The Copenhagen cholera epidemics 1853 in contemporary Danish newspaper prose: Some back-of-an-envelope calculations made during April 2020

Sigfrid Lundberg

The Royal Danish Library

Introduction

Inspired by the fact that we are all hiding away from the Covid-19 I wanted to take a closer look at some other epidemic. My hope was to find patterns of change in language use mirroring sentiments and attitudes expressed in words, bigrams and trigrams and frequency distributions. My first impulse was to analyze the Spanish flu 1918, but since we have as yet little data in our public corpora from that period I turned my attention to the Cholera epidemic in Copenhagen 1853. At the time, Copenhagen had around 130.000[1] inhabitants out of which 7.219 were diagnosed as having the disease out of which 4.737 (56,7%) died. Christian Molbech writes in a letter to Christian Knud Frederik Molbech, the July 3, 1853:

> The first cholera case was found as early as the 12th June (i.e., the day before the opening of the parliament in June) — a young man from the area close to the naval baracks, who for quite some time had worked on a dredge. He has recovered, and was released from hospital the 25th. The authorities were able to conceal the earliest few cases for a fortnight.[2]
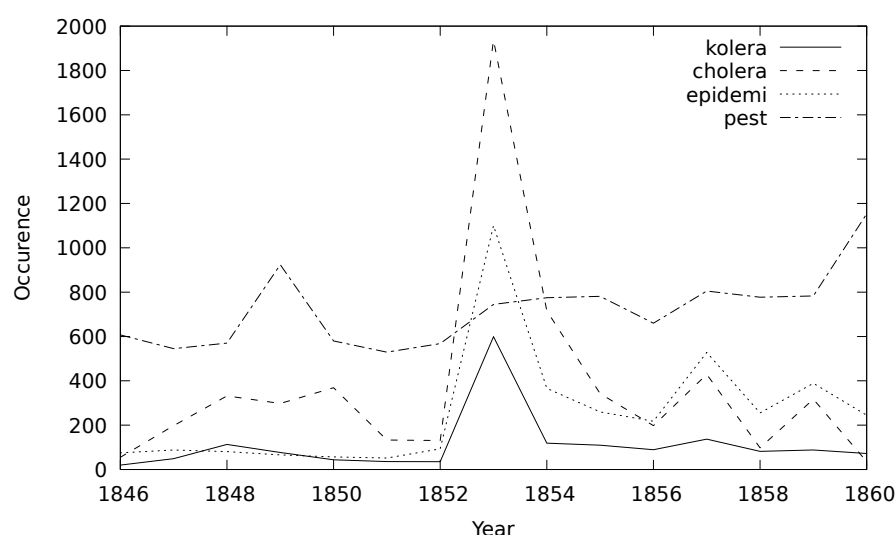

Figure 1. Number of newspaper pages mentioning any of the words *kolera, cholera, epidemi* or *pest*.

---

[1] http://www2.kb.dk/udstillinger/medhist/kolera/koebenhavn1853.htm

[2] Translated by the author. https://bit.ly/3cOzlAg

The epidemic spread from Copenhagen to other areas. Outside the capital 1.951 fatal cases were reported.[3]

The spread of an infectious disease does not occur in a vacuum; the Copenhagen epidemic is a part of the cholera pandemic (1846-1860)[4] which is the third out of seven global outbreaks.[5]

Here I present some data analyses: (1) The number of newspaper articles mentioning *epidemi, pest, kolera* and *cholera* through the period 1846 to 1860 (Figure 1). (2) I repeated that analysis with the frequencies of words rather than articles in weekly aggregations of texts. (3) Finally I visualize a set of bigrams extracted from the corpus.

The data

The data was retrieved from Royal Danish Library's LOAR Repository.[6] It comes in csv format, aggregated to yearly chunks. Each line is basically corresponding to one page of text in the newspaper.

