Exploring the limits of EM algorithm for tracking sense evolution in the COHA corpus

Josefa Kubitová

Final Year Project, 2021

Supervisor: Dr. Martin Emms

Declaration

I hereby declare that this thesis is entirely my own work and that it has not been submitted as an exercise for a degree at any other university.

_____August 20, 2021

Permission to Lend

I agree that the Library and other agents of the College may lend or copy this thesis upon request.

_____ August 20, 2021

Acknowledgments

Thank you to my parents, my cats and professor Emms for being very patient and supportive during the writing of this thesis.

Josefa Kubitová

University of Dublin, Trinity College

August 2021

Contents

1) Introduction

In the question of human-machine communication, ascertaining meaning in words with multiple senses can be a tricky task. This work aims to examine how well the EM algorithm performs in fulfilling that task, with special attention paid to semantic neologisms, as their pattern of sense evolution is unique compared to other polysemic words. Mapping of emergent senses is done using data from the Corpus of Historical American, or COHA for short, with the aid of pseudo-neologisms, which serve as artificially constructed semantic neologisms.

Historical evolution of meaning is extremely difficult to track without text processing, as it would require extensive work simply in cataloguing and labelling contexts in which certain words appear. The possibility to track it using unsupervised means is therefore one that should not be overlooked.

2) Overview

In this paper, we will first go over the necessary theoretical background. We will describe COHA, the corpus we are working with, we will define neologisms and their two subcategories of 'formal' and 'semantic' neologisms, and we will define the concept of pseudowords and their variation with a temporal aspect, pseudo-neologisms.

We will then go over the methods we used in examining the evolution of sense throughout time, how COHA was processed, how the EM algorithm was used to extract sense probabilities over time, and how we got the base truth for our pseudo-neologisms to verify our predictions.

In the Experiments chapter, we will go through examples of how well or poorly the EM algorithm performed under different conditions. The conclusions reached from these experiments will be summarized in the Results chapter, and the whole project reflected upon in the conclusion.

3) Background

a) COHA

In this paper, we are working with COHA; the Corpus of Historical American. It is, as the name suggests, a corpus of American texts sorted into decades from the 1810s to the 2010s. Compared to other text corpora, COHA is very well balanced with regards to time as well as genre. While recent decades are slightly better represented, they do not overwhelm the others. As for genre, COHA sources texts from newspapers, magazines, and both fiction and non-fiction books. These categories are balanced throughout the years so that about half of the corpus' texts are sourced from fiction (fiction books), half from non-fiction (newspapers, magazines, and non-fiction books). In total, COHA provides us with 475 million words through almost two centuries to work with.

Decade	Fiction	Magazines	Newspaper	NF Books	Total	Percent fiction
1810s	641,164	88,316	0	451,542	1,181,022	0.54
1820s	3,751,204	1,714,789	0	1,461,012	6,927,005	0.54
1830s	7,590,350	3,145,575	0	3,038,062	13,773,987	0.55
1840s	8,850,886	3,554,534	0	3,641,434	16,046,854	0.55
1850s	9,094,346	4,220,558	0	3,178,922	16,493,826	0.55
1860s	9,450,562	4,437,941	262,198	2,974,401	17,125,102	0.55
1870s	10,291,968	4,452,192	1,030,560	2,835,440	18,610,160	0.55

1880s	11,215,065	4,481,568	1,355,456	3,820,766	20,872,855	0.54
1890s	11,212,219	4,679,486	1,383,948	3,907,730	21,183,383	0.53
1900s	12,029,439	5,062,650	1,433,576	4,015,567	22,541,232	0.53
1910s	11,935,701	5,694,710	1,489,942	3,534,899	22,655,252	0.53
1920s	12,539,681	5,841,678	3,552,699	3,698,353	25,632,411	0.49
1930s	11,876,996	5,910,095	3,545,527	3,080,629	24,413,247	0.49
1940s	11,946,743	5,644,216	3,497,509	3,056,010	24,144,478	0.49
1950s	11,986,437	5,796,823	3,522,545	3,092,375	24,398,180	0.49
1960s	11,578,880	5,803,276	3,404,244	3,141,582	23,927,982	0.48
1970s	11,626,911	5,755,537	3,383,924	3,002,933	23,769,305	0.49
1980s	12,152,603	5,804,320	4,113,254	3,108,775	25,178,952	0.48
1990s	13,272,162	7,440,305	4,060,570	3,104,303	27,877,340	0.48
2000s	14,590,078	7,678,830	4,088,704	3,121,839	29,479,451	0.49
Total	207,633,395	97,207,399	40,124,656	61,266,574	**406,232,024**	0.51

Table 3-1: COHA textual sources by decade

COHA has many other features, the most relevant to this paper being that it is fully lemmatised; this means that all variations of a word will be treated as occurrences of various forms of its base form, that is the lemma. For example, all the variations 'swim', 'swam', 'swims' and others would all be counted as occurrences of the verb swim, lemmatized as swim.

b) Neologisms (formal and semantic)

This work deals with the tracking of sense evolution in semantic neologisms. Neologisms overall are simply described as new words, but they can be divided further into two categories: formal neologisms and semantic neologisms, which are both described below.

Formal neologisms are the more straightforward category – they are new words, those that didn't exist before a certain point. Typical examples include the names for new technologies, such as 'radio' or 'computer' or other new ideas. Formal neologisms are reasonably easy to detect in a corpus tagged for time since a new word simply appears at a certain point where there was none before.

Semantic neologisms on the other hand are new *meanings* acquired by already existing words. The most stereotypical example of a semantic neologism is 'mouse'. Up to a certain point (the 1960s), 'mouse' referred to a small furry animal, but from the 60s onwards, it gained the additional meaning of a computer accessory. Semantic neologisms are harder to detect since the word merely gains a new context in which it is used. Furthermore, even if we attempt to track them, it is extremely difficult to verify that our predictions are accurate, unless we go through our corpus to manually tag each occurrence of our word as either "original meaning" or "neological meaning." This problem can luckily be solved with the use of "pseudo-neologisms"

c) Pseudowords and Pseudo-neologisms

As mentioned above, there are certain complications with confirming the accuracy of our predictions of sense evolution in semantic neologisms. Luckily, this is not at all a new problem in the field of textual analysis, and there is a simple method to deal with it. To avoid having to manually tag every occurrence of a polysemic word to specify whichever meaning is being referred to, Schutze (1998) suggests the use of pseudowords.

A pseudoword is a combination of two words that are then treated as one word by the algorithm analysing a given text. For example, with the pseudoword banana-dog, all instances of 'banana' and all instances of 'dog' will be treated as instances of banana-dog. Obviously, banana-dog will then seem to have two meanings – one being a yellow fruit, the other an animal.

A pseudo-neologism is a pseudoword with an added diachronic element. If one of the words chosen as part of a pseudoword is a formal neologism, we can call that pseudoword a pseudo-neologism. For example, banana-radio is a possible pseudo-neologism. Pseudo-neologisms model semantic neologisms nicely, with one meaning being present throughout the examined period and the other emerging during that time. We can therefore assume that any algorithm that accurately describes the evolution of senses in a pseudo-word will also be able to accurately describe the evolution of senses in a semantic neologism.

4) Methods:

a) Preparing COHA

Before anything else, we had to process COHA data into a format that would be easy to work with. For this purpose, two main files were created: firstly a list file 'list_1850_to_2000', which enumerates the files containing COHA data for the time period examined and is referred to often by the EM algorithm.

```
.
.
/shared/teaching/CSLL/4thYrProjects/ForJosefa/CorpData/wlp/1920s/news_1922_688216.txt
/shared/teaching/CSLL/4thYrProjects/ForJosefa/CorpData/wlp/1920s/news_1928_695015.txt
/shared/teaching/CSLL/4thYrProjects/ForJosefa/CorpData/wlp/1920s/mag_1926_486229.txt
/shared/teaching/CSLL/4thYrProjects/ForJosefa/CorpData/wlp/1920s/mag_1928_164026.txt
.
.
/shared/teaching/CSLL/4thYrProjects/ForJosefa/CorpData/wlp/1980s/mag_1989_329460.txt
/shared/teaching/CSLL/4thYrProjects/ForJosefa/CorpData/wlp/1980s/news_1982_668398.txt
/shared/teaching/CSLL/4thYrProjects/ForJosefa/CorpData/wlp/1980s/mag_1984_487769.txt
/shared/teaching/CSLL/4thYrProjects/ForJosefa/CorpData/wlp/1980s/news_1983_668464.txt
.
.
```

Figure 4-1: Excerpt from list file

Secondly, a call to make_coha_counts created a per_year_counts file which contains the occurrences per decade for every word with over 400 occurrences in total. This comes into play when calculating the ground truth for our pseudo-neologisms.

