similarly details on the second level merge is also provided in the table.

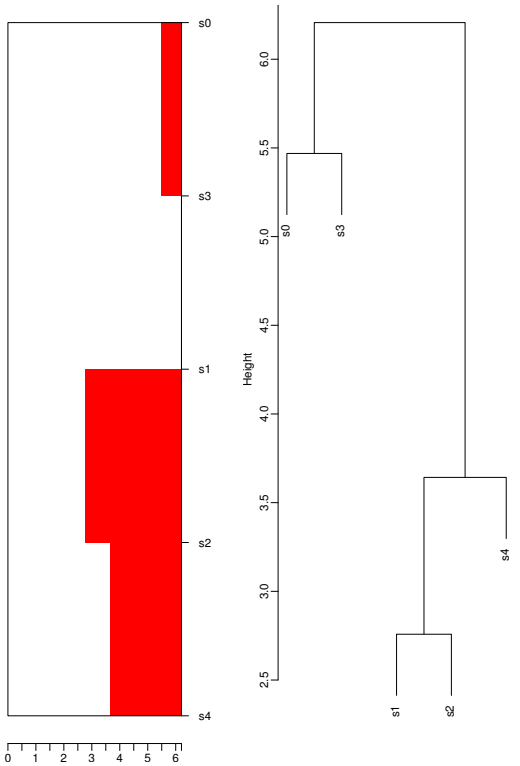


Figure 6.30 – Hierarchical cluster plots based on KL-divergence distances between $\boldsymbol{\theta}_k$ distributions for *surf*. The sense numbers are represented as $s0, \ldots, s4$ and the height provided in the plot gives the distance between the clusters.

inf senses	1^{st} **merge**	2^{nd} **merge**
k_0	k_1	k_1
k_1	k_0	k_0
k_2	k_0	k_0
k_3	k_2	k_2
k_4	k_3	k_0

Table 6.40 – Sense numbers allocated based on merging inferred $\boldsymbol{\theta}_k$ distributions for *surf* – see text for further details.

Figure 6.31 shows the plots based on the initializations from merging the senses (here senses k_1 and k_2 are merged to k_0 and k_3 the neologism sense [as can be seen from the first plot of figure 6.13] is assigned to sense k_2) and inferred sense distributions $\pi_t[k]$ from EM. The inferred sense distributions $\pi_t[k]$ seems to have learned the sense distributions very close to the initialized distributions. Additionally, a second level merge is also done and the initializations from this merge are provided in figure 6.32. The inference procedure seems to have the learned the sense distributions again very close to the initializations from the second merge, however it is still not very clear to whether the learned distributions are still right – this can be verified

with the ‘gist’ words.

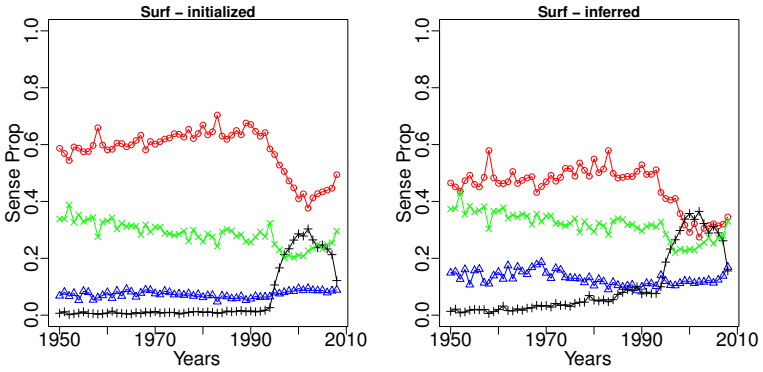


Figure 6.31 – The plots show the initialized and inferred sense distributions $\pi_t[k]$ for the target *surf* – shows first level merge

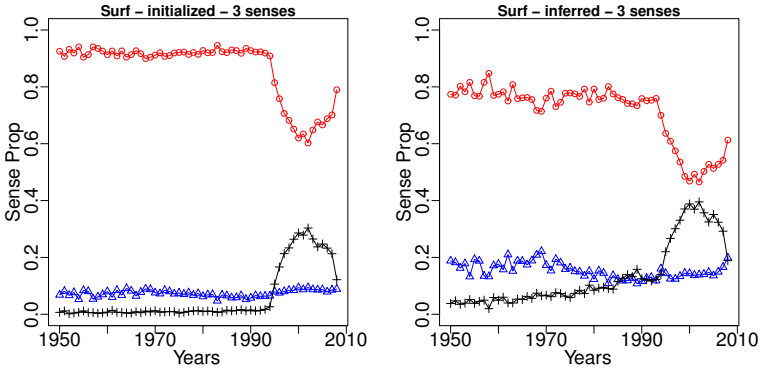


Figure 6.32 – The plots show the initialized and inferred sense distributions $\pi_t[k]$ for the target *surf* – shows second level merge

gist words - EM outcome
gist(sense 0): *of, sound, into, out, in, zone, pounding, edge, roar, on, line, beach, crashing, by, thunder, breaking, ;, beyond, through, ocean, outside, distant, upon, from, shore, clam, along, at, above, noise*
gist(sense 1): *_START_, was, as, is, The, hear, so, could, I, which, but, when, high, it, be, can, that, you, not, too, to, broke, there, no, beat, When, a, If, listening, heavy*
gist(sense 2) **neologism**: *Internet, Web, ”, Net, or, for, net, mail, -, web, wind, swimming, fishing, diving, ’, Wide, World, turf, sailing, ,, skiing, you, to, e, ?, sand, games, scuba, channel, while*

Table 6.41 – Top 30 *gist* words for the target *surf* ranked by comparing word distributions to corpus Probabilities, where the word distributions are obtained from the new merged distributions.

Table 6.41 provides the top 30 ‘gist’ words for all the senses learned from the inferred word distributions $\boldsymbol{\theta}_k$ with 3 senses. The ‘gist’ words for ‘sense 0’ representing k_0 unanimously represents the ‘water sport’ sense of the word *surf*, however the gist words for the neologism sense seem to have got rather more unexpected words *wind, swimming, fishing, diving, turf, sailing, skiing, sand, games, channel* in the context of ‘exploring internet’ usage – most of these words are related to the ‘water sport’ usage. Further, a tracks for these words in figure 6.33, these shows the words to have increasing probability of occurrence in the context of *surf* in 1960’s and goes up later which can seem to have disturbed the sense inference for the

neologism sense $\pi_t[k=2]$ as well.

