# Dynamic Generative model for Diachronic Sense Emergence Detection

**Martin Emms**
Dept of Computer Science
Trinity College, Dublin
Ireland
martin.emms@scss.tcd.ie

**Arun Jayapal**
Dept of Computer Science
Trinity College, Dublin
Ireland
jayapala@scss.tcd.ie

## Abstract

As time passes words can acquire meanings they did not previously have, such as the 'twitter post' usage of 'tweet'. We address how this can be detected from time-stamped raw text. We propose a generative model with senses dependent on times and context words dependent on senses but otherwise eternal, and a Gibbs sampler for estimation. We obtain promising parameter estimates for positive (resp. negative) cases of known sense emergence (resp non-emergence) and adapt the 'pseudo-word' technique (Schütze, 1998) to give a novel further evaluation via 'pseudo-neologisms'. The question of ground-truth is also addressed and a technique proposed to locate an emergence date for evaluation purposes.

## 1 Introduction

It is widely noted that a single word can have several senses. The diachronic aspect of this is that the set of senses possessed by a word changes over time. (1a) and (1b) below illustrate this:

(1) (a) *she was a* **gay** *little soul, enjoying everything and always trilling with laughter* (1905)
(b) *applying heightened scrutiny to discrimination against* **gay** *men and lesbians* (1990)
(a′) *sie war ein* **Homosexuell** *kleine Seele, alles zu genießen und immer rollen vor Lachen*

(1a), from 1905, illustrates a 'being happy' sense of *gay*, while (1b), dating from 1990, illustrates a 'homosexual' sense, a sense which the word did not possess in 1905, and came to possess at some time since. The advent of a new sense for an existing form is sometimes called a *semantic* neologism (Tournier, 1985), in contrast to the simpler *formal* neologism, where simply a new form arrives (eg. *selfie*). The concern of this paper is to propose an unsupervised algorithm for detecting semantic neologisms, an algorithm which can be given time-stamped but plain-text examples of a particular word and detect whether (and when) the word gained a sense.

Information about such lexical change could be useful to NLP tasks. For example, if a SMT system is trained on data from particular times and is to be applied to texts from different times, either later or earlier, advance warning of sense changes could be of use. To illustrate, (1a′) gives the English→German translation via Google translate[1] of (1a), mistranslating the 1905 usage of *gay* as *Homosexuell*, probably due to the newer sense predominating in training data.

We will propose a diachronic sense model where a target's sense is conditioned on time and the context words are conditioned just on the target's sense, and not the time. We use the Google n-gram data set (Michel et al., 2011) which provides time-stamped data but *no* sense information and develop a Gibbs sampling algorithm (Gelfand and Smith, 1990) to estimate the parameters in an unsupervised fashion. We will show that the algorithm is able to provide an accurate date of sense emergence (true positives), and also to detect the absence of sense emergence when appropriate (true negatives). We adapt also the 'pseudo-word' technique first proposed by Schütze (1998) to give a further means of algorithm assessment. We also make a number of points concerning difficulties and possibilities evaluating such a sense-emergence system.

[1] Executed on Apr 13, 2016





*Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1362–1373, Osaka, Japan, December 11-17 2016.

## 2 A Diachronic Sense Model

Assume we have a corpus of $D$ *time-stamped* occurrences of some particular target expression $\sigma$, with each time-stamp shared by many items. Let $\boldsymbol{w}$ be the sequence of words in a window around $\sigma$, and $Y$ be its time-stamp, ranging from 1 to $N$. Assume $\sigma$ exhibits $K$ different senses and let $S$ be the sense of a particular item. $S$ will be treated as a hidden variable: the actual data provides values just for $Y$ and $\boldsymbol{w}$. We will now propose a particular model of the joint probability $p(Y, S, \boldsymbol{w})$.

It may be factorised as $P(Y) \times P(S|Y) \times p(\boldsymbol{w}|S, Y)$ without loss of generality. We now make two independence assumptions: (a) that conditioned on $S$ the words $\boldsymbol{w}$ are independent of $Y$, so $p(\boldsymbol{w}|S, Y) = P(\boldsymbol{w}|S)$ and (b) that conditioned on $S$ the words are independent of each other, so $P(\boldsymbol{w}|S) = \prod_i p(w_i|S)$. With these assumptions the equation for a single data item is

$$P(Y, S, \boldsymbol{w}; \boldsymbol{\pi}_{1:N}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\tau}_{1:N}) = P(Y; \boldsymbol{\tau}_{1:N}) \times P(S|Y; \boldsymbol{\pi}_{1:N}) \times \prod_i p(w_i|S; \boldsymbol{\theta}_{1:K}) \quad (2)$$

where we have also introduced explicit notations for model parameters: (i) for every time $t$, $\boldsymbol{\pi}_t$ is a length $K$ vector of sense probabilities (ii) for every sense $k$, $\boldsymbol{\theta}_k$ is a length $V$ vector of context word probabilities for target sense $k$ — $V$ is the size of the vocabulary encountered in all the data and (iii) for every $t$, $\tau_t$ is a 'time' probability reflecting simply the abundance of target $\sigma$ at $t$.

The fact that for every time $t$ there is a parameter $\boldsymbol{\pi}_t$ directly models that a sense's likelihood varies temporally; for an *established* sense this variation might be due to changes in the world and in people's concerns, but for an *emergent* sense, part of the variation represents genuine *language* change and for some $k$, for some early range of times, $\pi_t[k]$ should ideally be *zero*. Assumption (a) reflects an expectation that the vocabulary co-occuring with a particular *sense* is comparatively time independent. This particular time-independence is perhaps plausible but certainly is not absolute and its viability as an assumption can only be confirmed or disconfirmed by our later experiments.