Figure 4-2: Excerpt from per_year_counts file

Formal neologisms were filtered out from this file as words with no representation for at least two decades in a row[1]. They ranged from the less common well-represented neologisms (the largest being 'soviet' with over 34 000 occurrences) to the cut-off point of 400 occurrences.

Figure 4-3: Excerpt from poss_formal_neols file

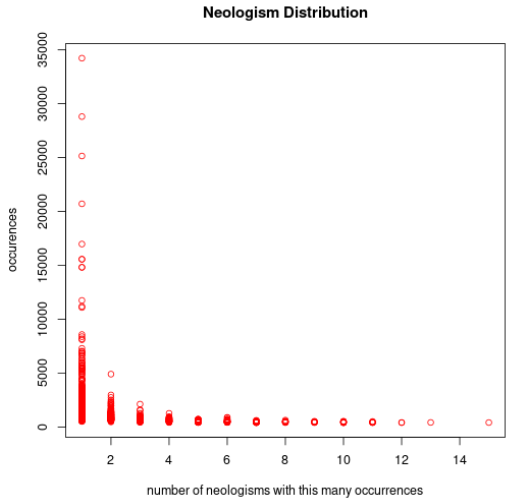


Figure 4-4: Distribution of neologisms from poss_formal_neols file

As we can see in the graph above, formal neologisms with more than 8000 occurrences throughout COHA are rather rare; work on these can be patchy at times as a result, simply because there are no intermediate neologisms between, for example, 22 000 and 25 000 occurrences. Another point to keep in mind is that some of these 'neologisms' are awkward to use as models for a second meaning coming into being, because they are names of important figures that tend to peak in the decade they were in power and only get a few mentions afterwards. Others can be names of countries (such as Vietnam with 8183 occurrences) or even just numbers (like 1980 with 3759 occurrences).

[1] grep '0 0 ' per_year_counts > poss_formal_neols

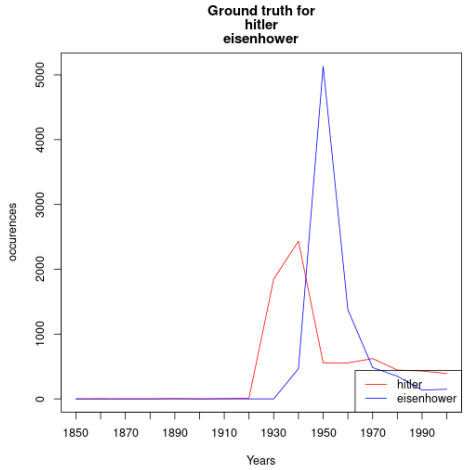


Figure 4-5: Occurrences in COHA of Hitler, Eisenhower

b) EM algorithm

This paper processes COHA using the EM algorithm, which calculates the relative frequencies for either of the pseudo-senses of our pseudo-neologism that make the data as likely as possible. Since EM is an unsupervised algorithm, we must note that the senses it produces are not labelled as to which sense they represent, and we are merely making an educated assumption when labelling them as one or the other.

When going through this process, we make use of the programs 'corpus_cereal_maker', 'dynamicEM', and 'inspector'. Firstly, corpus_cereal_maker reads in the corpus and creates and saves an 'archive' - a data structure containing all the pertinent information on whichever pseudo-neologism we are examining - which dynamicEM and inspector can make further use of. This step is not strictly necessary, as dynamicEM when run without an extant archive will create one on its own, but dynamic EM does not store this archive, which means that when re-running experiments after some time, one would be forced to go through the lengthy text-processing step again. An example corpus_cereal_maker call would look something like:

```
./corpus_cereal_maker \
-targ pray/movie \
-corpus_type 3 \
-files /users/ugrad/kubitovj/CohaExpt/list_1850_to_2000 \
-left 4 \
-right 4 \
-whether_pad no \
-whether_csv_suffix no \
-group_size 10 \
-wlp 1 \
-whether_archive yes \
-archive_name /users/ugrad/kubitovj/CohaExpt/pray_movie_1850_2000_archive
```

Figure 4-6: corpus_cereal_maker call

This call specifies that we want to create an archive for the pray-movie pseudo-neologism between the years 1850 and 2000 and store it at /users/ugrad/kubitovj/CohaExpt/pray_movie_1850_2000_archive. Note that in all this and all subsequent calls, we refer to the pseudo-neologisms in play as 'word1/word2', which symbolises a logical disjunction (OR). Simply said, this means that all occurrences of word1 and all occurrences of word2 will be considered targets.

Moving on, we would use the dynamicEM program to run all the EM calculations necessary and give us (among other documentation) a 'senseprobs_final' and a 'wordprobs_final' file, containing the final parameter estimation for our n senses and the words associated with them, respectively. In the example call below (one which does not make use of an archive), note the disjunctive notation pointed out above, the list file we are using to specify the time span we want to work with, and that we have to specify that we are looking for two senses for this pseudo-neologism.

```
/shared/teaching/CSLL/4thYrProjects/Software/Neologisms/DynamicEM/dynamicEM \
-targ automobile/cure \
-corpus_type 3 \
-expts /users/ugrad/kubitovj/CohaExpt/automobile \
-files /users/ugrad/kubitovj/CohaExpt/list_1850_to_2000 \
-left 4 \
-right 4 \
-whether_csv_suffix no \
-whether_pad no \
-group_size 10 \
-wlp 1 \
-num_senses 2 \
-senseprobs_string 0.45/0.55 \
-sense_rand no \
-word_rand yes \
-set_seed yes \
-rand_mix no \
-rand_word_mix yes \
-rand_word_mix_level 1e-05 \
-words_prior_type corpus \
-unsup yes \
-max_it 150 \
-time_stamp yes
```

Figure 4-7: dynamicEM call

The sense distribution file is also used by a plotting function, which features heavily when illustrating examples in the Experiments section. This plotting function does not name any of the senses outright, simply labelling them with numbers, as the EM algorithm only provides the number of senses it is told to look for, but cannot label any of them further. By convention, senses 0 through 4 will be represented by red, blue, black, green, and purple lines respectively. For example, in the automobile-cure relative sense distribution plot below, we know that sense 0 is represented by the red line (starting to grow only in 1880) and sense 1 by the blue.

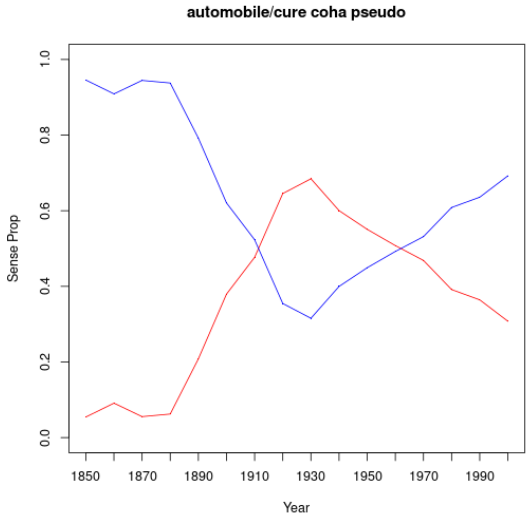


Figure 4-8: Automobile-cure EM prediction of relative sense distribution

Both 'senseprobs_final' and a 'wordprobs_final' can then be processed into a more accessible format by the inspector program, which can for example read out the most common associated n words for our senses, or even examine those words' most common contexts. Continuing on with the previous example of automobile-cure, looking at the most commonly associated words with each sense below, we can now associate sense 0 with the 'automobile' pseudo-sense and sense 1 with 'cure'.

- SENSE 0:
  worker,united,industry,drive,an,on,accident,association,in,into,production,manufacturer,company,american,street,dealer,new,club,from,steel,truck,park,two,speed,motor,tire,big,car,down,factory

- SENSE 1:
  you,i,can,disease,",n't,not,do,it,me,could,will,no,if,but,would,cancer,?,know,think,ill,my,only,him,what,cure,evil,she,effect,that

c) Ground Truth

As mentioned earlier, we decided to use pseudo-neologisms for testing the accuracy of the EM algorithm because of an absence of ground truth for semantic neologisms. With pseudo-neologisms, we can easily extract the factual number of occurrences of either pseudo-sense from our per_year_counts file. We can then either normalize the data to get results comparable to the EM output, or not if we simply want to take a look at the actual evolution of the number of occurrences for our pseudo-senses in any given decade. The plotting function for the ground truth, unlike the EM prediction plots, provides labels for all our pseudo-senses, making it very useful for double-checking our results. See appendix A for further detail on the plotting function.