Some low-hanging fruits

Using the Unix `grep` command makes it easy to count the number of pages containing a given word. So the graph in Figure 1, is simple plot of what is extracted using the shell script in Figure 2.

```
#!/bin/sh
echo "#year kolera cholera epidemi pest farsot"
for file in artikler*
do
    year=$(echo $file| tr -d '[:alpha:][:punct:]')
    kolera=$(grep -i kolera $file | wc -l)
    cholera=$(grep -i cholera $file | wc -l)
    epidemi=$(grep -i epidemi $file | wc -l)
    pest=$(grep -i pest $file | wc -l)
    farsot=$(grep -i farsot $file | wc -l)
    echo "$year $kolera $cholera $epidemi $pest $farsot"
done
```

Figure 2. Script for extracting the data for plotting Figure 1

Three things spring to my mind: (1) The obvious one. People wrote about this cholera outbreak in Copenhagen 1853. They wrote a lot. (2) The effect on the public discourse of the epidemic lasted for years and it did not return to pre-epidemic levels that decade. You see this on the shape of the curves; the incidence of words related to the epidemic rise rapidly from very low levels and then fade out slowly. The fact that the article count in Figure 1 never really decreases to counts before the cholera outbreak. Almost certainly the people felt like we do in relation to the Covid-19 epidemic: Will it ever be the same again? (3) People hadn't settled on how to spell the name of the disease yet, cholera and kolera. (4) The word *pest* seems

---

[3] https://da.wikipedia.org/wiki/Koleraepidemien_i_K%C3%B8benhavn_1853

[4] https://en.wikipedia.org/wiki/1846%E2%80%931860_cholera_pandemic

[5] https://en.wikipedia.org/wiki/Cholera_outbreaks_and_pandemics

[6] Newspapers from Royal Danish Library https://loar.kb.dk/handle/1902/157

to be widely used in Danish at the time, possibly as something annoying but also as epidemic disease in general.

I also made another analysis, where I aggregated text by week, and used text tokenized by word, see Figure 3. The week numbering starts with 1 at 1846-01-01 and ends at 1860-12-31 with week 781. The huge peak in the graph starts at 1853-06-24 at week 389. It reaches its highest point at the end of July, 1853-07-29, week 395. From what I can tell it continues until the beginning of October, but the mentions of *cholera* in the corpus is higher than the pre-outbreak into March the following year, week 425. There is a discrepancy between the two analyses, in that the article level analyses (Figure 1) implies a longer effect of the outbreak on the discourse than the per word one (Figure 3.)
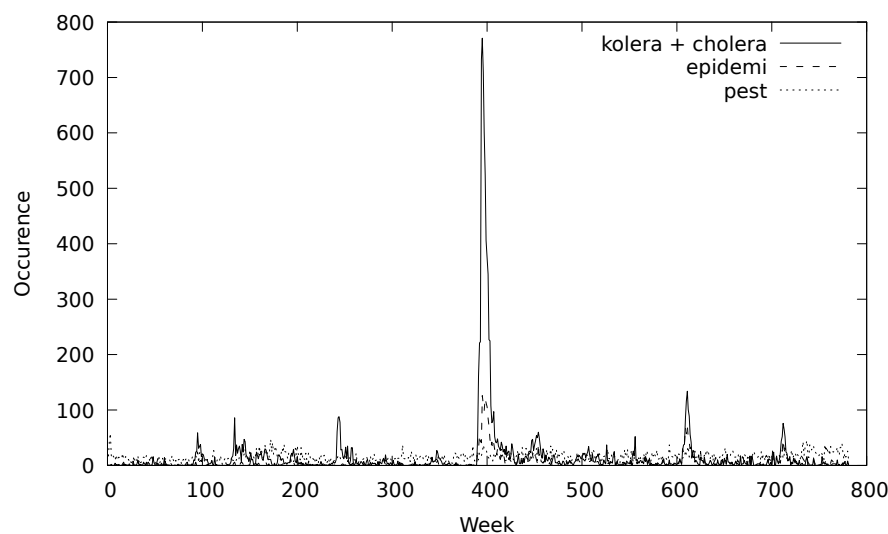
Figure 3. Here I use tokenized text, i.e., there is only one word per line, and I also aggregated the texts by week.