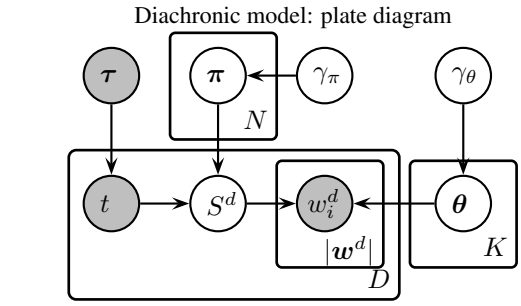
To develop the model further to the level of a corpus and incorporate parameter priors, let $\boldsymbol{t}^{1:D}, \boldsymbol{s}^{1:D}, \boldsymbol{w}^{1:D}$ be the values of $Y$, $S$ and $\boldsymbol{W}$ on $D$ items, and let the $\boldsymbol{\pi}_t$ sense probability vectors have a $K$-dimensional Dirichlet prior with parameter $\gamma_\pi$ and the $\boldsymbol{\theta}_k$ word probability vectors have a $V$-dimensional Dirichlet prior with parameter $\gamma_\theta$. We consider the joint probability $P(\boldsymbol{t}^{1:D}, \boldsymbol{s}^{1:D}, \boldsymbol{w}^{1:D}, \boldsymbol{\pi}_{1:N}, \boldsymbol{\theta}_{1:K}; \boldsymbol{\gamma}_\pi, \boldsymbol{\gamma}_\theta, \boldsymbol{\tau}_{1:N})$ and its formula under our assumptions is equation (3) in Figure 1, above which the model is depicted as a plate diagram.

From a generative perspective, the $\boldsymbol{\pi}_{1:N}$ and $\boldsymbol{\theta}_{1:K}$ are generated from $N$ and $K$ samplings and then used for *all* $D$ data items. In actual data, the time-stamps and words are known, so for any fixed setting of $\boldsymbol{\gamma}_\pi, \boldsymbol{\gamma}_\theta$ there is a defined *posterior* distribution on $\boldsymbol{s}^{1:D}, \boldsymbol{\pi}_{1:N}, \boldsymbol{\theta}_{1:K}$. A Gibbs sampling algorithm (Gelfand and Smith, 1990) can be derived, generating a large set of samples of this $(D + N + K)$-tuple, representative of this posterior, from which *mean* values of the model parameters $\boldsymbol{\pi}_{1:N}$ and $\boldsymbol{\theta}_{1:K}$ can be derived. To arrive at the Gibbs sampler, sampling distributions for $s^d$, $\boldsymbol{\pi}_t$ and $\boldsymbol{\theta}_k$ are needed, in each case conditioned on all *other* parts of the sample tuple. The formulae for these sampling distributions are shown as (4 – 6) in Figure 1 : in these formulae $\mathbb{S}_t[k]$ is the number of data items with time-stamp $t$ and sampled sense $k$ and $\mathbb{V}_k[v]$ is the number of times word $v$ occurs in data items with sampled sense $k$. The derivation of these formulae is relatively straightforward given well-known conjugacy properties of Dirichlet priors – Appendix A gives an outline derivation for $\boldsymbol{\pi}_t$. The sampling algorithm is given in pseudo-code in Figure 1.

### 2.1 Ground truth for semantic neologisms?

Given a large-scale, time-stamped and *sense*-labeled corpus for a target expression $\sigma$, it would be easy to determine a true emergence date — call it $C_0$ — at which a new sense for $\sigma$ first departed from zero frequency (and continued to climb from zero). It has been noted (Lau et al., 2012; Cook et al., 2013) that such reference corpora do not exist, thus posing the question of what can serve as ground truth instead.

One option, sometimes adopted though far from ideal, is simple speaker intuition, which is subjective, of low temporal resolution and at best applicable to recent innovations. It is natural to consider dictionaries for something better. Form/meaning pairings are added to dictionaries at some particular time, so inspecting a series of dictionary editions, though labour intensive, can give a *first inclusion* date – call




$$P(\boldsymbol{t}^{1:D}, \boldsymbol{s}^{1:D}, \boldsymbol{w}^{1:D}, \boldsymbol{\pi}_{1:N}, \boldsymbol{\theta}_{1:K}; \boldsymbol{\gamma}_\pi, \boldsymbol{\gamma}_\theta, \boldsymbol{\tau}_{1:N}) = \prod_t \text{Dir}(\boldsymbol{\pi}_t; \boldsymbol{\gamma}_\pi) \times \prod_k \text{Dir}(\boldsymbol{\theta}_k; \boldsymbol{\gamma}_\theta) \times \prod_d [\, P(t^d; \boldsymbol{\tau}_{1:N}) \times P(s^d | t^d; \boldsymbol{\pi}_{1:N}) \times \prod_i P(w_i^d | s^d; \boldsymbol{\theta}_{1:K}) \,] \quad (3)$$

$$P(s^d) = \frac{\pi_{t,k} \prod_{i=1}^{|\boldsymbol{w}^d|} \theta_{k,w_i^d}}{\sum_{k'} \pi_{t,k'} \prod_{i=1}^{|\boldsymbol{w}^d|} \theta_{k',w_i^d}} \quad (4)$$

$$P(\boldsymbol{\pi}_t) = \text{Dir}(\boldsymbol{\pi}_t; \boldsymbol{\gamma}_\pi + \mathbb{S}_t) \quad (5)$$

$$P(\boldsymbol{\theta}_k) = \text{Dir}(\boldsymbol{\theta}_k; \boldsymbol{\gamma}_\theta + \mathbb{V}_k) \quad (6)$$

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**
    $k \sim P(s^d)$ see (4);
    $S[d] = k$;
    $\mathbb{S}[t^d][k]$ += 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**
    $\boldsymbol{\pi}_t \sim P(\boldsymbol{\pi}_t)$ see (5);
  **end**
  **for** *k:=1 to K* **do**
    $\boldsymbol{\theta}_k \sim P(\boldsymbol{\theta}_k)$ see (6);
  **end**
**end**

**Algorithm 1:** Gibbs sampling estimation

Figure 1: From top left in anti-clockwise shows: plate digram for diachronic model, Gibbs sampler updates, and pseudo-code for Gibbs sampler

this $D_0^i$. The time resolution of this is low and the subtle criteria involved in inclusion decisions make it a non-ideal approximation of $C_0$ (Sheidlower, 1995; Simpson, 2000; Barnhart, 2007). Some researchers (Lau et al., 2012) use this, though we will not. More accessibly, an historically oriented dictionary (eg. the Oxford English Dictionary (OED)) strives to include the earliest known use of a word in a particular sense — the so-called *earliest citation*. If we call this $D_0^c$, it seems to makes sense to use $D_0^c$ as a *lower* bound on $C_0$, and we will do this. $D_0^c$ represents a first use, which might be followed by a long interlude before the usage is really taken up: the experiments in Section 3.2 will highlight examples of this.