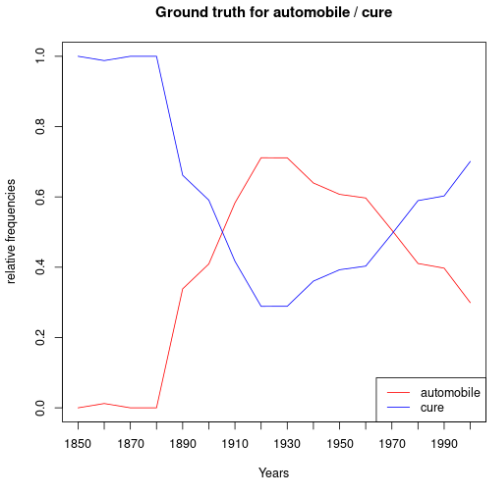


Figure 4-9: Actual relative representation of automobile-cure

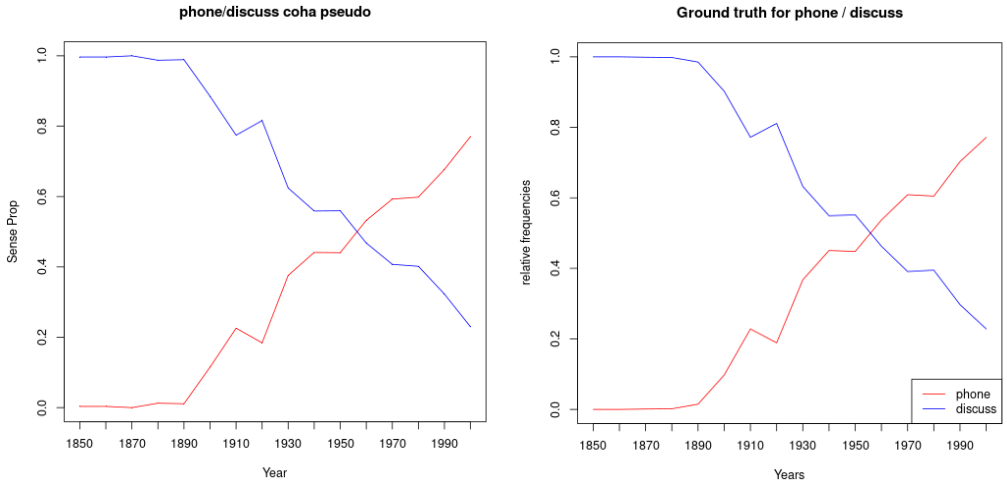
5) Experiments

When examining the evolution of meanings in neologisms, we had several basic questions that we hoped to answer. Firstly, we wanted to know how well-represented a meaning must be for the algorithm to catch it, secondly, we asked whether this is affected by the ratio of our two meanings' representation in the corpus, and finally, we experimented a little with the best choice of complimentary words to pair with a neologism to produce a pseudo-neologism that would be easy to work with.

a) Testing the algorithm

Before we moved on to more concrete tests, several pseudo-neologisms were selected to simply test how the algorithm worked. Large, balanced neologisms were used under the assumption that those would be easier for the algorithm to parse, as they provide plenty of data for both senses. Precise counts for the pseudo-neologisms as well as associated word senses are all provided in appendix B.

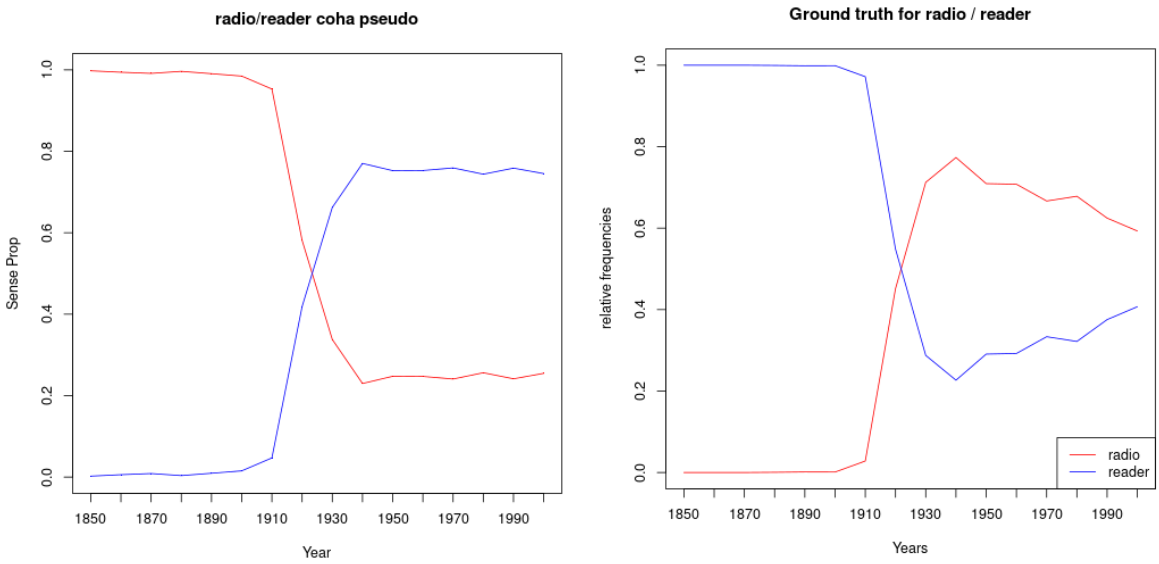
Starting from the better-represented end of things, we have phone-discuss; with phone occurring 28805 times throughout the text and discuss occurring 28800 times.



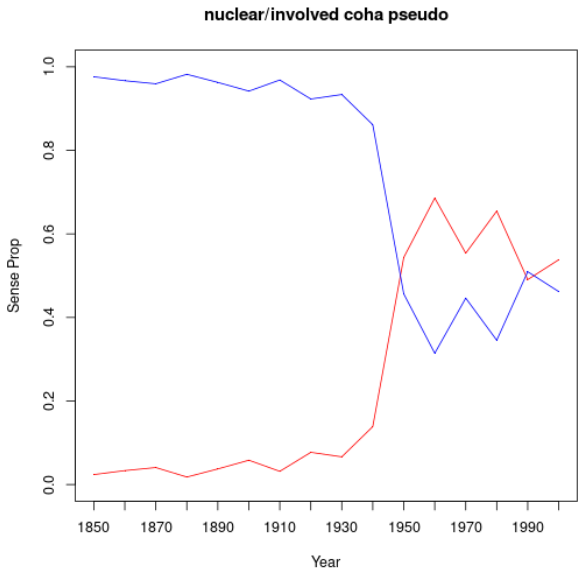
As we can see above, with a well-represented pseudo-neologism like phone-discuss, the EM algorithm is very reliable. It has inferred that there are two senses, one of which appears in the 1890s and grows steadily until the 2000s. Looking at the associated words below, we can assume that sense 0 (red) refers to the 'phone' sense of phone-discuss, while sense 2 (blue) refers to 'discuss'. This is further confirmed simply by comparing the predicted graph to the ground truth.

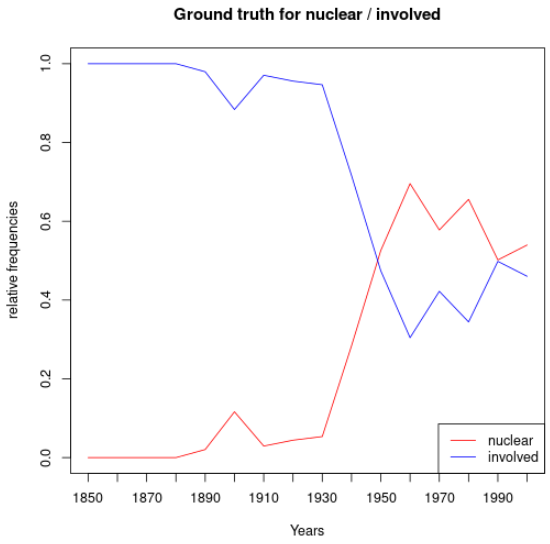
- SENSE 0 (red):
  ring,up,(,),into,pick,cell,call,on,get,she,answer,number,booth,hang,<p>,her,dial,me,down,-,off,hello,--,tell,back,hand,pay,go,?
- SENSE 1 (blue):
  matter,question,subject,not,of,which,problem,we,chapter,these,this,plan,they,situation,issue,in,detail,such,their,with,point,meet,far,be,much,refuse,topic,possibility,meeting,some

Moving on to radio-reader, both with around 25 000 occurrences, we see below that the predictions continue to be rather accurate, though there is some imprecision after 'radio' comes into use, possibly because of the sheer number of contexts that 'reader' can be used in.