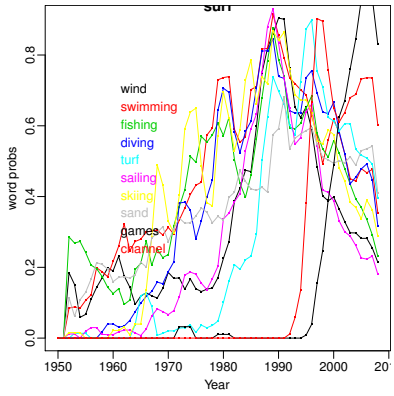


Figure 6.33 – The plots show the initialized and inferred sense distributions $\pi_t[k]$ for the target *surf* – shows second level merge

The experiment on the target *surf* is suggestive of a possible approach to the somewhat unsatisfactory situation that a neologism was sometimes detectable with some larger number for K, and so possibly with the results that some of the detected sense represent distinctions a human would make. Through an automated process, having found a neologism at high K, repeatedly close senses could be merged and EM re-run. If the neologism remains detected at a lower K then probably the other detected senses will be more intuitive. At the moment however it is just a conjecture that such an approach would generally work.

6.8 Comparison with the prior work

Prior work in terms of alternative models and algorithms and evaluation procedures has been described in sections 2.2 and 2.3. As has already been pointed out, due to the different data sets, targets, approaches to ground truth and evaluation criteria, it is not possible to make direct and quantitative comparisons of prior work with the results presented in the chapter[10] – the closest we have come to thi was in section 6.5, where we adapted the approach of [Cook et al., 2013, 2014, Lau et al., 2012] of using a 'novelty score' to try discriminate between known neologisms and non-neologisms. There are nonetheless some observations that is seems reasonable to make in comparing their experimental outcomes to ours.

Recall that in the work of [Cook et al., 2013, 2014, Lau et al., 2012] a LDA/HDP topic modeling approach. This has the theoretically attractive feature of deciding for itself how many senses there should be. Whereas in the experiments reported above, the numbers of senses was between 3 and 5, in the examples which they give in their paper, their algorithm seems to settle on a somewhat larger number.

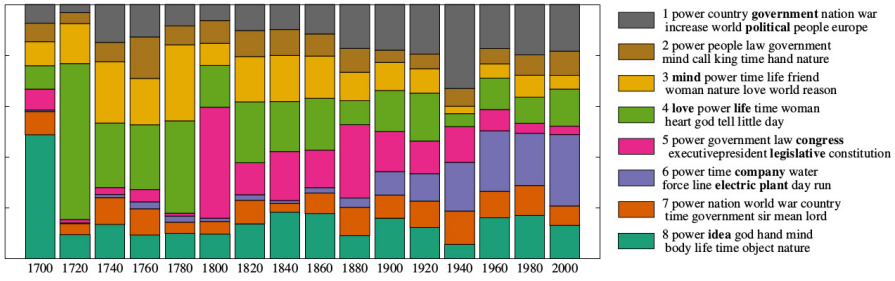
Figure 6.34 shows the example 'cheat' that has 9 senses, several of them seemingly rather close (they suggest 1,3 and 4 really represent the same sense), and several seemingly very hard to interpret. This listing of the terms points out one other contrast that they have used lemmatized

[10] Frermann and Lapata [2016] make the same point at several points of their paper as well.

Sense Num	Top-10 Terms
1	cheat think want ... love feel tell guy cheat#nsubj#include find
2	cheat student cheating test game school cheat#aux#to teacher exam study
3	husband wife cheat wife_#1 tiger husband_#-1 cheat#prep_on#wife ... woman cheat#nsubj#husband
4	cheat woman relationship cheating partner reason cheat#nsubj#man woman_#-1 cheat#aux#to spouse
5	cheat game play player cheating poker cheat#aux#to card cheated money
6	cheat exchange china chinese foreign cheat_#-2 cheat_#2 china_#-1 cheat#aux#to team
7	tina bette kirk walk accuse mon pok symkyn nick star
8	fat jones ashley pen body taste weight expectation parent able
9	euro goal luck fair france irish single 2000 cheat#prep_at#point complain

Figure 6.34 – Screen-shot of Table 2 from Lau et al. [2012] showing top-10 terms for each of the senses induced by HDP for the word 'cheat'.

words, position features and syntactic dependencies, whereas the current work uses the raw text without any processing. In Frermann and Lapata [2016], in the part of the paper where they do discuss the trajectories obtained for the sense probabilities for certain target words, it is clear that they ran their experiments also seeking a relative high number of senses, namely 8. The picture below shows their outcome for the target *power*



and akin to Lau et al. [2012] they suggest that several seem to represent the same sense (eg. 1, 2, 5, and 7 all representing an institutional sense of power). This issue of number of senses remains a difficult one. Simply looking at the sense emergence issue, a direction for further research is to examine more systematically the effect of the number of senses on the effectiveness in detecting novel senses.

We have already discussed the theoretical design of model used in Frermann and Lapata [2016]. Recall that they have a specific prior whose intention is to encourage *smooth* change. It seems worth noting that whilst there was no prior used in the proposed EM and Gibbs algorithms to encourage smooth change of the $\boldsymbol{\pi}_t$ values, nonetheless relatively smooth change *is* obtained, and sense emergence was successively detected in a number of cases. This suggests that for the n-gram data at least, the more complex system of Frermann and Lapata [2016] is not required. It may be that they required this smoothing prior because the dataset they used contains approximately 100 times fewer occurrences for a given target per time-period compared to the n-gram dataset we have used: they used the *Corpus of Historical American English* Davies [2010] dataset, and have a time-resolution of 10-year time spans.

6.9 Discussion

This chapter reported the results obtained using our diachronic model, and the proposed EM and Gibbs sampling parameter estimation approaches. Besides the preliminary 'pseudo-neologism' tests, experiments were carried out on both neologism targets, which were anticipated to show sense emergence (relative to the period looked at), and also non-neologism targets that were anticipated to not show semantic emergence.

Summarising the outcomes obtained, we can say that of the 10 neologism targets looked at section 6.3, all 10 were identified to have an emergent sense (as established according to the `EmergeTime` algorithm of section 4.1.4). In 6 out of 10 cases this happened with the number of senses K set to 3. In the other 4 cases it happened with the number of senses K set to 5. In all cases the inferred emergence time was later than D_0^c, the OED first citation date, which was argued to be a minimum requirement on accuracy. We also proposed the 'tracks'-based procedure via which to establish C_0, the sense emergence date within the corpus, and in all cases the sense emergence dates which were obtained via the EM and GS unsupervised estimation procedures agreed with 'tracks'-based emergence date to within 10% of the time-span considered. So under these two assessments of dating accuracy, in all 10 cases the emergence date of the apparent semantic neologism was 'accurate'. There does not seem to have been other work which is addressed to an extended time-line[11] and which attempts in this way to assess the accuracy of an apparent sense emergence.