We propose to use a different technique to establish $C_0$ more precisely. If there are words which it is intuitive to expect in the vicinity of a target word $\sigma$ in the novel sense, and not in other senses, then by consulting a time-stamped corpus one should see the probability of finding these words in $\sigma$'s context start to climb at a particular time. For example, *mouse* has come to have a 'computer pointing device' sense, and in this usage it is intuitive to expect words like *click*, *button*, *pointer* and *drag* in it's context. For any word $w$ and target $\sigma$ let $P_t(w|\sigma)$ be $w$'s probability of occurring in $\sigma$'s context in data from time $t$, and let $track_\sigma(w)$ to be the sequence these values. If when $track_{mouse}(w)$ is plotted for the above words, they all show a sharp increase at the *same* time point, this is good evidence that this is $C_0$ – the right-hand plot in Figure 3(a) is an example of this. This combines co-occurrence intuitions with corpus data, and does *not* rely on somewhat unreliable speaker intuitions of recency. To forestall any possible confusion, this procedure of inspecting the probabilities of words thought especially associated with a particular sense is *not* being advanced as a proposed unsupervised algorithm to locate sense emergences. It is advanced as a way to establish a ground-truth concerning emergence by which to evaluate our proposed unsupervised algorithm.

## 3 Data and Experiments

In the experiments reported below we use the Google N-gram dataset (Michel et al., 2011). This is a data-set based on Google's digitized publication holdings and it provides per-year counts of $n$-grams, for

| Target | Years | Lines | New sense | OED | Tracks | GS-Date | $< 10\%$ |
|---|---|---|---|---|---|---|---|
| mouse | 1950-2008 | 910k | computer pointing device | 1965 | 1983 | 1982 | yes |
| gay | 1900-2008 | 1253k | homosexual person | 1922 | 1966 | 1969 | yes |
| strike | 1800-2008 | 5052k | industrial action | 1810 | 1880 | 1866 | yes |
| bit | 1920-2008 | 7393k | basic unit of information | 1948 | 1965 | 1954 | no |
| paste | 1950-2008 | 318k | duplicate text in computer edit | 1975 | 1982 | 1981 | yes |
| compile | 1950-2008 | 689k | transform to machine code | 1952 | 1966 | 1971 | yes |
| surf | 1950-2008 | 182k | exploring internet | 1992 | 1994 | 1993 | yes |
| boot | 1920-2008 | 1285k | computer start up | 1980 | 1980 | 1984 | yes |
| rock | 1920-2008 | 4136k | genre of music | 1956 | 1965 | 1965 | yes |
| stoned | 1930-2008 | 79k | under drug influence | 1952 | 1970 | 1979 | no |

| Target | Years | Lines |
|---|---|---|
| ostensible | 1800-2008 | 130k |
| present | 1850-2008 | 56333k |
| cinema | 1950-2008 | 305k |
| promotion | 1930-2008 | 1681k |
| theatre | 1950-2008 | 1125k |
| play | 1950-2008 | 13726k |
| plant | 1900-2008 | 8175k |
| spirit | 1930-2008 | 11573k |

Table 1: Google 5 gram dataset - the left table provides the information for targets that are neologisms while the right one has the targets for non-neologisms – see text for explanation of columns

$1 \leq n \leq 5$: so for any given $n$-gram, $\boldsymbol{x}$, and any year, $t$, it gives the total frequency[2] of occurrences of $\boldsymbol{x}$ across all books dated to year $t$. For a given target word $\sigma$ we use the subset of all the data consisting of the *5*-grams that contain $\sigma$; we use the 5-grams as they provide most context around the target $\sigma$. For the target *mouse* the following is an example of a line of data from the corpus

```
Enter or click the mouse 1990 9 7
```

The first of the final three numbers is a year. The penultimate number is the number of occurrences of the 5-gram in all publications from that year – this is the significant count data for the algorithm. The final number is the number of publications from the year that contain the 5-gram, which is not significant for the algorithm.

For the experiments reported in Section 3.2 two sets of targets were chosen. The first set {*mouse, gay, strike, bit, paste, compile, surf, boot, rock, stoned*} are words which, relative to particular time periods, are known to exhibit sense emergence. The second set {*ostensible, present, cinema, promotion, theatre, play, plant, spirit*} are words which, relative to particular time periods, are thought *not* to exhibit sense emergence. Following Lau et al. (2012) the idea is that these should provide both positive and negative tests for the algorithm. Table 1 lists the targets. For each target, the 'Years' and 'Lines' columns give the range of years used and the total number of 5-gram lines of data for that year-range. For the positive targets the 'New sense' column gives an indication of the emergent sense and the next two columns give two kinds of reference dating information – see Section 2.1 – the OED first citation date and the 'tracks'-based date that is apparent from 'tracks' plots for words that are intuitively associated with the emergent sense (the right-hand plot in Figure 3(a) is an example). The 'GS date' column gives the emergence date inferred when the inference algorithm was run and will be discussed further in Section 3.2.

Before describing the experiments it is necessary to emphasize the Google n-gram data-set is best thought of as a frequency table giving per-year counts associated with 5-gram *types*. It is not really a corpus of text tokens. For brevity Algorithm 1 was formulated assuming that each data item represented a single target *token*. Any original publication token of a target $\sigma$ could have contributed to several different 5-gram *type* counts (up to 5) but the data-set makes it impossible to know to what extent this is so. We therefore effectively treat each 5-gram data entry $d$ listed with frequency of $n_d$ as if it derives from $n_d$ tokens of $\sigma$ which contributed to no other 5-gram counts. This leads to changing the count increment operations in Algorithm 1 to add $n_d$ rather than 1, that is, $\mathbb{S}[t][k]{+}{=}n_d$ and $\mathbb{V}[k][w]{+}{=}n_d$.