Looking now at nuclear-involved, both with just over 11 000 occurrences, we see some mild imprecision around 1900 and 1950, mostly because of the algorithm overestimating the peaks that occur in those decades.

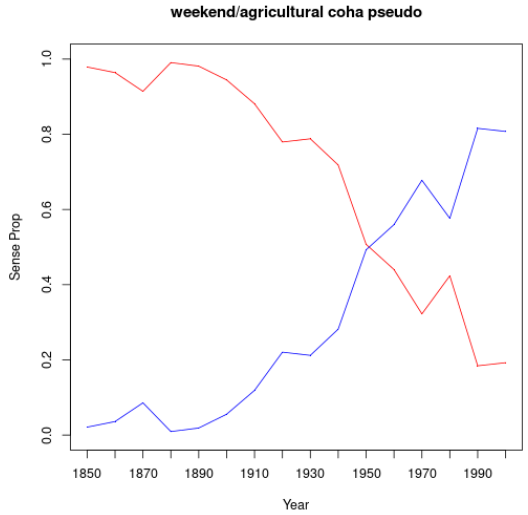


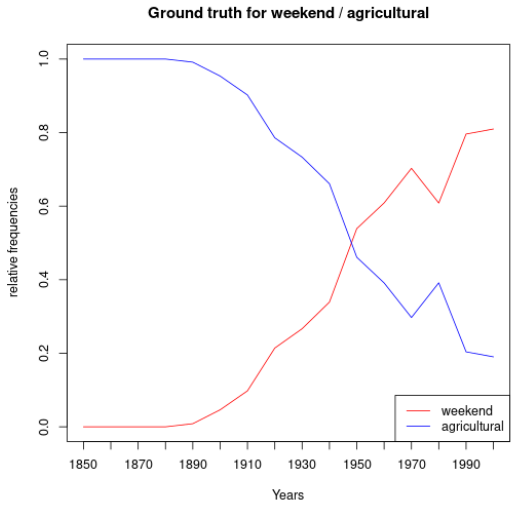


b) Meaning Representation

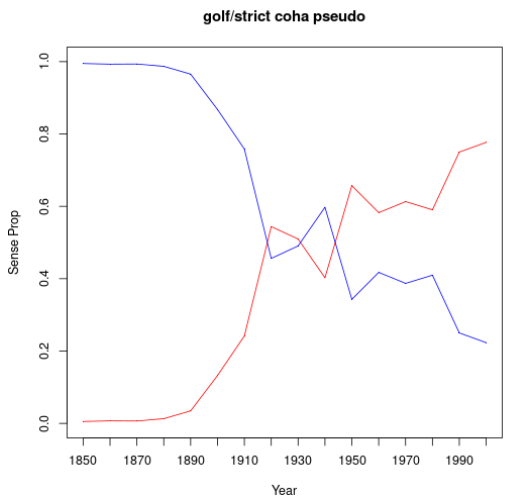
Every algorithm needs a certain amount of data to be able to accurately predict the desired outcome. Considering balanced pseudo-neologisms, several were selected that were less and less represented in the corpus in order to examine at which point the quality of predictions made by the EM algorithm would deteriorate. To do this we started with the best-represented neologisms and progressed downwards, checking the algorithms' accuracy along the way. A detailed distribution of the used pseudo-neologisms' occurrences can be found in appendix C.

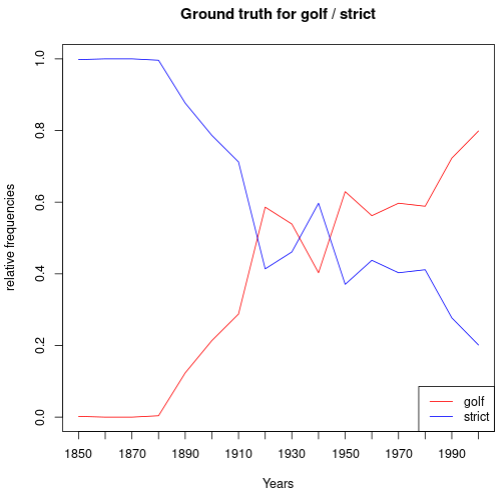
To start with, we have weekend-agricultural, at a little over 8500 occurrences. The predicted values are a little less smooth, especially because of an unexpected peak around 1870, but still very accurate overall.



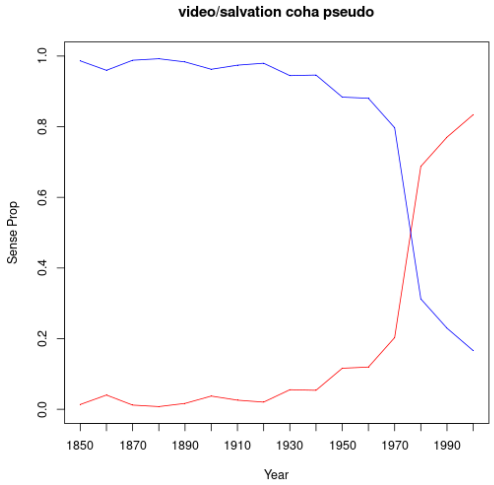


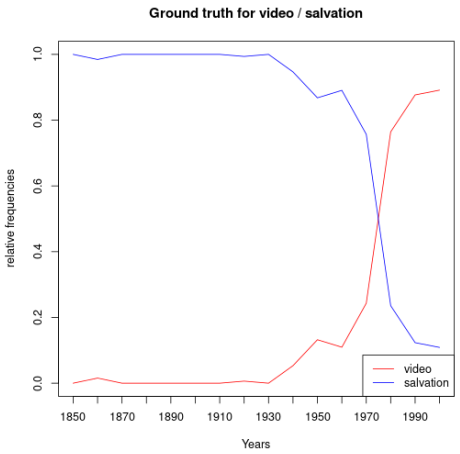
Next we have golf-strict at around 7000 occurrences, which turned out nearly perfect again, with only some slight miscalculation of the peaks throughout the first half of the 20th century.



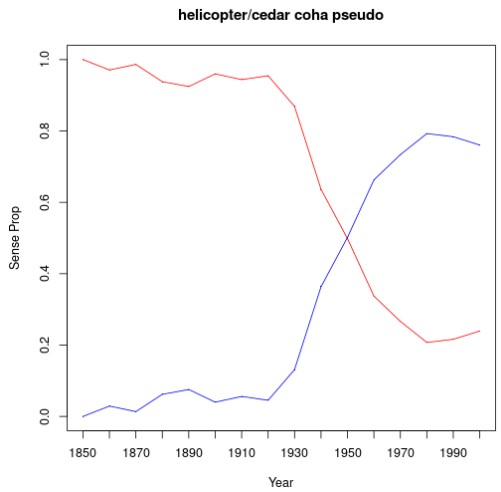


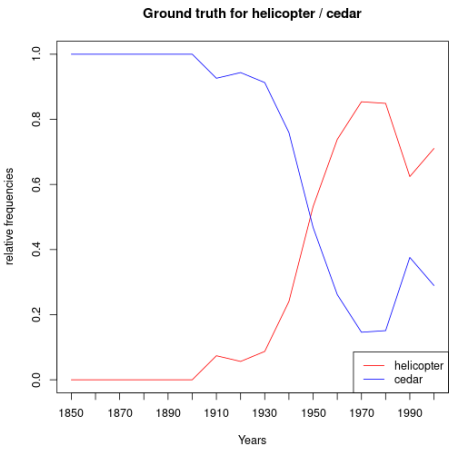
Moving on to video-salvation, which is interesting in that 'video' is a relatively new formal neologism, appearing in the 1940s and then taking off at breakneck speed to gain almost 5000 occurrences by the 2010s. The algorithm still performs well overall though, despite this unusual circumstance.



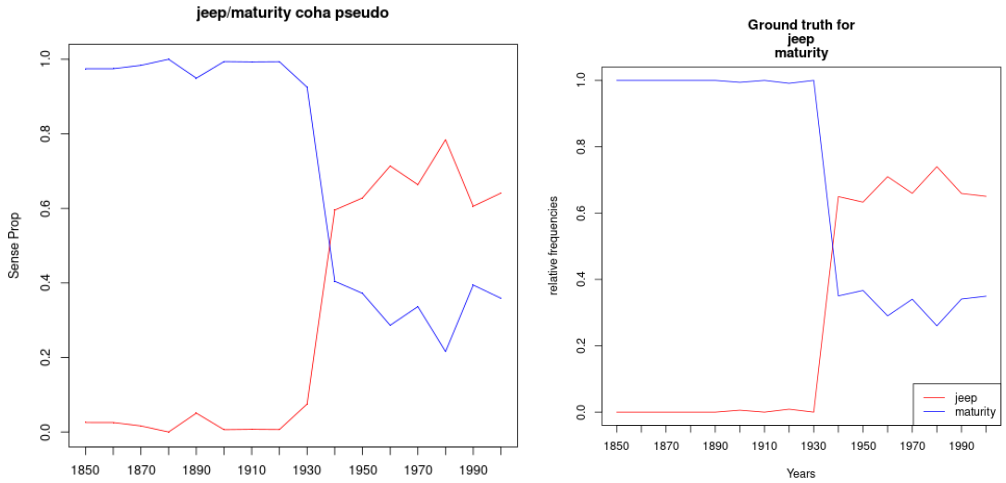


Next we have helicopter-cedar, at around 3500 occurrences. We can at this point see some imprecision in the pre-helicopter stage, and the 1990 dip is wholly ignored, but the overall tendency of the meanings' evolution is captured by our algorithm.