If we include the 5 further neologism targets that were discussed in section 6.6, then the success rate would drop to $\frac{2}{3}$, though as was discussed in section 6.6, there is quite a lot of evidence that suggests that the anticipated senses are objectively absent from the 5-gram data.

Concerning the non-neologism targets which were looked at in section 6.4, when the number of senses was set to 2 or 3, in 8 out of 8 cases no neologism was detected, whilst when the number of senses was set to 5, this happened in 7 out of 8 cases.

For both neologism targets and non-neologism targets it would be unjustified to extrapolate from these experiments to what might be the success rates with far larger samples of neologisms and non-neologisms. So no particularly strong claims can be made about the likely success of a system based on the current model if deployed as the kind of semantic neologism warning system that was alluded to in section 1.4. All that we can really say is that the method used has shown some promise.

Two aspects of this probably deserve to be emphasized here. One is that the model made the strong assumption that senses, seen as probability distributions on context words, be treated as time *independent*. It might have been the case that this assumption is completely inconsistent with the facts of language. The results show, at least for a certain collection of targets, that this assumption is *not* completely inconsistent with the facts. A second point is that the 5-gram data provides only 4 context words around a target. One might also think that such short contexts would make it impossible to infer useful information – for example Frermann and

[11] as opposed to large eras

Lapata [2016] use 10 context words. Again the results show, at least for a certain collection of neologism targets, that it is possible to make inferences on such a small amount of context. This may be because of the fact that the Google 5-gram datasets for most of the targets were large and the number of vocabulary items for each target turned out to be far smaller in number (this can be seen from table 6.5).

The experiments involved the two parameter estimation techniques, EM and Gibbs sampling. EM should give an estimate which is a mode of a posterior Dirichlet, and the Gibbs sampling variant should give a mean of a posterior Dirichlet. For the most part, the two methods gave similar outcomes. On the one hand that fact that when the EM outcome indicates a neologistic sense it is also the case that the Gibbs outcome does, and vice versa, gives a kind of reassurance that the one or the other outcome is *not* a fluke. On the other hand, on theoretical grounds the EM-based estimate could be somewhat different to the GS-based estimate, and one might argue the GS-based estimate to be a more motivated one. There were a few indicators that on occasion the GS approach arrived at more convincingly separated senses but it remains to be seen whether on a more substantial set of tests a greater difference is found between the EM and GS outcomes. On grounds of execution time, on the basis of the current test items one could argue for using just the much faster EM-based approach.

The question of the setting of the number of senses remains unresolved. The merge tests that were reported in section 6.7.2 looked at one aspect of this. A neologism was detected at a particular K, and by merging senses which had small Kullback-Leibler divergence, and re-running EM from a start point defined by the merge, the neologism remained detected at $K-1$. Another aspect of this is the setting of the number of senses so as to most reliably and accurately identify cases of sense emergence. Roughly speaking, the experiments indicated that with senses in the 3 to 5 range, true semantic neologisms were detected, and true semantic non-neologisms were identified as such. One possibility following the intuition of the Kullback-Leibler divergences between the inferred per-sense word distributions might be to do repetitions of experiments with increasing K while monitoring the divergences and cease to increase K once components are created with a divergence falling below some threshold. Another possibility would be to try an approach which balances the data probability achieved at a particular K with the model complexity associated with that K, using for example the Akaike Information Criterion [Akaike, 1974] or a variant of it. These might provide a way to have an automatic way to have the size of K vary with the target. At the moment this remains a possibility for future work. Besides the additional computational overhead of such an approach, to draw meaningful conclusions about the success of such strategies will probably require a larger sample of neologisms and non-neologisms than those that were addressed here.

Chapter 7

Conclusions and future work

The diachronic model proposed in this thesis (section 3.2) models $P(S|Y)$ and $P(\boldsymbol{w}|S)$, where $P(S|Y)$ represents probability of sense given year and $P(\boldsymbol{w}|S)$ represents probability of context words given the sense. This is the fundamental contribution made in identifying a 'novel' sense from the given dataset containing occurrences of a semantic neologism. The proposed model is a simple one that has never been previously used for this task, although there are works (discussed in section 2.2) that use forms of static model without the time dependency parameter $P(S|Y)$. It is from this proposal the other contributions of the thesis emerge such as deriving the parameter updates for EM (section 3.3) and Gibbs (section 3.4) algorithms. In the area of evaluation, one contribution made here is the proposed use of a 'tracks' plot (section 4.1.3) to find the corpus emergence date. Another contribution in that area is the use of 'pseudo-neologisms'.

The proposed model is tested with experiments conducted on a number of semantic neologisms and non-neologism targets in chapter 6. The simplifying assumptions made in the model may not have been convincing until the experiment outcomes based on the model are discussed. Further, the experiments were conducted without any sophisticated data processing (such as stemming, lemmatization, stop word removal) and the diachronic model is still able to find the novel sense. Additionally, there were limitations with the data sources – (1) just 4 context words were available with 5-gram data which is considered as a strong limitation (2) the Google search data set had a number of impurities and limited to a maximum of 100 data items per year; considering all the said limitations and simplicity in the model, the experiments showcase nonetheless plausible outcomes. A summarization of the key contributions made in this thesis and the future work that could be built over these contributions are presented further in this chapter.

7.1 Summary of the contributions

Diachronic model A generative model $P(Y,S,\boldsymbol{w}) = P(Y) \times P(S|Y) \times \prod_i P(w_i|S)$ is proposed in section 3.2, with the terms $P(Y)$ provides the relative abundance of data items with year Y

in the dataset; $P(S|Y)$ term represents time-dependent sense likelihood; and $\prod_i P(w_i|S)$ term expresses that the context words are independent of time given the target's sense. The dynamic nature of this model is given by the $P(S|Y)$ parameter directly expressing the idea that the sense varies with time (years). This only model that closely resembles is the recent work by [Frermann and Lapata, 2016]. Further, this model is a simple one compared to the other works (section 2.2) that can identify neologism sense.

EM and Gibbs procedure To estimate the model parameters of the proposed diachronic model, the EM and Gibbs sampling procedures are proposed in sections 3.3 and 3.4 where the parameter updates are also derived. The EM procedure repeatedly calculates the expected completions of the incomplete data, and derive new parameters by maximum likelihood estimation of the expected completions through E-step and M-step. This provides a one-point estimate for the parameters. But Gibbs sampling procedure provides a chain of parameter estimates. The idea here is to get the desired posterior distribution after iterating through a number of sampling steps from the conditional distribution. A one-point parameter estimate is made by computing a mean of the samples.

Tracks plot For a given target T and a dataset containing the targets (data extraction procedures are discussed in chapter 5), K number of senses are discovered by the EM and Gibbs procedures. For some sense $S = k$ to be identified a neologism sense, it is expected the values for $P(S|Y)$ are close to 0 during the initial period and continue to climb thereafter. To evaluate the proposed model, the date of emergence C_0 of the neologistic sense is identified in the corpus. C_0 is the time at which the neologistic sense for the word departed from close to zero and continued to climb thereafter (section 4.1.1).