Bigrams and trigrams

N-grams generation is not an end in itself. They are a starting point for many kinds of statistical and machine learning methods and applications in natural language processing.[7] You extract n-grams by "sliding" through your tokenized text word by word, and sample n words. Bigrams and trigrams are special cases where you sample two or three words. I am a great fan of simplicity and elegance and hence I love Kenneth Ward Church's tool box which he described in his wonderful classic *Unix for Poets*[8] where you are given methods to do a text analyses using simple tools that you find on all Unix/Linux machines. Here are some examples of trigrams containg the word *hovedstaden* across the weeks, i.e., the capital (in decreasing frequency):

```
cholera i hovedstaden
hovedstaden bortriver cholera
hovedstaden er cholera
hovedstaden af cholera
hovedstaden forekomne cholera
```

---

[7] https://en.wikipedia.org/wiki/N-gram

[8] http://doc.cat-v.org/unix/for-poets/

```
choleraepidemien i hovedstaden
epidemien i hovedstaden
hovedstaden udbrudte cholera
```

For some reason, I don't understand why, the actual name of the city Copenhagen, København (or whatever spelling) seem to occur rarely in comparison with the word hovedstaden, which is frequent, in particular 1853. Here is another sequence (containing *cholera er*) and sampled such that they begin with that phrase:

```
cholera er af
cholera er afgaaet
cholera er aftagen
cholera er aldeles
cholera er alter
cholera er aner
cholera er at
cholera er atter
cholera er begjeert
cholera erboldes som
cholera er borlfalden
cholera er bortfalden
cholera er cadetfiibs
cholera ere altsaa
cholera ere berovede
```

Figure 4-7 are of a visualizations of the frequencies of various trigrams.

1. Trigrams starting with *kolera i* or *cholera i* ("kolera in" or "cholera in" in English). The third word is often a place name, such as *cholera i hovedstaden* (i.e., "cholera in the capital"). There are other constructs like *cholera i begyndelsen* (i.e., "cholera in the beginning"). Repetitions are excluded, so the heights of the bars are more measures of diversity rather than abundance. See Figure 4.

2. Trigrams starting with *kolera er* or *cholera er* ("kolera" or "cholera" followed by "is") in English). These are constructs like *cholera er udbrudt* ("cholera is broken out"), *cholera er fra* ("cholera is from") etc. See Figure 5.

3. Trigrams starting with *kolera har* or *cholera har* ("kolera" or "cholera" followed by "has") in English). The words following *har* (i.e., "has") is similar to those follwing "cholera is". After all they are both conjugations of the same verb *have*; "has" is still present tense just like "is" but the cholera has been going on for a while when using "has". For instance, we have *cholera har hersket* ("cholera has ruled") as opposed to the "cholera is broken out" as mentioned in point 2 above. See Figure 6.

4. Finally, trigrams starting with *kolera havde* or *cholera havde* ("kolera" or "cholera" followed by "had") in English). One example is *cholera havde vist*, i.e., "cholera had shown". There are only handful words following this past tense construct of the verb *have*. They only appear in five of the ten years I have in my time series. Since the numbers are so low, I am hesitant to draw any conclusions upon the set of words here. However, only one occurence of *cholera havde* is before 1853, in 1849. The rest of them are in 1853, 1854, 1855 and 1856 (See Figure 7).

The *cholera havde* constructs seem to be discussions on why things happened as they did in 1853 as in this text

> Saavel paa St. Thomas som paa St. Croix vedligholdtes streng Karantæne, och intet Tilfælde af **Kolera havde vist** [authors emphasis] sig paa nogen av Øerne.[9] (On St. Thomas as well as on St. Croix there was stringent quarantine, and no cases were encountered on the islands.)