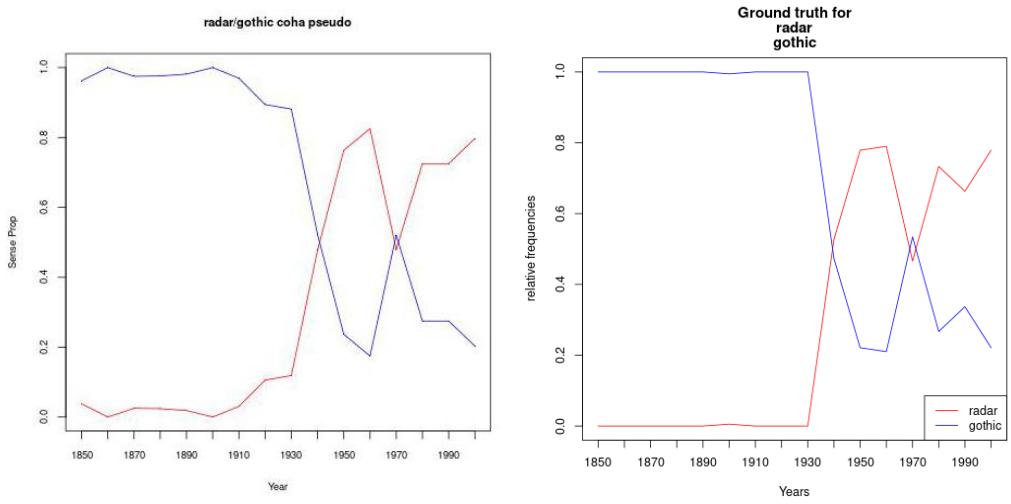




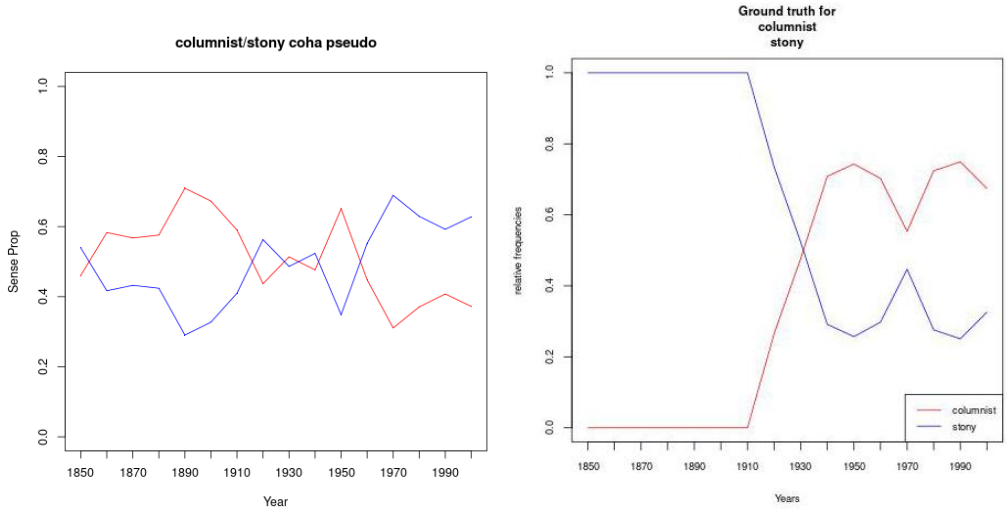
Moving on to jeep-maturity at just over 3000 occurrences in COHA, we seem to revert back to more precise predictions with just some exaggeration of the 1980 peak.



Next up is radar-gothic, with only just over 2400 occurrences, but while the imprecision prior to the formal neologisms' rise does return, the overall evolution of meanings remains sound throughout our predicted timeline.



Finally, we arrive at columnist-stony, with merely 2066 and 2056 occurrences for columnist and stony respectively. It is here that our predictions break down completely.



For a clearer look at how the predicted probabilities got so muddled, we can check the words associated with either detected sense:

- SENSE 0 (red):
  its,where,way,through,walter,path,on,**commentator**,bed,steep,**author,beach**,over,ground,draw,joseph,**radio**,also,**road**,up,john,pearson,**mountain**,week,winchell,field,along,against,**tv**,francisco

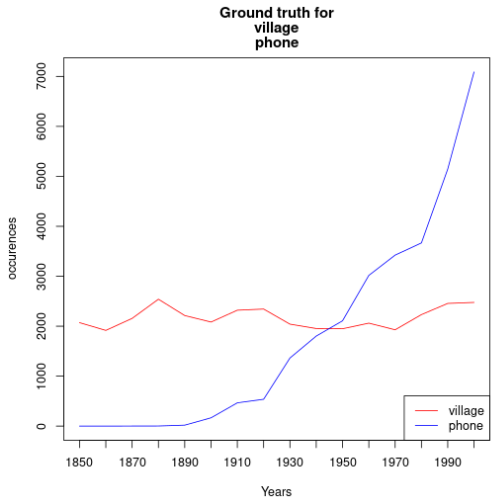
- SENSE 1 (blue):
  new,time,brook,york,silence,not,at,**island**,he,say,have,if,old,tell,you,as,when,avenue,do,heart,fifth,will,would,**conservative,sport**,**gossip**,those,she,william,smith

We can see that both senses contain words we would associate with both 'columnist' and 'stony'. Sense 0 has beach, road and mountain to suggest it might represent stony, but also radio, commentator or tv to suggest that its meant to represent the 'columnist' sense of columnist-stony. Sense 1 is no better as it mixes brook and island with conservative, sport and gossip, similarly making it unclear which of our two 'senses' it represents.

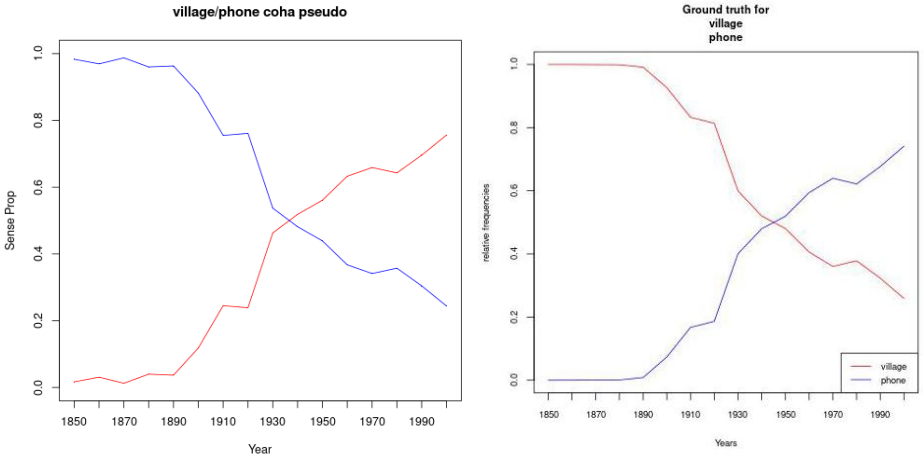
c) Unbalanced Pseudo-neologisms

Now that we've explored how many occurrences the EM algorithm needs to detect a sense in a balanced pseudo-neologism (that is, when both 'senses' are roughly equally represented), we can examine the limits we face with unbalanced pseudo-neologisms, as it is more likely than not that any naturally occurring semantic neologism will not be perfectly balanced in its representation. We have therefore chosen several neologisms in descending order of magnitude to be paired with 'village' (with nearly 35 000 occurrences) in order to examine what ratio our two meanings must have at the very least for the smaller meaning to be detectable. A breakdown of the exact number of occurrences of 'village' as well as all the chosen formal neologisms throughout COHA is available in appendix D.