For all of the experiments sampling is done according to Algorithm 1, for 10000 iterations, with the first 1000 discarded as 'burn-in' samples and then means are determined for the model parameters $\boldsymbol{\pi}_{1:N}$ (sense-given-year) and $\boldsymbol{\theta}_{1:K}$ (word-given-sense) from the sampled values. The parameters $\boldsymbol{\gamma}_\pi$ and $\boldsymbol{\gamma}_\theta$ of the Dirichlet priors are set to have 1 in all positions to make them non-informative priors so uniform over all possible $\boldsymbol{\pi}_t$ and $\boldsymbol{\theta}_k$. The sampler is initialised with values $\boldsymbol{\pi}_{1:N}$ and $\boldsymbol{\theta}_{1:K}$ in the following way. Let $P_{corp}$ be the observable corpus word probabilities in $\boldsymbol{w}^{1:D}$. Each $\boldsymbol{\theta}_k$ is set to $(1-\alpha)P_{corp} + \alpha P_{ran}$, where $P_{ran}$ is a random word distribution and $\alpha$ is a mixing proportion, here set to $10^{-1}$. The $\pi_t$ are set to some shared set of sense probabilities. Thus initially the word distributions for each sense $k$ are almost identical, and the sense distributions are the same at all times, so far from the neologism situation.

[2]They exclude 5-grams with total count $< 40$.

The procedure was implemented in C++. To obtain the code or data see `www.scss.tcd.ie/Martin.Emms/SenseDynamics`.

### 3.1 Experiments with 'pseudo'-neologisms

The '*pseudo-word*' technique was introduced in Schütze (1998) as a possible means to test unsupervised word-sense discrimination. It can be given a diachronic twist to furnish what might be called '*pseudo-neologisms*' in the following way. Relative to some period of time select two words, $\sigma_1$ and $\sigma_2$, both unambiguous, with $\sigma_1$ in use throughout the time period, but with $\sigma_2$ first emerging at some point, $t_e$, in the period. If the 5-grams for $\sigma_1$ and $\sigma_2$ are then all treated as examples of the fake word '$\sigma_1$-$\sigma_2$' this functions as an artifical *semantic* neologism, manifesting $\sigma_2$'s sense only from $t_e$ onwards. Furthermore, if we say $f_t(\sigma_i)$ gives the true empirical probability of target $\sigma_i$ in pooled $\sigma_1$,$\sigma_2$ data for time $t$, then ideally the outcome of inference when run for $K = 2$ should be that for each $k$, the trajectory of the $\pi_t[k]$ values is very similar to that for one of the $f_t(\sigma_i)$. We tested this, for the time-period 1850–2008, with 'ostensible' for $\sigma_1$ (present throughout), and 'supermarket', 'genocide' and 'byte' as possibilities for $\sigma_2$ (which emerged as new words over this time frame) and indeed obtained the desired correspondence between inferred $\pi_t[k]$ and empirical $f_t(\sigma_i)$ trajectories – Figure 2 shows the outcomes for the first two. For the first case the succession of $\pi_t[1]$ values matches closely the succession of $f_t$('supermarket') values, and in the second case the $\pi_t[0]$ values match the $f_t$('genocide') values. To get an insight into the inferred $\boldsymbol{\theta}_k$ values, we defined $gist(S)$ to be the top 20 words when ranked according to the ratio of $P(w|S)$ to $P_{corp}(w)$. For the apparently neologistic sense $S$, Figure 2 also shows $gist(S)$ and it can be seen that these sets of words seem very consistent with relevant parts of the pseudo-neologisms.

Thus on these pseudo-neologisms, the proposed model and algorithm has been successful, identifying an emerging 'sense' in an unsupervised fashion. Moving on from this first test of the algorithm, the next section considers outcomes on authentic words.

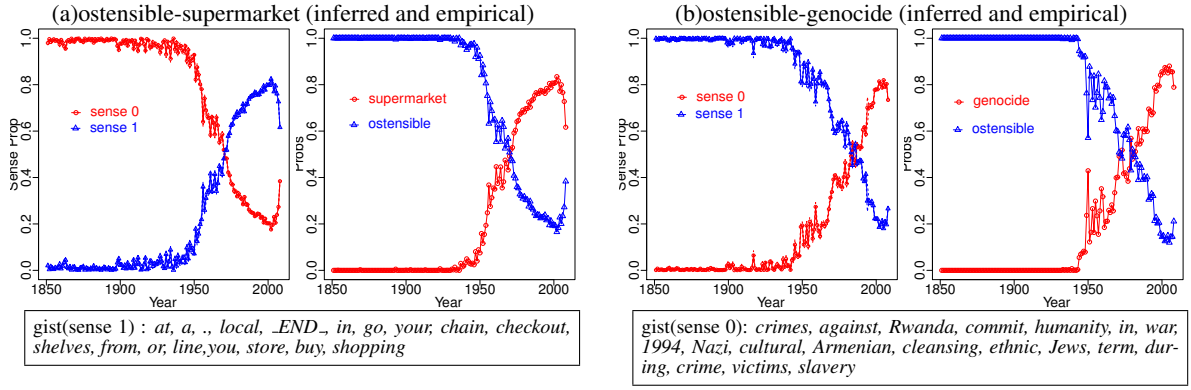



Figure 2: For (a-b), the left-hand plots show the inferred $\pi_t[k]$ sense parameters for a pseudo-neologism $\sigma_1$-$\sigma_2$, and right-hand plot shows the known $\sigma_1$ and $\sigma_2$ proportions. Below the plots are 'gist(S)' words associated to the apparent neologism sense – see text.