To evaluate this, a novel method is proposed using a so-called 'tracks' plot (section 4.1.3). For a target word T, there are words which it is intuitive to expect in the vicinity of T in its neologistic sense of T, but not in its vicinity in its other senses. The idea behind this 'tracks' plot is if the per-year probabilities for such words in the dataset $P(w|Y)$ are plotted, they are expected to be at close to 0 during an initial period and take off at C_0. To identify C_0 from sense parameters $P(S|Y)$ and 'tracks' we propose an automatic, objective procedure that can be to calculate an emergence time (if any) from a time series.

As a further level of verification, there is a proposal provided in section 4.1.1 to use the earliest citation date of the word-sense pair – call this D_c^0. Oxford English Dictionary (OED) records the earliest citation date D_c^0 of word-sense pairs. These dates are not the same as C_0, and D_c^0 being the earliest citation date, it is expected that $D_c^0 <= C_0$. This seems a reasonable assumption made and the experiment outcomes were assertive about this assumption.

EM and Gibbs sampling experiments on Google 5-gram datasets In chapter 6 the EM and Gibbs sampling experiments based on Google 5-gram are reported and discussed in detail. Maximum likelihood and mean estimates were obtained using EM and Gibbs sampling exper-

iments for all the targets reported in chapter 6. For all the experiments reported in chapter 6, non-informative priors were used. For most of the neologism targets, the neologistic sense outcomes of EM closely resembled the Gibbs outcomes. Although the number of senses required to find the neologistic sense slightly varied for different targets, both algorithms were successful in finding them. Similar experiments were conducted on 'negative' targets (section 6.4) – targets known not to exhibit sense emergence. For them, no novel sense was detected. There were a few semantic neologism targets for which the model did not discover an expected novel sense, but there are valid reasons associated with such outcomes and those were discussed in section 6.6.

Comparisons with prior work The experiments conducted using the diachronic model have produced convincing outcomes in identifying a neologistic sense. But this does not rule out other possibilities to approach this task. So relevant comparisons from the prior work are carried out in section 6.8.

To summarize the model comparisons, it was identified that many of the prior works have considered one of these two design options: (i) *pool* all training data together to apply some form of *sense induction* algorithm which is time-unaware, then assign the likeliest senses to examples, and then to finally check for a correlation with time (ii) to separate the data into eras, perform independent *sense induction* algorithm on each subset and then seek to consider how the sense representations from each era may (or may not) be linked to each other.

In terms of the models used for sense induction, many of the works can be classified under two different modeling schemes (i) uses probabilistic generative models (ii) clustering approaches. The actual approaches used in the prior works are discussed in section 2.2.

With no large scale time-stamped sense-labeled corpora available to find the time at which a neologistic sense emerged in a particular corpus C_0, the approach to ground-truth concerning sense emergence is discussed in section 4.1. Many of the prior works have considered *dictionary first inclusion* (section 4.1.2) D_0^i as their evaluation option.

7.2 Future work

Experiments using data from social media The experiments reported in this thesis are based on Google 5-gram dataset, a digitized books corpus and Google search dataset that comes from Google time-line search. Google 5-gram dataset has a very formal usage of semantic neologism targets but it is possible to get data that are more informal in nature from Google time-line search. This way experiments were carried out on single-word and multi-word targets, but experiments need not necessarily be limited to these. This work can be extended to idioms as well. For example, the idiom "have bigger fish to fry" in normal language use refers to "have other big fish to fry", while the novel sense here would be "to have more important things to do". A machine translation engine may just be able to perform a literal translation of the idiom, and not in the idiom sense. To find novel usages of idioms, it may be a good idea to explore the data from social media such as blogs and twitter.

Tests on a different language In this thesis, the EM and Gibbs experiments are reported based on raw text contexts from English datasets (chapters 6). The outcomes are plausible and the proposed evaluation also works reasonably well. However, there is a possibility of extending this work to datasets from different languages and see the impact of diachronic model on morphologically rich languages such as Tamil and Sanskrit (south Asian languages). It may turn out the model works just as expected as it did on English datasets, however there is also a possibility that just the raw-text contexts may not be enough for diachronic sense discrimination in finding a neologistic sense. This work may further require investigating other context possibilities such as morpho-syntactic features.

Using a collapsed Gibbs sampler The current work includes a simple Gibbs sampler constructed to estimate the parameters of the diachronic model. In the future work, constructing a collapsed Gibbs sampler is a possibility. The motivation behind using a collapsed Gibbs sampler is, it is considered to be computationally faster than a simple Gibbs sampler.

Data extrapolation Google 5-gram dataset as discussed in section 5.3, is not really a corpus but a (per-year) frequency table for 5-gram *types* that gives time-stamped counts on 5-gram types arising by sliding a window over the original texts, a window in which a succession of token sequences appear; basically the window contents will contribute to a count if the tokens do not span certain boundaries such as sentence or paragraph endings. This way a maximum of 4 context words are available. Each occurrence of a target word can contribute to 5 different 5-gram counts according to its position in the 5-gram (as each 5-gram come from sliding a window over original text). The thesis, however, treats all 5-grams featuring a given target as independent. But there is a possibility that these 5-grams can be pieced together using some form of *extrapolation* procedure to get a dataset with longer context around the target. This way, experiments can be conducted on a dataset with more context and with closer to actual count of the target from Google's digitized book holdings.

Language changes may happen with locales Language changes over time are explored in this thesis, but there is a possibility that languages change also with locales. As an example, consider the word *intimate* which provides the sense of 'in a private relationship', but this word takes a predominantly different sense in Indian English to mean 'to inform in advance'. With such language changes, one can foresee a mistranslation in SMT systems. This form of language change over locale can be explored with a joint probability model of locale L, sense S and contexts $\boldsymbol{w}$ is given by:

$$P(L,S,\boldsymbol{w}) = P(\boldsymbol{w}|S,L) \times P(S|L) \times P(L) \tag{7.1}$$

On the joint probability in 7.1, consider the following conditional independence assumptions

(i) words W are conditionally independent of locale L and (ii) every word is independent of each other, to get:

$$P(L,S,\boldsymbol{w}) = \prod_i P(w_i|S) \times P(S|L) \times P(L) \tag{7.2}$$

Having said this, intuition suggests that such a new sense was acquired at some time. These form of language changes with time over locales can be explored with a joint probability model of locale L, year Y, sense S and contexts $\boldsymbol{w}$ given by:

$$P(L,Y,S,\boldsymbol{w}) = P(\boldsymbol{w}|S,Y,L) \times P(S|Y,L) \times P(Y|L) \times P(L) \tag{7.3}$$

For the joint probability in 7.3, consider the following conditional independence assumptions (i) words W are conditionally independent of year Y and (ii) year Y is conditionally independent of locale L, to get:

$$P(L,Y,S,W) = P(W|S,L) \times P(S|Y,L) \times P(Y) \times P(L) \tag{7.4}$$

This can be further reduced, by considering every word is independent of each other to get:

$$P(L,Y,S,W) = \prod_i P(w_i|S,L) \times P(S|Y,L) \times P(Y) \times P(L) \tag{7.5}$$

The parameters in equation 7.5 can be inferred with using the parameter estimation procedures (EM and Gibbs sampling) already used in this thesis. But the challenge in this further exploration would lie in identifying a suitable dataset that has time-stamped raw text coming from different locales.