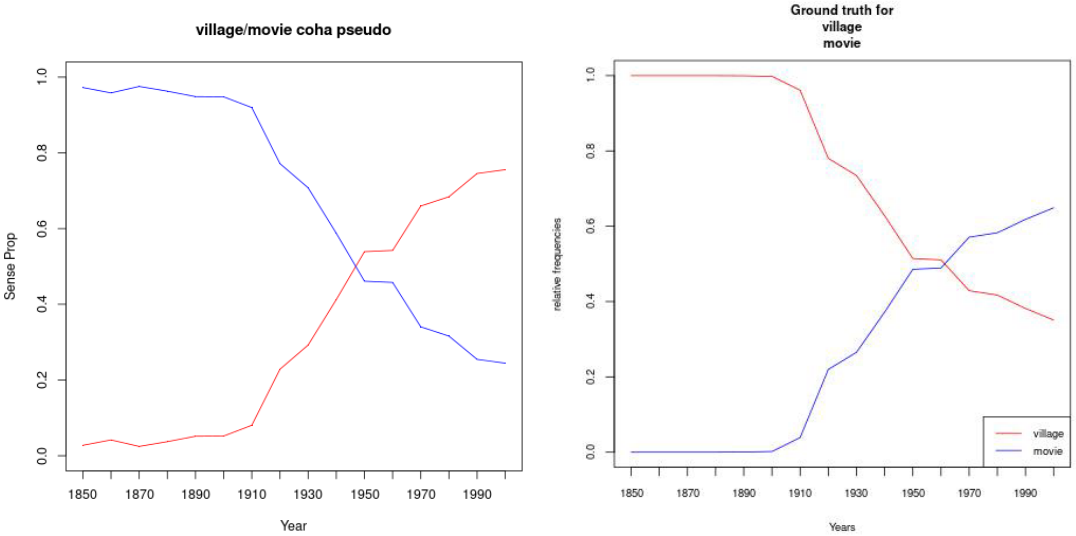
First up, we are looking at village-phone, where 'phone' has just over 28 800 occurrences, therefore being about 83% the size of 'village'. To be able to more clearly imagine what that means, we can look at the graph below, detailing the actual representation of both 'phone' and 'village' throughout the examined time period. We can see that use of 'village' is very stable. 'Phone', on the other hand, starts being used in the 1890s, overtakes 'village' around 1930 and is used about three times as much as 'village' by 2000.



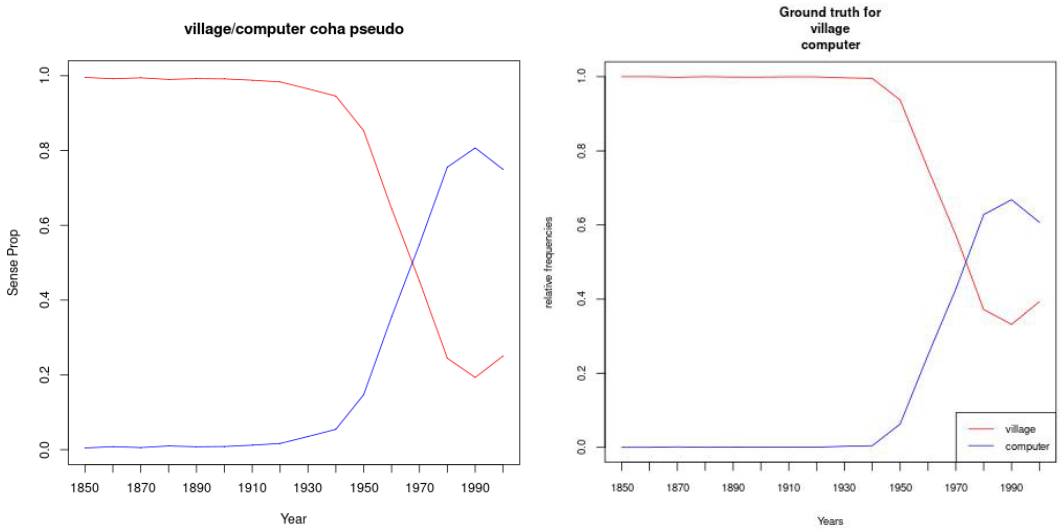
Looking now at the relative representation, we're doing well; the senses' evolution is modelled nicely, despite some shakiness before 1890.



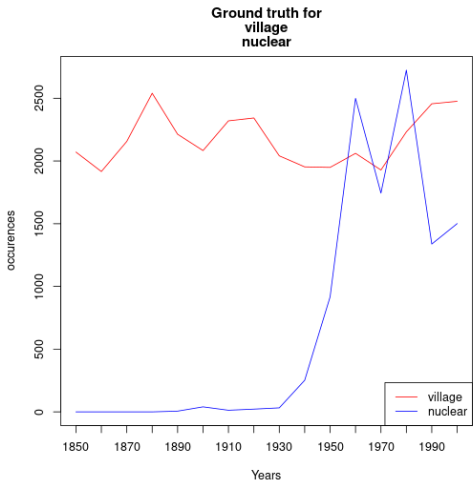
Moving on to village-movie, 'movie' occurring 20 711 times, therefore being 60% the size of 'village', we start seeing some slight inaccuracies, but overall we still have a very good approximation of the senses' evolution.



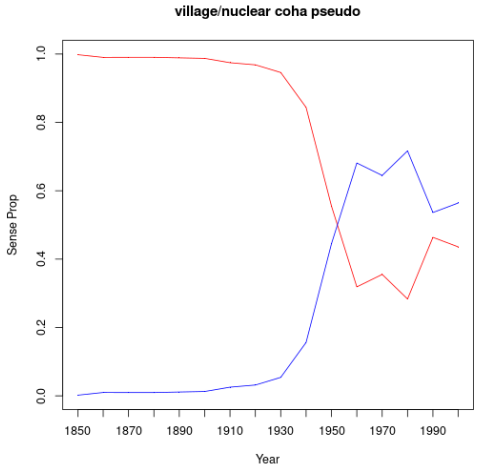
Next we have village-computer, with ‘computer’ at just over 14 800 occurrences being 43% the size of ‘village’. We can clearly see the exaggeration of the 1990 peak, but the shape direction of our senses’ shift is modelled well.

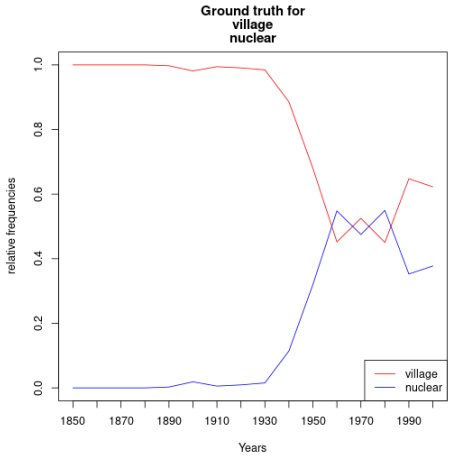


Moving on to village-nuclear, ‘nuclear’ having just over 11 000 occurrences has 32% of the representation of ‘village’. It is also worth noting that we are at a point where ‘nuclear’ just barely ever reaches the same level of use as ‘village’.

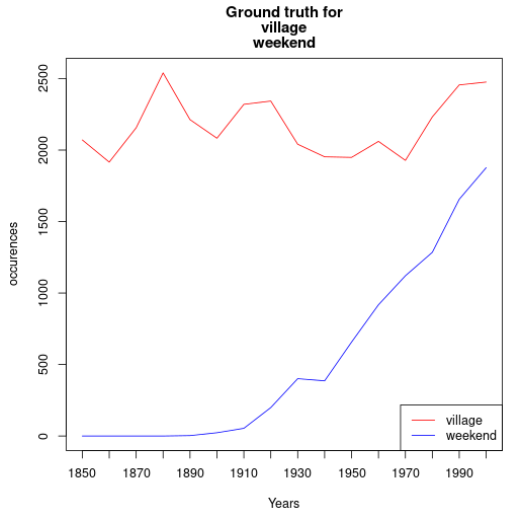


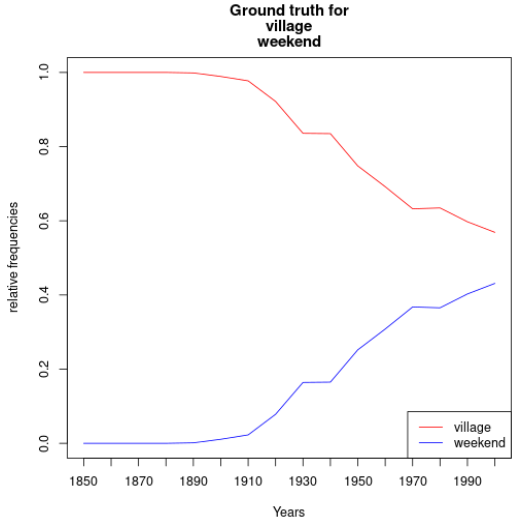
And indeed, we can see that the predictions, while preserving the overall characteristic shape, are becoming unreliable at this point.