---

[9] http://hdl.handle.net/109.3.1/uuid:dd96317c-bf59-463f-9e63-01ce718ae521

Words

1848	1849	1850	1851	1852	1853	1854	1855	1856	1857	1858
					det					
					den					
					de					
					kjobenhavn					
					en					
					christiania					
					skagen					
					aalborg				portugal	
					stockholm				sverrig	
					skovshoved				kjobenhavn	
					sidste				europa	
					provindserne				svendborg	
					lubeck				slagelse	
					liverpool				skjclskor	
					hovedstaden				s	
					byen				rorsor	
					aarhuus				rjobenhavu	
					a				rbhavn	
					zkovsboved				ognene	
					verrig				og	
		hamborg			uger	til			nr	
		berlin			ugen	stockholm			lorsor	
		udbrud			tydskland	konstantinopel			korsser	
		truende			tverrig	wilster			korsør	
	den	toplitz			tvcrrig	vort			korsor	
	paris	tiltagende			tk	sroctbolm			korsoer	
	dublin	tbplch			ten	selve			kors	
	warfch	t			temmelig	sebastopol			klvrsvr	
	ver	sverrig			tel	rutschuk			kjobcnhavn	
	store	staden			te	regelen			kjøbsnliavn	
de	starkt	saa			taarnby	newyork			kj	
hamborg	sl	panama			taarnbv	metiko			kbhavn	
rusland	peterssy	malms			taarbek	mcrtcv			jtzehoe	
petersborg	petersborg	malmo			sverrig	madrid			jrvrsvr	
stadig	pari	malmø			svendborg	maancder			idornsr	
smitte	op	malm	sommeren		svanlke	london	wien		hjemsogte	
sammenligning	lang	lubeck	smit		svanike	jor	saa		hjemsøgte	
peters	gravciand	loverdags	lcrngere	von	stokholm	ict	rhkjpbiug		hjems	tirsdags
london	galizteu	loll	en	saa	stock	frygtelig	og		hjcmsogte	norsk
honnette	det	henhold	cannot	posen	smetmester	extending	neworlean	st	gliickstadt	lybek
helsingfors	deres	halberstadt	bergen	polen	sinde	dtommelse	n	lu	gliicksladt	lemberg
et	dcn	ben	band	kredsene	sig	den	kjobenhavn	liibeck	fre	gluckstadt
cn	berg	begyndelsen	afdode	hvlsteen	rytiobing	christiania	kbbavn	en	byen	gliickstadt
alle	asta	bandholm	afdede	flere	rusland	buen	gefle	egnen	april	casymirdalen
a	alt	b	aarhuus	aftagende	rjobenl	armeen	berlin	byen	aabenraa	byen

Year

Figure 4. Trigrams per year with the text begining with `cholera i`. The highest bar is slightly massaged. I removed repetions and incomprehensible OCR errors in order to make it fit into the page.