### 3.2 Experiments with genuine neologisms

Table 1 listed both targets expected to show sense emergence and targets expected to not show sense emergence. For several of the sense emergence targets, Figure 3(a-d) depicts various aspects of the outcomes. In each case the leftmost plot for a target $\sigma$ shows for each $k$ the succession of inferred $\pi_t[k]$ values – the sense-given-year values – plotted as a solid line[3]; the rightmost plot in each case is a 'tracks' plot (see Section 2.1), showing for some collection of words considered to be associated with the novel sense the succession of their probabilities of occuring in n-grams for the target $\sigma$, $P_t(w|\sigma)$. These are the basis for the 'tracks' column in Table 1.

**mouse** Figure 3(a): The algorithm was run looking for 3 sense variants on data between 1950 and 2008. The blue line for the $\pi_t[1]$ sequence in the left-hand plot shows a neologistic pattern, starting near 0 and

[3]also shown is the HPD interval around the mean as dotted lines

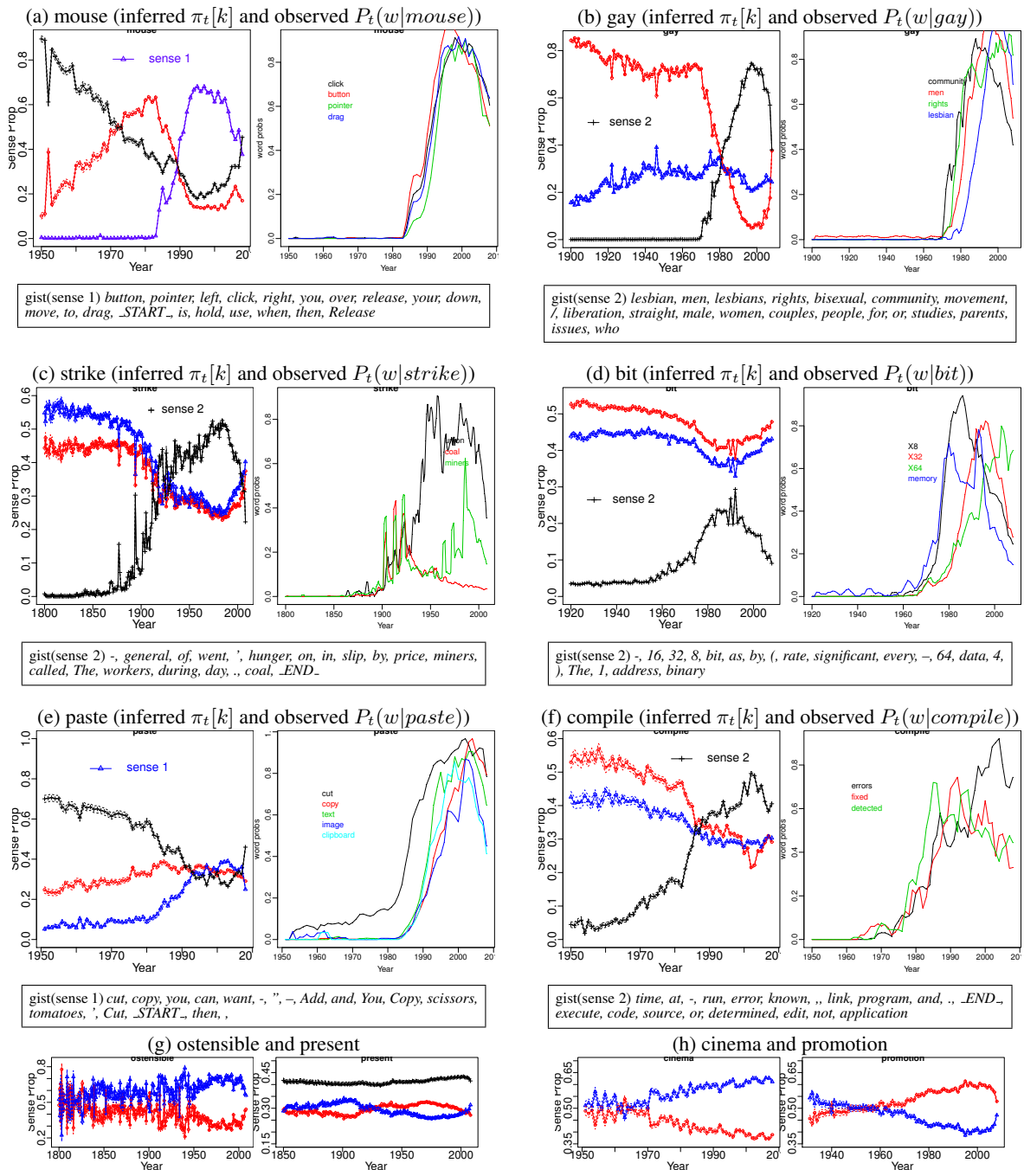



Figure 3: For (a-f), the left-hand plot shows the inferred $\pi_t[k]$ sense parameters, with the sense number $S$ of the potential neologism labeled; the right-hand plot show probability 'tracks' for some words intuitively associated with the neologism (see text for further details). The box below the plots has top 20 $gist(S)$ words for the neologism sense $S$. (g-h) show the inferred $\pi_t[k]$ sense parameters for negative targets

departing from 0 around 1983. The 'tracks' plot also shows that several words intuitively associated with the neologistic sense, also drastically increase their probability conditioned on *mouse* around the same time. The 'GS-date' column of Table 1 gives the time $t$, if any, in a $\pi_t[k]$ sequence where it appears to depart from, and continues to climb from, zero. The '$< 10\%$' column records whether this agrees with the tracks-based date to within $10\%$ of the time-span considered – which it does in this case. Notably in this case, the GS-based emergence date, though close to the tracks-based date, is more than 20 years *later* than the OED first citation date. The OED first citation comes from a research paper in 1965, but the mouse computer peripheral only became popular considerably later and it is not unexpected that the

date at which this use of the term *mouse* departed and continued to climb from zero in the n-grams books based data is substantially later. We take this to illustrate why simply taking the OED first citation date, $D_0^c$, as a gold standard for the true corpus emergence date, $C_0$, would be a mistake[4]. The box below the plots in Figure 3(a) tries to give some insight into the estimated $\boldsymbol{\theta}_1$ parameter concerning word-given-sense probabilities by showing the words belonging to $gist(1)$ (see definition in section 3.1). They seem mostly consistent with the 'pointing device' sense.