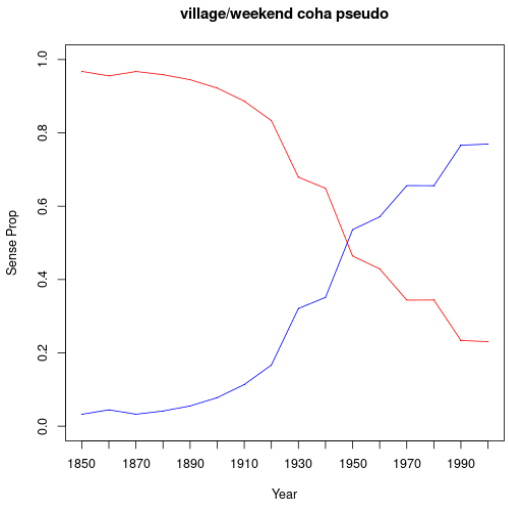




And finally, we have village-weekend, 'weekend' being 25% the size of of 'village' with 8580 occurrences. 'Weekend' never reaches the same level of representation as 'village' and the sense probability predictions break down completely at this point.







It is interesting to note though, that our associated words still seem to be strongly associated with one or the other expected sense.

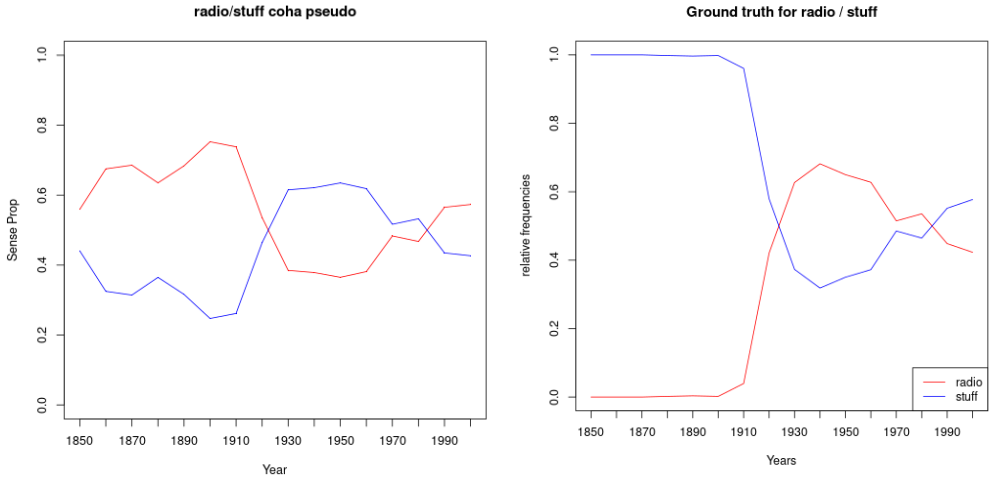
- SENSE 0:
  little,indian,which,town,upon,native,where,;,street,through,large,church,of,country,small,man,village,pass,far,mile,walk,reach,our,from,boy,way,young,very,beyond,find

- SENSE 1:
  <p>,spend,last,for,(,),",over,on,'s,:,day,during,long,holiday,n't,say,work,next,this,you,home,will,night,voice,plan,year,summer,york,july

We can easily associate sense 0 with 'village' and sense 1 with 'weekend', so perhaps even a very unbalanced pseudo-neologism like village-weekend, while not ideal, can still hold some value.

d) Word Choice

Finally, it is worth mentioning that in choosing a word to be paired with a neologism to form a pseudo-neologism, there are several factors to keep in mind. Firstly, the word must have a well defined meaning of its own; otherwise, it will be difficult to distinguish from our neologism. As an example of a badly defined word, we can look at 'stuff' and the pseudo-neologism radio-stuff that it can form.



As we can see from the graphs above, because 'stuff' is difficult to define and doesn't appear in an easily definable context, its meaning gets muddled with 'radio' and we can't get a clear reading of the meanings' evolution.

In addition to this, while it might seem obvious, we must keep in mind that even a word with a clear meaning of its own could get confused with the neologisms' meaning if the two are sufficiently similar. As an example, imagine the pseudoword mud-dirt; while there will be some difference between the two meanings, there are countless better words to pair with mud that will clarify its use better.

6) Results

The EM algorithm, as applied in this paper, can indeed effectively predict the evolution of various senses within a pseudo-neologism, and therefore within a semantic neologism as well. For best results, one should examine words with more than 2400 occurrences. This is an achievable goal, as while COHA does not provide us with many words with tens of thousands of occurrences, words with less than some 8000 occurrences are plentiful.

At the same time, the neologisms' less represented meaning should have at least 30% of the larger ones' representation and should be rather distinct from the first. These parameters are less definite and could perhaps use further investigation.

7) Conclusion

In this paper we successfully applied the EM algorithm to the problem of tracking the evolution of senses in semantic neologisms, using pseudo-neologisms as a model and COHA as our source of data. Similar tests could certainly be run using a different algorithm.

Being able to map changes in a words' meaning is interesting not only from a historical linguistic perspective, but also with regards to the possibility of being able to track new emergent meanings as they emerge, though that would require another work altogether.

8) Bibliography

Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora* , pp. 121-157.

Emms, M., & Jayapal, A. (2016). Dynamic generative model for diachronic sense emergence detection. *Proceedings of COLING 2016.* Osaka.

Jurafsky, D., & Martin, J. H. (2008). *Speech and Language Processing.* Prentice Hall.

Schutze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics* , 97–123.

10) Appendices

a) Appendix A: Ground truth

```
plot_actual_relative_representation <- function(counts_per_yr,words, normalize=TRUE){

#we may plot up to 5 words
count=length(words);
if(count>5){
print('only operating on first 5 words given');
count=5
}

#read in table of counts per year
Table <- read.table(counts_per_yr,sep=' ',header=T,check.names=F);

#prep yrs data
yrs=colnames(Table);
yrs= yrs[3:length(yrs)];
yrsnum=as.numeric(yrs);
yrs_cnt=length(yrs);

#prep some variables
colors=c('red','blue','green','purple','orange');
colors=colors[1:count]
title=c("Ground truth for", words);
data<-array(dim=c(count, yrs_cnt));

#select required data from table, put into an array
for(i in c(1:count)){a=subset(Table, Table$WORD==words[[i]]);
a=unlist(a[3:length(a)]);
data[i,]=a;
}

#normalize if requested
if(normalize){for(i in c(1:yrs_cnt)){data[,i]=data[,i]/sum(data[,i])}}

#create the plot
if(normalize){
plot(0,type="n",xlim=c(min(yrsnum),max(yrsnum)), col='blue', ylim=c(0,1), ylab=c('relative frequencies'),
xlab="Years", xaxt="n", main=title);
}
else{
plot(0,type="n",xlim=c(min(yrsnum),max(yrsnum)), col='blue', ylim=c(0,max(data)), ylab=c('occurences'),
xlab="Years", xaxt="n", main=title);
}
axis(1, at=seq(min(yrsnum),max(yrsnum),by=10), labels=TRUE);

#plot the data
for(d in c(1:count)){
```

```
Line=data[d,];
lines(yrsnum,Line, type='l', col=colors[d]);
}

legend(legend=words, x="bottomright", col=colors, lty = c(1, 1));
}
```

b) Appendix B: First few pseudo-neologisms

Word	Total	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
phone	28805	0	0	1	2	19	166	466	537	1363	1802	2108	3017	3423	3667	5145	7089
discuss	28800	703	847	900	1060	1267	1534	1577	2308	2345	2196	2597	2600	2198	2394	2176	2098
radio	25148	0	0	0	1	2	2	49	1228	2897	4159	3437	2850	2219	2654	2798	2852
reader	25239	2345	1909	1984	1771	1420	1650	1685	1492	1169	1219	1409	1177	1110	1259	1682	1958
nuclear	11088	0	0	0	0	6	40	13	22	32	252	917	2499	1744	2724	1338	1501
involved	11073	246	295	291	307	287	304	430	476	567	636	827	1094	1274	1431	1328	1280