Appendix A

Appendix

A.1 Sense definitions

For the neologistic senses noted in section 6.3 excerpts from the OED are given below concerning this sense, giving the OED's definition of the sense and the first of its list of citations.

computer peripheral sense of 'mouse'

def: A small hand-held device which is moved over a flat surface to produce a corresponding movement of a pointer on a monitor screen or to delimit an area of the screen, and which usually has fingertip controls to select or initiate a computer function, or to place a cursor at the pointer's position.

cit: 1965 W. K. ENGLISH et al. *Computer-aided Display Control: Final Rep.* (Stanford Res. Inst.) 6 Within comfortable reach of the user's right hand is a device called the 'mouse' which we developed for evaluation..as a means for selecting those displayed text entities upon which the commands are to operate.

homosexual sense of 'gay'

def: orig. *U.S. slang*. (a) Of a person: homosexual; (b) (of a place, milieu, way of life, etc.) of or relating to homosexuals.

cit: 1941 G. LEGMAN *Lang. Homosexuality* in G. W. Henry Sex Variants II. 1167 Gay, an adjective used almost exclusively by homosexuals to denote homosexuality, sexual attractiveness, promiscuity..or lack of restraint, in a person, place, or party. Often given the French spelling, gai or gaie by (or in burlesque of) cultured homosexuals of both sexes.

(*note: the OED is somewhat tentative about this as the first citation and gives a number of citations prior to the above (1922–1941), but in parenthesis and with remarks that they do not feel they are conclusive.*)

industrial action sense of 'strike'

def: A concerted cessation of work on the part of a body of workers, for the purpose of obtaining some concession from the employer or employers.

cit: 1810 *Docum. Hist. Amer. Industrial Soc.* (1910) III. 370 The Society, in November 1809, ordered a general strike.

computer related sense of 'bit'

def: A unit of information derived from a choice between two equally probable alternatives or 'events'; such a unit stored electronically in a computer.

cit: 1948 C. E. SHANNON in *Bell Syst. Techn. Jrnl.* July 380 The choice of a logarithmic base corresponds to the choice of a unit for measuring information. If the base 2 is used the resulting units may be called binary digits, or more briefly bits, a word suggested by J. W. Tukey.

computer related sense of 'compile'

def: To produce (a machine-coded form of a program), orig. from existing subroutines but now from a source program in a high-level language; also, more commonly, to translate from a high-level source language into machine language, usually by means of a program written for the purpose.

cit: 1952 *Proc. Assoc. Computing Machinery* 1/2 UNIVAC compiled the program in one and one half minutes.

computer related sense of 'paste'

def: To insert (text or graphics) into a document by copying it from elsewhere in a single operation.

cit: 1975 *Business Week* 30 June 82 Hit a button called 'cut', and the word or paragraph disappears. Punch another button labeled 'paste' and the paragraph or word is inserted into the text where the pointer is located.

internet related sense of 'surf' has transitive and intransitive entries

def: *trans.* To visit successively (a series of Internet sites); to use (the Internet); to seek information about (a topic) on the Internet.

cit: 1992 *Re: Size Limits for Text Files?* in *alt.gopher* (Usenet newsgroup) 25 Feb. There is a lot to be said for..surfing the internet with gopher from anywhere that you can find a phone jack.

def: *intr.* To move from site to site on the Internet, esp. to browse or skim through web pages. Also: to go to a particular website.

cit: 1993 *San Francisco Chron.* 1 June c1/2 Millions of the world's most plugged-in people spend hours each week surfing at near-warp speed on a wave of information called the Internet.

computer related sense of 'boot'

def: To prepare (a computer) for operation by causing an operating system to be loaded into its memory from a disc or tape, esp. by a bootstrap routine; to cause (an operating system or a program) to be loaded in this way; to load the program on (a disc) into a computer's memory. Also to boot up.

cit: 1980 M. E. SLOAN *Introd. Minicomputers & Microcomputers* vi. 158 We turn the power knob to on, and depress the control and boot switches. We call this procedure booting the system ... The computer is now in the machine language mode, in which machine language programs can be entered and run.

music related sense of 'rock'

def: Originally (also with capital initial): = rock 'n' roll n. 2a. Now chiefly: a genre of popular music which evolved from rock 'n' roll during the mid to late 1960s, characterized by a strong beat, the use of the (esp. electric) guitar, and (esp. initially) musical experimentation, having a harsher sound than pop and often regarded as more serious or complex.

cit: 1956 *Daily Defender* (Chicago) 18 Dec. 14 Easily the most socksational exponent to hit the byways with 'Rock' is Presley.

intoxicated sense of 'stoned' has drink-related and drug-related sub-entries

def: Drunk, extremely intoxicated orig. U.S.

cit: 1952 *Life* 29 Sept. 67/2 Like boiled snails, bop jokes certainly are not everybody's dish, but those who acquire the taste for them feel cool, gone, crazy and stoned.

def: In a state of drug-induced euphoria, 'high'; also, incapacitated or stimulated by drugs, drugged. orig. U.S.

cit: 1953 H. J. ANSLINGER & W. F. TOMPKINS *Traffic in Narcotics* 315 Stoned, under the influence of drugs.