**gay** Figure 3(b): In this case the procedure was run on data from 1900 to 2008, for 3 senses. In the left-hand plot the black line, for the $\pi_t[2]$ sequence, shows sense emergence, appearing to depart from near zero first around 1969. The 'tracks' plots to the right seem to increase around around 1966. The OED first citation date of 1922 predates both considerably. The 'gist' words for $S = 2$ also seem mostly consistent the 'homosexual' sense.

Similar to *mouse* and *gay*, the detailed outcomes for *strike, bit, paste, compile* are shown in figures (c - f) with the procedures run on data for 3 senses. For space reasons, these details are not shown for *boot, surf, strike, rock* and *stoned* but Table 1 summarizes all outcomes: in each case the inferred date was later than the OED first citation date, and in all cases close to the tracks-based date, just missing the $10\%$ margin in two cases.

Turning to the words which were *not* expected to exhibit an emergent sense, Figure 3(g-h) shows the plots of the inferred $\pi_t[k]$ sequences for the targets *ostensible, present, cinema* and *promotion*. None show a clear neologistic pattern, in line with expectations. Though the details are not shown in Figure 3 the same kind of outcome was found for the other negative targets listed in Table 1.

The value of $K$ varied somewhat between the experiments. That the number of senses possessed by the different targets varies is somewhat to be expected and in some cases where a neologistic trend was less clear with $n$ senses, it became clearer with $n + 1$. The automatic setting of this parameter remains an area for future work.

## 4 Comparisons to related work and conclusions

We have looked at the detection that a word has acquired the possibility to express a meaning which it could not hitherto (eg. *mouse* as pointing device). One can also look at senses themselves as changing over time, perhaps widening or narrowing, and there has been prior work addressing this issue (Sagi et al., 2009). We would like to treat this as a separate issue, though drawing a conclusive line between the two is tricky.

Concerning sense emergence specifically, it has to be stressed there is no strict quantitative state-of-the-art, because it is not the case that prior works share the same targets, use the same data, or address in the same way the tricky ground-truth issue (see Section 2.1). Bearing this in mind we have tried to organise the discussion below around major design options and papers that exemplify these.

There have been some proposals concerning sense emergence detection *without* modelling senses at all (Cook and Hirst, 2011; Kim et al., 2014). Though able to detect a difference between corpora from different eras, these systems tend to lack a capacity to pick out instances exemplifying a putative novel sense, which is arguably a desirable feature.

Moving on to systems which do involve some kind of modelling of senses, a noteworthy characteristic of many is that they often apply a WSI algorithm which is time-*unaware*. One design option is to *pool* all training data for the WSI phase, then assign likeliest senses to examples, and then to finally check for a correlation with time, such as a sense only being assigned after a particular time. Another design option is to *separate* the data into eras, perform independent WSI on each subset and then seek to consider how the sense representations from each era may (or may not) be *linked* to each other.

The *pooling* design option is exemplified in (Lau et al., 2012; Cook et al., 2013; Cook et al., 2014). Their time-unaware WSI system is based on LDA (Blei et al., 2003), and treats the $I$ words of a context as generated from $I$ topics, and then identifies a target's *sense* with the most frequent amongst the $I$ topics of the context words. It is furthermore an HDP variant of LDA (Teh et al., 2004) in which the

[4] other than as a lower bound

number of topics is self-determined by the training process. The equating of senses with topics could be questioned (Wang et al., 2015) and also the self-determined sense number in their illustrative examples seems strikingly high (10), with many unintuitive components included. Rather than a year-by-year time-line, their data is time-stamped to just two time *eras*, $\mathcal{E}_1$ and $\mathcal{E}_2$ (eg. in one of their papers $\mathcal{E}_1$ is the late $20^{th}$ century (BNC) and $\mathcal{E}_2$ is 2007 (UkWaC)), and so they attempt a much lower resolution of emergence dating than we do. Their approach to ground-truth on sense emergence was different also, being that using $D_0^i$ (dictionary *inclusion date*, mentioned in section 2.1), and so has some of the drawbacks that were noted there. 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 a 'novelty' score: the max over $k$ of the ratio of $\mathcal{E}_2$ to $\mathcal{E}_1$ frequency of assigned sense $k$. 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). As a possible generalisation of this score to a time-line, consider a 'novelty' score computed in the following way: from the sequence of $\pi_t[k]$ values, find 'min' and 'max' values and divide the temporally later value by the temporally earlier one, letting the novelty score be the max over $k$ of this ratio. On our targets this gives a ranking: { **stoned**($10^5$), **strike**(5442.7), **gay**(2791.1), **mouse**(1485.9), **surf**(156.7), **compile**(26.6), **bit**(10), **rock**(7.4), **boot**(7), *ostensible*(3.5), *plant*(1.89), *play*(1.8), *promotion*(1.4), *cinema*(1.3), *theatre*(1.1), *spirit*(1) }, separating the positive from the negative targets. Due to the data and target differences it would not make sense to compare these rankings. Earlier work by Rohrdantz et al. (2011) also instantiates the *pooling* option to exploit a time-unaware system. Their system was again LDA-based, their ground-truth approach was also $D_0^i$-based and their data was news articles between 1987 and 2007.