Associated words	
Phone-discuss	
SENSE 0: presumed 'phone' ring,up,(,),into,pick,cell,call,on,get,she,answer,number,booth,hang,<p>,her,dial,me,down,-,off,hello,--,tell,back,hand,pay,go,?	SENSE 1: presumed 'discuss' matter,question,subject,not,of,which,problem,we,chapter,these,this,plan,they,situation,issue,in,detail,such,their,with,point,meet,far,be,much,refuse,topic,possibility,meeting,some
Radio-reader	
SENSE 0: presumed 'reader' will,may,our,not,my,know,must,who,many,interest,if,mind,most,attention,that,should,any,understand,refer,find,this,remember,your,general,some,feel,think,young,to,book	SENSE 1: presumed 'radio' station,television,on,broadcast,over,<p>,turn,tv,(,),car,program,off,music,listen,radio,voice,play,talk,wave,report,up,digest,set,city,press,telephone,show,hear,news
Nuclear-involved	
SENSE 0: presumed 'automobile' worker,united,industry,drive,an,on,accident,association,in,into,production,manufacturer,company,american,street,dealer,new,club,from,steel,truck,park,two,speed,motor,tire,big,car,down,factory	SENSE 1: presumed 'cure' you,i,can,disease,",n't,not,do,it,me,could,will,no,if,but,would,cancer,?,know,think,ill,my,only,him,what,cure,evil,she,effect,that

c) Appendix C: Decreasing pseudo-neologisms

Word	Total	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
weekend	8580	0	0	0	0	3	23	54	199	401	386	657	918	1121	1285	1656	1877
agricultural	8520	377	242	288	391	349	473	498	730	1102	752	562	590	474	827	424	441
golf	6996	1	0	0	2	61	119	158	524	526	299	720	587	606	648	1174	1571
strict	7036	459	453	497	513	433	438	391	370	450	443	424	457	409	453	450	396
video	4967	0	6	0	0	0	0	0	2	0	12	30	29	98	772	1913	2105
salvation	5006	371	380	260	411	324	275	710	316	244	213	197	236	305	238	269	257
helicopter	3511	0	0	0	0	0	0	13	19	34	64	208	556	573	686	661	697
cedar	3510	255	154	162	225	181	212	163	318	356	202	183	197	98	122	398	284
jeep	3058	0	0	0	0	0	1	0	2	0	393	459	521	337	549	433	363
maturity	3043	160	159	141	124	131	171	171	225	284	212	266	213	174	193	224	195
radar	2434	0	0	0	0	0	1	0	0	0	216	384	421	163	574	305	370
gothic	2421	144	105	141	129	164	184	212	144	127	194	109	112	187	209	155	105
columnist	2066	0	0	0	0	0	0	0	45	124	265	292	283	197	244	293	323
stony	2056	121	141	169	156	129	136	107	125	136	109	101	120	159	93	98	156

Associated words	
Weekend-agricultural	
SENSE 0: presumed 'agricultural' product,industrial,production,state,college,land,society,experiment,station,department,of,interest,region,development,which,commodity,implement,research,adjustment,population,price,an,industry,increase,economics,machinery,act,produce,district,export	SENSE 1: presumed 'weekend' spend,i,she,<p>,you,go,last,on,day,he,next,night,over,this,home,every,",come,long,during,holiday,after,n't,her,his,visit,my,say,do,when
Golf-strict	
SENSE 0: presumed 'golf' play,club,course,ball,<p>,tennis,tournament,at,game,round,championship,cart,like,'s,(,golf,n't,link,hole,),or,pro,get,win,swing,miniature,open,sport,around, day	SENSE 1: presumed 'strict' sense,law,under,order,discipline,rule,keep,control,regulation,accordance,neutrality,enforcement,not,attention,account,economy,adherence,observance,government,by,confidence,upon,very,justice,maintain,term,hold,regard,word,supervision
Video-salvation	
SENSE 0: presumed 'video' game,camera,screen,store,play,music,watch,on,show,monitor,image,tape,digital,clip,computer,audio,use,film,video,movie,tv,book,display,a,home,television,violent,recorder,shoot,mtv	SENSE 1: presumed 'salvation' soul,army,man,own,our,god,work,hope,only,way,his,my,no,mean,christ,eternal,!,for,depend,out,world,of,country,seek,upon,their,necessary,believe,not,faith

Helicopter-cedar	
SENSE 0: presumed ‘cedar’ pine,tree,lebanon,spruce,creek,swamp,bough,grow,stand,lodge,old,grove,under,wood,red,fir,tall,oak,branch,little,so,green,of,clump,dark,balsam,smell,behind,top,forest	SENSE 1: presumed ‘helicopter’ -,<p>,fly,overhead,bend,pilot,slow,waltz,),hover,police,land,int,attack,plane,use,(,crash,army,iowa,circle,jet,gunship,rapids,night,u.s.-,hear,wait,shoot,force
Jeep-maturity	
SENSE 0: presumed ‘jeep’ drive,back,get,out,stop,down,truck,go,road,park,pull,climb,start,",up,stand,front,i,over,car,ride,behind,wheel,head,side,away,walk,driver,around,along	SENSE 1: presumed ‘maturity’ year,reach,grow,age,youth,early,than,full,attain,bond,not,its,sexual,this,time,child,may,power,experience,before,which,period,or,;,date,of,only,more,stage,average
Radar-gothic	
SENSE 0: presumed ‘radar’ screen,radio,system,station,off,signal,pick,on,control,missile,equipment,aircraft,'s,up,<p>,detect,beam,use,installation,contact,their,operator,get,",search,room,plane,track,under,n't	SENSE 1: presumed ‘gothic’ architecture,cathedral,church,style,period,arch,art,window,romanesque,spire,tower,building,early,form,renaissance,french,old,of,gothic,roman,vault,chapel,between,great,greek,english,france,in,structure,stone
Columnist-stony	
SENSE 0: its,where,way,through,walter,path,on,commentator,bed,steep,author,beach,over,ground,draw,joseph,radio,alsop,road,up,john,pearson,mountain,week,winchell,field,along,against,tv,francisco	SENSE 1: new,time,brook,york,silence,not,at,island,he,say,have,if,old,tell,you,as,when,avenue,do,heart,fifth,will,would,conservative,sport,gossip,those,she,william,smith

d) Appendix D: Unbalanced Pseudo-neologisms

Word	Total	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
village	**34738**	**2071**	**1916**	**2156**	**2540**	**2213**	**2083**	**2320**	**2343**	**2041**	**1953**	**1949**	**2061**	**1928**	**2232**	**2456**	**2476**
phone (83%)	28805	0	0	1	2	19	166	466	537	1363	1802	2108	3017	3423	3667	5145	7089
movie (60%)	20711	0	0	0	0	1	4	94	660	737	1157	1840	1974	2567	3117	3979	4581
computer (43%)	14817	0	0	3	0	2	2	1	1	6	9	131	684	1438	3769	4946	3825
nuclear (32%)	11088	0	0	0	0	6	40	13	22	32	252	917	2499	1744	2724	1338	1501
weekend (25%)	8580	0	0	0	0	3	23	54	199	401	386	657	918	1121	1285	1656	1877

Associated words	
Presumed village	**Presumed neologism sense**
SENSE1:of,town,little,small,which,indian,in,where,near,mile,street,city,live,their,village,country,church,through,every,native,some,people,pass,life,many,large,whole, old,man,upon	SENSE 0 (phone): ring,),(,pick,cell,<p>,up,call,you,answer,get,i,n't,me,", number, she, say, ?, hang,her,on,dial,-,my,your,hello,voice,talk,again
SENSE 1: town,little,indian,through,small,near,street,city,mile,,where,;, country, native,pass,down,church,come,of,upon,their,reach,every,from, which,farm, village,large,along,road	SENSE 0 (movie): <p>,star,n't,",you,make,do,tv,like,watch,?,show,television,see, theater,say,i,film,book,'s,play,movie,hollywood,(,want,what,can,about,',:
SENSE 0: little,town,indian,street,where,mile,near,city,man,church,house, native, ;, come,pass,country,small,there,go,every,reach,walk,live, upon,through, in,road,whole,from,farm	SENSE 1 (computer): <p>,use,system,program,personal,screen,game, 's, can, software, company, computer,your,science,network,industry, technology,work, data,model,for,n't,will,),(,design,business,electronic, information,control
SENSE 0: little,town,where,street,indian,through,come,man,every,she,i,go,his,her, down,house,church,walk,he,native,people,see,back,live,whole,village,pass,small, leave,my	SENSE 1 (nuclear): weapon,power,war,test,plant,<p>,force,arm, use, energy, reactor,bomb,warhead,missile,soviet,attack,program,explosion, control,waste,u.s.-,(,),ban,fuel,'s,against,nuclear,united,family
SENSE 0: little,indian,which,town,upon,native,where,;,street,through,large,church, of,country, small,man,village,pass,far,mile,walk,reach,our,from,boy,way,young, very, beyond,find	SENSE 1 (weekend): <p>,spend,last,for,(,),",over,on,'s,:,day,during,long,holiday,n't,say, work, next, this, you, home,will,night,voice,plan,year,summer,york,july