For the words discussed in section 6.6 having a neologistic sense which was undetected again below there are excerpts from the OED for the anticipated neologistic sense.

stylish sense of 'hip' OED identifies this was alternative form of *hep* and defers to that for the definition

def: slang (orig. U.S.). Well-informed, knowledgeable, 'wise to', up-to-date; smart, stylish.

cit: 1904 G. V. HOBART *Jim Hickey* i. 15 At this rate it'll take about 629 shows to get us to Jersey City, are you hip?

computer related sense of 'export'

def: To transmit (data) out of (part of) a computer for processing elsewhere.

cit: 1982 *Electronics* 10 Mar. 124/1 DJC allows the user of any work station to export a batch job to the NRM for remote execution.

data duplication sense of 'mirror' As first citation does not illustrate network-related usage, have included one of the later citations from the entry also

def: To write (data) on to two separate devices (esp. two hard disks) simultaneously, to protect against the possible failure of one; to create (a duplicate disk) in this manner. Also: to copy (a website) on to a different server

cit: 1993 *UNIX World May* 120/1 The data is striped to multiple controllers and then mirrored at each controller.

cit: 1999 *Sunday Times* 16 May 13 The message..had been 'mirrored' – copied onto other web sites.

network related sense of 'domain'

def: A subset of locations on the Internet or other network which share a common element of their IP address (indicating a geographical, commercial or other affiliation), or which are under the control of a particular organization or individual

cit: 1982 Z. SU & J. B. POSTEL *Request for Comments* (Network Working Group) (Electronic text) No. 819. 1 The name of a domain consists of a concatenation of one or more 'simple names'

intoxication sense of 'high'

def: Under the influence of, stimulated by, an illicit drug or drugs. Frequently with on.

cit: 1932 *Evening Sun* (Baltimore) 9 Dec. 31/4 High, under the influence of a narcotic.

Bibliography

Adams, Ryan Prescott and Mackay, David J.C. Bayesian online changepoint detection. arXiv preprint arXiv:0710.3742, 2007.

Agirre, Eneko and Edmonds, Philip. *Word Sense Disambiguation: Algorithms and Applications*. Springer Publishing Company, Incorporated, 1st edition, 2007. ISBN 1402068700, 9781402068706.

Akaike, Hirotugu. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19(6):716–723, December 1974.

Aminikhanghahi, Samaneh and Cook, Diane J. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367, 2017. ISSN 0219-3116. doi: 10.1007/s10115-016-0987-z. URL `http://dx.doi.org/10.1007/s10115-016-0987-z`.

Barnhart, David. A calculus for new words. *Dictionary*, 28:132–138, 2007.

Basseville, Michèle and Nikiforov, Igor V. *Detection of Abrupt Changes: Theory and Application*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993. ISBN 0-13-126780-9.

Biemann, ; C., ; Riedl, , and M., . Text: Now in 2d! a framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, 1(1):55–95, Apr 2013.

Biernacki, Christophe; Celeux, Gilles, and Govaert, Gérard. Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Comput. Stat. Data Anal.*, 41(3-4):561–575, January 2003. ISSN 0167-9473. doi: 10.1016/S0167-9473(02)00163-9. URL `http://dx.doi.org/10.1016/S0167-9473(02)00163-9`.

Bishop, Christopher M. *Pattern Recognition and Machine Learning*. Springer, February 2006.

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

Box, George E.P. and Tiao, George C. *Bayesian Inference in Statistical Analysis*. Wiley Classics Library, 1992.

Brody, Samuel and Lapata, Mirella. Bayesian word sense induction. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, page 103–111. Association for Computational Linguistics, 2009.

Burnard, Lou. *Reference Guide for the British National Corpus*. 2007. URL http://www.natcorp.ox.ac.uk/docs/URG/.

CHEN, Ming-Hui and SHAO, Qi-Man. Monte carlo estimation of bayesian credible and hpd intervals. *Journal of Computational and Graphical Statistics*, 8:69–92, 1999.

Choe, Do Kook and Charniak, Eugene. Naive bayes word sense induction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1433 – 1437. Association for Computational Linguistics, October 2013. URL http://www.aclweb.org/anthology/D13-1148.

Clear, Jeremy H. The digital word. chapter The British National Corpus, pages 163–187. MIT Press, Cambridge, MA, USA, 1993. ISBN 0-262-12176-x. URL http://dl.acm.org/citation.cfm?id=166403.166418.

Cook, Paul and Stevenson, Suzanne. Automatically identifying changes in the semantic orientation of words. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 28–34, Valletta, Malta, May 2010.

Cook, Paul; Lau, Jey Han; Rundell, Michael; McCarthy, Diana; , and Baldwin, Timothy. A lexicographic appraisal of an automatic approach for detecting new word senses. In *Proceedings of eLex 2013*, 2013.

Cook, Paul; Lau, Jey Han; McCarthy, Diana, and Baldwin, Timothy. Novel word-sense identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, page 1624–1635. ACL, August 2014. URL http://anthology.aclweb.org/C/C14/C14-1154.pdf.

Csörgö, M. and Horváth, L. *Limit theorems in change-point analysis*. Wiley series in probability and statistics. Wiley, 1997. ISBN 9780471955221. URL https://books.google.ie/books?id=iyXvAAAAMAAJ.

Davies, Mark. The corpus of historical american english: 400 million words, 1810-2009, 2010. available online at http://corpus.byu.edu/coha.

Dempster, A.P.; Laird, N.M., and Rubin, D.B. Maximum likelihood from incomplete data via the em algorithm. *J. Royal Statistical Society*, B 39:1–38, 1977.

Emms, Martin. Dynamic EM in neologism evolution. In Yin, Hujun; Tang, Ke; Gao, Yang; Klawonn, Frank; Lee, Minho; Weise, Thomas; Li, Bin, and Yao, Xin, editors, *Proceedings of IDEAL 2013*, volume 8206 of *Lecture Notes in Computer Science*, pages 286–293. Springer, 2013.

Emms, Martin and Jayapal, Arun. Detecting change and emergence for multiword expressions. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 89–93, Gothenburg, Sweden, 2014. Association for Computational Linguistics.

Emms, Martin and Jayapal, Arun. An unsupervised em method to infer time variation in sense probabilities. In *ICON 2015: 12th International Conference on Natural Language Processing*, pages 266–271, Trivandrum, India, December 2015.

Emms, Martin and Jayapal, Arun. Dynamic generative model for diachronic sense emergence detection. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, page to be published, Osaka, Japan, December 2016.

Fellbaum, Christiane. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

Ferraresi, Adriano; Zanchetta, Eros; Baroni, Marco, and Bernardini, Silvia. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *In Proceedings of the 4th Web as Corpus Workshop (WAC-4*, 2008.

Firth, J. *A synopsis of linguistic theory 1930-1955*. Studies in Linguistic Analysis, Philological. Longman, 1957.

Fisher, R. A. *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh, second edition, 1959.

Frermann, Lea and Lapata, Mirella. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45, 2016. ISSN 2307-387X. URL `https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/796`.

Gelfand, Alan E. and Smith, Adrian F. M. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, jun 1990. ISSN 0162-1459 (print), 1537-274X (electronic).

Goldberg, Yoav and Orwant, Jon. A dataset of syntactic-ngrams over time from a very large corpus of english books. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 241–247, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

Granjon, Pierre. The cusum algorithm - a small review. Technical Report, 2013.