The *separate-then-link* strategy for deploying time-unaware WSI to nonetheless attempt to detect sense dynamics is exemplified in (Mitra et al., 2014; Mitra et al., 2015). The time-unaware WSI system in this case is a so-called 'Distributional Thesaurus' clustering approach (Rychlý and Kilgarriff, 2007; Biemann and Riedl, 2013), which starting from a word(type) co-occurrence graph where edges reflect co-occurrence, induces sets of words to represent a sense. Their data set consists of 'syntactic dependency n-grams' as produced by Goldberg and Orwant (2013) from the same digitised books as those from which the Google n-gram data is derived. They divide the entire time-line into eras $\mathcal{E}_1 \ldots \mathcal{E}_8$ of ever shortening duration but containing equal amounts of data (eg. $\mathcal{E}_2$ = 1909–53, $\mathcal{E}_7$ = 2002-05). For a given target, for each era they run their clustering to induce sense-representing word sets, and then they propose ways to link the clusters for $\mathcal{E}_i$, $\{s_1^i, \ldots s_m^i\}$ to the clusters of later era $\mathcal{E}_j$, $\{s_1^j, \ldots s_n^j\}$. Roughly speaking a cluster in $\mathcal{E}_j$ is judged a 'birth' (ie. sense emergence) if sufficiently few of its member words belong to the any of the clusters for the earlier era $\mathcal{E}_i$. In the paper they discuss outcomes concerning apparent 'births' when comparing the 1909-1953 and 2002-2005 eras. They do not test with respect to known positive and negative examples. Instead they apply the procedure to *all* words, obtain a very large set of candidate 'births', apply a relatively complex multi-stage filtering process to this and then on a randomly selected 48 cases from the filtered 'births' they find 60% are correct. Their approach to ground-truth concerning sense emergence (cf. Section 2.1) is somewhat varied but essentially was author intuition in (Mitra et al., 2014) and dictionary first citation $D_0^c$ in (Mitra et al., 2015), though as we have noted this should only serve as lower bound[5].

Unlike these proposals, the experiments in this paper concern a model which is *not* time-unaware: the model has variables and parameters referring to time. Earlier versions of this idea were discussed in (Emms, 2013; Emms and Jayapal, 2014; Emms and Jayapal, 2015) though differing from the work presented here in number of respects (such as the estimation approach (EM), data used (text snippets via Google custom date search) and the targets considered (multiword expressions)). This aspect of including time explicitly in a probabilistic model seems to have been considered much less often than the above-mentioned essentially time-unaware approaches. The work of Wijaya and Yeniterzi (2011) is one example. They do not propose a sense-emergence detection algorithm per-se but do make some

[5]Without getting too lost in case-by-case details, it is worth noting that some seem incorrectly judged true 'births' relative to the eras considered, such as an assailant sense of *thug*, a calculus sense of *derivative*

analyses on the Google n-gram data to seek indicators of sense change. They sought to apply the *Topics Over Time* variant of *LDA* (Wang and McCallum, 2006), to do which they somewhat curiously collapse a year's worth of n-grams for a target into a *single* 'document' for that year. They found for example that for *gay*, training for 2 topics, there is a switch from a strong preference for one topic to preference for the other around 1970.

The recent work of Frermann and Lapata (2016) is a further example of a time-aware probabilistic model, in fact one having much in common with the model we have been discussing. They, as we have done, consider a generative model in which for a given time $t$ a sense $k$ is chosen, according to some discrete distribution[6] $\boldsymbol{\pi}_t$, and then, again as we have assumed, the context words in $\boldsymbol{w}$ are generated independently of each other. Whereas we have assumed that word choices are conditionally independent of the time $t$ given the sense $k$, and so have for each sense $k$ a parameter $\boldsymbol{\theta}_k$ of word probabilities, they do *not* assume this independence, and so for each *time* $t$ and sense $k$ have a parameter $\boldsymbol{\theta}_{tk}$ of word probabilities. The key further feature of their model is their use of *intrinsic Gaussian Markov Random Fields* (iGRMFs) to have priors which control how the distributions $\boldsymbol{\pi}_t$ and $\boldsymbol{\theta}_{tk}$ change over time: basically there is a precision hyper-parameter $\kappa$ such that a *high* $\kappa$ favours *small* changes in successive values. For the succession of $\boldsymbol{\theta}_{tk}$ values, they set $\kappa$ to a high value, so that although $\boldsymbol{\theta}_{tk}$ does not have to be constant over time, only small variation is anticipated by the prior. The succession of $\pi_t$ values is allowed greater variation. This use of iGMRF-based priors requires in its turn a more sophisticated Gibbs sampling approach to parameter estimation than that which we have used – which they achieve following ideas of Mimno et al. (2008). From the perspective of their model, our model is more or less what would be arrived at by (i) letting $\kappa$ for $\boldsymbol{\theta}_{tk}$ tend to $\infty$, preventing any change of word-given-sense probabilites in successive times and (ii) letting $\kappa$ for $\boldsymbol{\pi}_t$ tend to 0, allowing arbitrary change of sense-given-time probabilities. They evaluated their model in a variety of ways, the most comparable of which was to consider particular target words in the *Corpus of Historical American English* (Davies, 2010) with number of senses set to 10 and a time-resolution of 10-year time spans. As with the other papers already discussed, their use of different targets and a different data-set means again there is not the possibility at the moment of a quantitative comparison. In our work whilst we do not have a prior 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, suggesting that for the n-gram data at least, the more complex system of Frermann and Lapata (2016) is not required. It may be of interest in future work to investigate to what extent this is dependent on the data-set used: the data-set they used contains ~100 times fewer occurrences for a given target per time-period compared to the n-gram data-set we have used and it could be that with less data the priors they propose become more necessary.

In conclusion we have proposed a simple generative model, with a $P(S|Y)$ term for time-dependent sense likelihood, and a $p(\boldsymbol{W}|S)$ term expressing that the context words are independent of time given the target's sense. The fact that intuitive outcomes were obtained on our pseudo-neologisms, and on some authentic cases of sense emergence and non-emergence is indicative at least that the model's assumptions are tolerable. It remains for future work involving further targets to test the limits of these assumptions. Amongst several possibiities for further investigation it would be of interest to reformulate the model to refer not just to plain words but rather to syntactic annotations, as well as to consider data sources representing other and more recent text types, such as social media posts.