Griffiths, Thomas L. and Steyvers, Mark. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004. doi: 10.1073/pnas.0307752101. URL `http://www.pnas.org/content/101/suppl_1/5228.abstract`.

Gulordava, Kristina and Baroni, Marco. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, GEMS '11, pages 67–71, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-16-9. URL http://dl.acm.org/citation.cfm?id=2140490.2140498.

Hall, David; Jurafsky, Daniel, and Manning, Christopher D. Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 363–371, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1613715.1613763.

Harman, Donna and Liberman, Mark. Tipster complete ldc93t3a, 1993. Available:DVD.

Heinrich, Gregor. Parameter estimation for text analysis. Technical report, 2004.

Huang, Jonathan. Maximum likelihood estimation of dirichlet distribution parameters. Technical report, Stanford University, 2005. URL http://web.stanford.edu/~jhuang11/research/dirichlet/dirichlet.pdf.

Ide, Nancy and Véronis, Jean. Introduction to the special issue on word sense disambiguation: The state of the art. *Comput. Linguist.*, 24(1):2–40, March 1998. ISSN 0891-2017. URL http://dl.acm.org/citation.cfm?id=972719.972721.

Jin, Peng; Zhang, Yihao, and Sun, Rui. Lstc system for chinese word sense induction. In *CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 2010. URL http://aclweb.org/anthology/W10-4167.

Kilgarriff, Adam; Rychly, Pavel; Smrz, Pavel, and Tugwell, David. The sketch engine. In *Proceedings of EURALEX*, 2004.

Kim, Yoon; Chiu, Yi-I.; Hanaki, Kentaro; Hegde, Darshan, and Petrov, Slav. Temporal analysis of language through neural language models. *CoRR*, abs/1405.3515, 2014. URL http://arxiv.org/abs/1405.3515.

Koehn, Philipp. Europarl: A parallel corpus for statistical machine translation. Online, 2005. URL http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdfx.

Kulkarni, Vivek; Al-Rfou, Rami; Perozzi, Bryan, and Skiena, Steven. Statistically significant detection of linguistic change. *CoRR*, abs/1411.3315, 2014. URL http://arxiv.org/abs/1411.3315.

Lau, Jey Han; Cook, Paul; McCarthy, Diana; Newman, David, and Baldwin, Timothy. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 591–601, Avignon, France, April 2012.

Lorden, G. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908, 1971. ISSN 00034851. URL http://www.jstor.org/stable/2240115.

Michel, Jean-Baptiste; Shen, Yuan Kui; Aiden, Aviva Presser; Veres, Adrian; Gray, Matthew K.; Team, The Google Books; Pickett, Joseph P.; Hoiberg, Dale; Clancy, Dan; Norvig, Peter; Orwant, Jon; Pinker, Steven; Nowak, Martin A., and Aiden, Erez Lieberman. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011. doi: 10.1126/science.1199644. URL http://www.sciencemag.org/content/331/6014/176.abstract.

Mikolov, Tomas; Chen, Kai; Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. URL http://arxiv.org/abs/1301.3781.

Mimno, David; Wallach, Hanna, and McCallum, Andrew. Gibbs sampling for logistic normal topic models with graph-based priors. In *NIPS Workshop on Analyzing Graphs*, 2008.

Mitra, Sunny; Mitra, Ritwik; Riedl, Martin; Biemann, Chris; Mukherjee, Animesh, and Goyal, Pawan. That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, page 1020–1029. Association for Computational Linguistics, June 2014.

Mitra, Sunny; Mitra, Ritwik; Maity, Suman Kalyan; Riedl, Martin; Biemann, Chris; Goyal, Pawan, and Mukherjee, Animesh. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, 21(5):773–798, 2015. URL http://dblp.uni-trier.de/db/journals/nle/nle21.html#MitraMM0BGM15.

Page, E. S. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527, 1955. ISSN 00063444. URL http://www.jstor.org/stable/2333401.

Renouf, Antoinette; Kehoe, Andrew, and Banerjee, Jay. 'webcorp: an integrated system for web text search. In *Corpus Linguistics and the Web*, University of Birmingham, 2007. Research and Development Unit for English Studies, University of Birmingham.

Resnik, Philip and Hardisty, Eric. Gibbs sampling for the uninitiated. Lamp-tr-153, University of Maryland, June 2010.

Rohrdantz, Christian; Hautli, Annette; Mayer, Thomas; Butt, Miriam; Keim, Daniel A., and Plank, Frans. Towards tracking semantic change by visual analytics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 305–310. ACL, June 2011.

Rue, Harvard and Held, Leonhard. *Gaussian Markov random fields: Theory and applications.* 2005.

Rychlý, Pavel and Kilgarriff, Adam. An efficient algorithm for building a distributional thesaurus (and other sketch engine developments). In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 41–44, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1557769.1557783.

Schütze, Hinrich. Automatic word sense discrimination. *Comput. Linguist.*, 24(1):97–123, mar 1998. ISSN 0891-2017. URL http://dl.acm.org/citation.cfm?id=972719.972724.

Sheidlower, Jesse T. Principles for the inclusion of new words in college dictionaries. *Dictionaries*, 16:32–43, 1995.

Simpson, John. Preface to the third edition of the oed, 2000. public.oed.com/the-oed-today/preface-to-the-third-edition-of-the-oed.

Sprott, D.A. *Statistical Inference in Science*. Springer Series in Statistics. Springer New York, 2000. ISBN 9780387950198. URL https://books.google.ie/books?id=3cGefthoP9gC.

Stephens, Matthew. Dealing with label switching in mixture models. In *Journal of the Royal Statistical Society*, volume 62 of *B (Statistical Methodology)*, pages 795 – 809. University of Oxford, UK, Blackwell Publishing for the Royal Statistical Society, 2000. URL https://stat.duke.edu/~scs/Courses/Stat376/Papers/Mixtures/LabelSwitchingStephensJRSSB.pdf.

Tang, Xuri; Qu, Weiguang, and Chen, Xiaohe. Semantic change computation: A successive approach. In *Business Media New York 2015*. Springer Science, 2015.

Teh, Yee Whye; Jordan, Michael I.; Beal, Matthew J., and Blei, David M. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101, 2004.

Tournier, Jean. *Introduction descriptive à la lexicogénétique de l'anglais contemporain*. Champion-Slatkine, 1985.

Véronis, J. Review of "Polysemy - Theoretical and computational approaches" by Yael Ravin and Claudia Leacock. *Computational Linguistics*, forthcoming, 2002.

Wang, Xuerui and McCallum, Andrew. 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 '06, pages 424–433, New York, NY, USA, 2006. ACM. ISBN 1-59593-339-5. doi: 10.1145/1150402.1150450. URL http://doi.acm.org/10.1145/1150402.1150450.

Wijaya, Derry Tanti and Yeniterzi, Reyyan. Understanding semantic change of words over centuries. In *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web*, DETECT '11, pages 35–40, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0962-2. doi: 10.1145/2064448.2064475. URL http://doi.acm.org/10.1145/2064448.2064475.

Wikipedia, . Computer mouse: Wikipedia, the free encyclopedia, 2016. URL https://en.wikipedia.org/wiki/Computer_mouse#In_the_marketplace. [Online; accessed 20-Aug-2016].

Yang, Miin-Shen; Lai, Chien-Yo, and Lin, Chih-Ying. A robust em clustering algorithm for gaussian mixture models. *Pattern Recogn.*, 45(11):3950–3961, November 2012. ISSN 0031-3203. doi: 10.1016/j.patcog.2012.04.031. URL http://dx.doi.org/10.1016/j.patcog.2012.04.031.

Yao, Xuchen and Durme, Benjamin Van. Nonparametric bayesian word sense induction. In *Proceedings of the TextGraphs-6 Workshop*, pages 10–14. Association for Computational Linguistic, June 2011.

Index

ablation, 79, 105, 123
AQUAINT, *see* corpus

bag of words, 37
Bayesian approach, 24
beta distribution
    see prior distribution, 25
bi-gram model, 37
bit, *see* targets
boot, *see* targets
burn-in, 49
byte-ostensible, *see* targets

clustering, 15
COCA, *see* corpus
COHA, *see* corpus
compile, *see* targets
context words, 12
corpus
    AQUAINT, 71
    BNC, 18, 74
    COCA, 72
    COHA, 72
    EUROPARL, 71
    Google 5-gram, 73, 79, 80, 118, 119, 121, 131, 134, 135
    Google n-gram, 71, 74
    Lexis-Nexis, 73
    NYT, 71
    TIPSTER, 71
    UKWAC, 74
    ukWac, 18
    WebCorp live, 72
credible interval, 53
    HPD, 53, 92

diachronic model, 32, 36, 37, 50, 52, 53, 57, 69, 79, 106, 123, 131, 133, 135
    hyper-parameters, 36
    mean, 45, 51
    plate diagram, 35, 36
    prior, 36
dictionary first inclusion, 135
Dirichlet distribution
    see prior distribution, 25
downloadable resource, 69, 71

EM, *see* parameter estimation
emergence date, 4, 67, 100
emerging sense, 79
empirical estimate, 81, 83
EUROPARL, *see* corpus

frequency change, 6
    opinion change, 7, 65
    world change, 7, 65

gay, *see* targets
generative model, 16, 135
genocide-ostensible, *see* targets
Gibbs sample, 46, 51, 52
Gibbs sampler, 45
gist words, 65, 66, 82, 89, 91, 95, 99, 101, 103, 106, 120, 121, 127
Google 5-gram, *see* corpus
Google n-gram, 6, 15, 33, *see also* corpus
Google search, 2

hierarchical clustering, 125
HPD interval, *see* credible interval

hyper-parameters, *see* prior distribution, *see also* diachronic model

KL-div, 99, 125

label switching, 50, 52, 53
Lagrange multipliers, 39
language change, 1, 6, 65
lexical change, 1, 3
  amelioration, 3
  broadening of word sense, 3, 16
  narrowing of word sense, 3, 16
  pejoration, 3
Lexis-Nexis, *see* corpus

MAP, *see* parameter estimation
MAP estimate, *see* parameter estimation
mean, *see* diachronic model, *see* parameter estimation
merge test, 79, 122, 123, 131
mouse, *see* targets
multi-nomial, *see* prior distribution

Naive Bayes model, 37
neologism, 116
  formal neologism, 1, 4
  semantic neologism, 1–4, 69, 70, 72, 79, 80, 135
neologism sense, 92
neologistic sense, 4, 57, 59, 60, 66, 67, 79, 83, 85, 86, 89, 91, 92, 94, 95, 97, 98, 100–103, 117, 122, 134, 135
non-neologism, 106, 116
novelty score, 114, 115
novelty-score, 79
NYT, *see* corpus

opinion change, *see* frequency change
Oxford English Dictionary
  OED, 18, 19, 58, 65, 80, 81, 97, 99, 105, 117, 134

parameter estimation, 23
  Expectation Maximization, 24, 31, 38, 45, 46, 49, 79–82, 86, 89, 90, 92, 94, 95, 98, 100, 102, 104, 117–119, 121, 123, 125, 134
  Gibbs sampling, 24, 31, 45, 46, 49, 53, 79–82, 86, 89, 90, 92, 94, 95, 98, 100, 102, 104, 134
  Maximum A Posteriori, 24, 44, 49
  Maximum Likelihood Estimate, 23, 38, 49, 134
  mean, 24, 25, 134
  mode, 24, 27
  non-informative prior, 24
  prior, 24
plate diagram, *see* diachronic model
point estimate, 46
posterior, 24, 28, 45, 48, 53
prior distribution
  beta, 25, 27
  conjugate prior, 28
  Dirichlet, 13, 24, 25, 27–29, 36, 48
  multi-nomial, 28, 36
  symmetric prior, 50
pseudo-neologism, 22, 79, 80, 83, 84, 122
pseudo-word, 22, 80

rock, *see* targets

sense emergence, 16, 59, 60, 85, 89, 96, 117, 135
sense granularity
  homonym, 7, 11
  polysemy, 7, 11
static model, 36, 37
Statistical Machine Translation
  SMT, 4, 71
stoned, *see* targets
strike, *see* targets
supermarket-ostensible, *see* targets
surf, *see* targets
symmetric prior, *see* prior distribution

targets
bit, 84, 85, 92
boot, 84, 85, 100, 123
byte-ostensible, 81
cinema, 106
compile, 84, 85
domain, 119
export, 118
gay, 84, 85, 89
genocide-ostensible, 82
high, 121
hip, 117
mirror, 119
mouse, 84–86, 123
ostensible, 106
paste, 95
plant, 106
play, 106
present, 106
promotion, 106
rock, 84, 85, 102
spirit, 106
stoned, 84, 85, 104
strike, 84, 85, 90
supermarket-ostensible, 83
surf, 84, 85, 125, 128
theatre, 106
time-stamped data, 69, 71
TIPSTER, *see* corpus
topic model, 13, 16, 18
HDP, 128
LDA, 12, 18, 128
TOT, 14
tracks, 85, 86, 88, 89, 92, 94, 98, 100, 102, 105, 118, 119, 134
tracks-plot, 58–60, 92

uni-gram model, 37

web-accessible resource, 69, 72
WebCorp live, *see* corpus
word sense disambiguation
WSD, 11
word sense induction
word sense discrimination, 80
WSI, 4, 11, 13
Wordnet synset, 15
world change, *see* frequency change