## Appendix A. Derivation of sampling formula for $\boldsymbol{\pi}_t$

For the sampling formula for $\boldsymbol{\pi}_t$ we need the conditional probability $P(\boldsymbol{\pi}_t|\boldsymbol{\pi}_{-(t)}, \boldsymbol{s}^{1:D}, \boldsymbol{w}^{1:D}, \boldsymbol{t}^{1:D}, \boldsymbol{\tau}_{1:N}, \boldsymbol{\theta}_{1:K}; \boldsymbol{\gamma}_\pi, \boldsymbol{\gamma}_\theta)$, where indexing by $-(t)$ is meant to indicate consideration of all indices *except* $t$. This conditional probability is given by

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

Recalling the model formula (3) given in Figure 1 the numerator in this fraction is

[6] Adapting their notation to make things as comparable as possible: they have $\Phi^t$ rather than $\boldsymbol{\pi}_t$ and $\Psi^{t,k}$ rather than $\boldsymbol{\theta}_{tk}$.

$$\prod_d \left[ \tau_{t^d} \times \pi_{t^d,s^d} \times \prod_{i=1}^{|\boldsymbol{w}^d|} \theta_{s^d,w_i^d} \right] \times \text{Dir}(\boldsymbol{\pi}_t;\boldsymbol{\gamma}_\pi) \times \prod_{-(t)} \text{Dir}(\boldsymbol{\pi}_{-(t)};\boldsymbol{\gamma}_\pi) \times \prod_{1:K} \text{Dir}(\boldsymbol{\theta}_k;\boldsymbol{\gamma}_\theta)$$

and the denominator differs only by the the integral over $\boldsymbol{\pi}_t$. $\boldsymbol{\pi}_t$ is involved in the $\text{Dir}(\boldsymbol{\pi}_t;\boldsymbol{\gamma}_\pi)$ term and in those parts of the product for $d$ where you have $t_d = t$. Because of this most terms in the denominator can be taken outside the scope of the integral and then cancel with corresponding terms in the numerator. Because of this, the fraction can be written

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

In the data items $\{d : t_d = t\}$ a variety of sense values have been sampled and the numerator can instead be expressed using $\mathbb{S}_{t,k}$, which counts the sampled sense values (see Section 2). Re-expressing the numerator in this way and using the definition of the Dirichlet (Heinrich, 2005), 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

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

where the last step uses the fact that in any Dirichlet $\text{Dir}(\boldsymbol{x}_{1:K};\boldsymbol{\alpha}_{1:K}) = \frac{1}{\beta(\boldsymbol{\alpha}_{1:K})} \prod_{k=1}^K x_k^{\alpha_k-1}$, the 'normalizing' constant $\beta(\boldsymbol{\alpha}_{1:K})$ is the integral of the main product term. Hence we finally obtain

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

which is the sampling formula given earlier as (6). The derivation of the sampling formula for $\boldsymbol{\theta}_k$ is similar, and that for the discrete $s^d$ is straightforward.

## Acknowledgments

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.

## References

David Barnhart. 2007. A calculus for new words. *Dictionary*, 28:132–138.

Chris Biemann and Martin Riedl. 2013. Text: Now in 2d! a framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, 1(1):55–95, Apr.

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

Paul Cook and Graeme Hirst. 2011. Automatic identification of words with novel but infrequent senses. In *Proceedings of the 25th Pacific Asia Conference on Language Information and Computation (PACLIC 25)*, pages 265–274, Singapore, December.

Paul Cook, Jey Han Lau, Michael Rundell, Diana McCarthy, and Timothy Baldwin. 2013. A lexicographic appraisal of an automatic approach for detecting new word senses. In *Proceedings of eLex 2013*.

Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel word-sense identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, page 1624–1635. ACL, August.

Mark Davies. 2010. The corpus of historical american english: 400 million words, 1810-2009. available online at http://corpus.byu.edu/coha.

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

Martin Emms and Arun Jayapal. 2015. 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.

Martin Emms. 2013. Dynamic EM in neologism evolution. In Hujun Yin, Ke Tang, Yang Gao, Frank Klawonn, Minho Lee, Thomas Weise, Bin Li, and Xin Yao, editors, *Proceedings of IDEAL 2013*, volume 8206 of *Lecture Notes in Computer Science*, pages 286–293. Springer.

Lea Frermann and Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Alan E. Gelfand and Adrian F. M. Smith. 1990. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, jun.

Yoav Goldberg and Jon Orwant. 2013. 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. Association for Computational Linguistics.

Gregor Heinrich. 2005. Parameter estimation for text analysis. Technical report, Fraunhofer Computer Graphics Institute.

Yoon Kim, Yi-I. Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. *CoRR*, abs/1405.3515.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. 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.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

David Mimno, Hanna Wallach, and Andrew McCallum. 2008. Gibbs sampling for logistic normal topic models with graph-based priors. In *NIPS Workshop on Analyzing Graphs*.

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. 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.

Sunny Mitra, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, 21(5):773–798.

Christian Rohrdantz, Annette Hautli, Thomas Mayer, Miriam Butt, Daniel A. Keim, and Frans Plank. 2011. Towards tracking semantic change by visual analytics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 305–310. ACL, June.

Pavel Rychlý and Adam Kilgarriff. 2007. 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. Association for Computational Linguistics.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the EACL 2009 Workshop on GEMS: GEometical Models of Natural Language Semantics*, page 104–111. Association for Computational Linguistics, March.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Jesse T. Sheidlower. 1995. Principles for the inclusion of new words in college dictionaries. *Dictionaries*, 16:32–43.

John Simpson. 2000. Preface to the third edition of the oed. `public.oed.com/the-oed-today/preface-to-the-third-edition-of-the-oed`.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2004. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101.

Jean Tournier. 1985. *Introduction descriptive à la lexicogénétique de l'anglais contemporain*. Champion-Slatkine.

Xuerui Wang and Andrew McCallum. 2006. 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. ACM.

Jing Wang, Mohit Bansal, Kevin Gimpel, Brian D. Ziebart, and Clement T. Yu. 2015. A sense-topic model for word sense induction with unsupervised data enrichment. In Hwee Tou Ng, editor, *Transactions of the Association for Computational Linguistics*, page 59–71. Association for Computational Linguistics, January.

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. 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. ACM.