Words

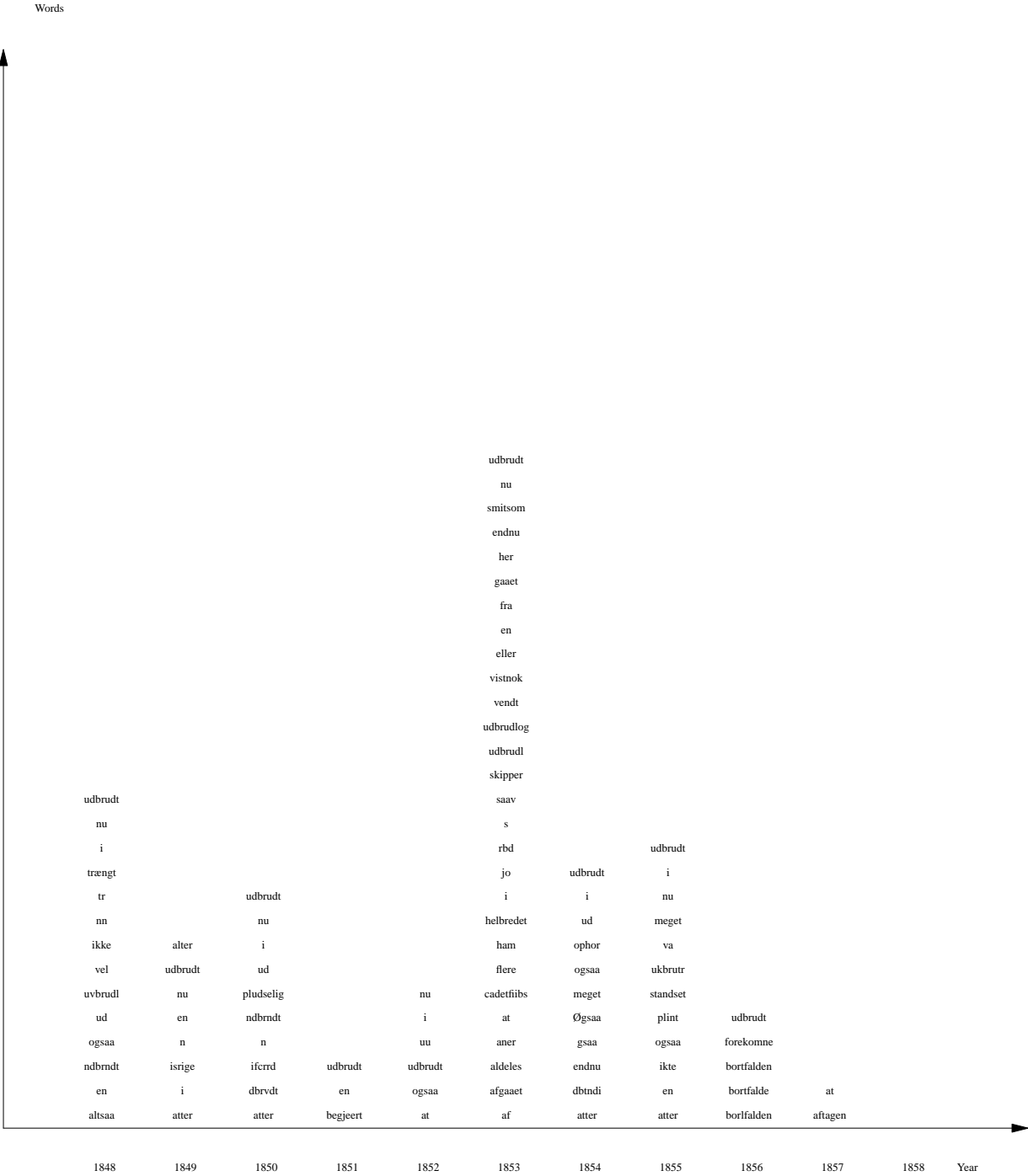|	1848	1849	1850	1851	1852	1853	1854	1855	1856	1857
						udbrudt				
						nu				
						smitsom				
						endnu				
						her				
						gaaet				
						fra				
						en				
						eller				
						vistnok				
						vendt				
						udbrudlog				
						udbrudl				
						skipper				
	udbrudt					saav				
	nu					s				
	i					rbd		udbrudt		
	trængt					jo	udbrudt	i		
	tr		udbrudt			i	i	nu		
	nn		nu			helbredet	ud	meget		
	ikke	alter	i			ham	ophor	va		
	vel	udbrudt	ud			flere	ogsaa	ukbrutr		
	uvbrudl	nu	pludselig		nu	cadetfiibs	meget	standset		
	ud	en	ndbrndt		i	at	Øgsaa	plint	udbrudt	
	ogsaa	n	n		uu	aner	gsaa	ogsaa	forekomne	
	ndbrndt	isrige	ifcrrd	udbrudt	udbrudt	aldeles	endnu	ikte	bortfalden	
	en	i	dbrvdt	en	ogsaa	afgaaet	dbtndi	en	bortfalde	at
	altsaa	atter	atter	begjeert	at	af	atter	atter	borlfalden	aftagen
	1848	1849	1850	1851	1852	1853	1854	1855	1856	1857	1858	Year

Figure 5. Trigrams per year with the text begining with `cholera er`. The highest bar is slightly massaged. I removed repetions and incomprehensible OCR errors in order to make it fit into the page.
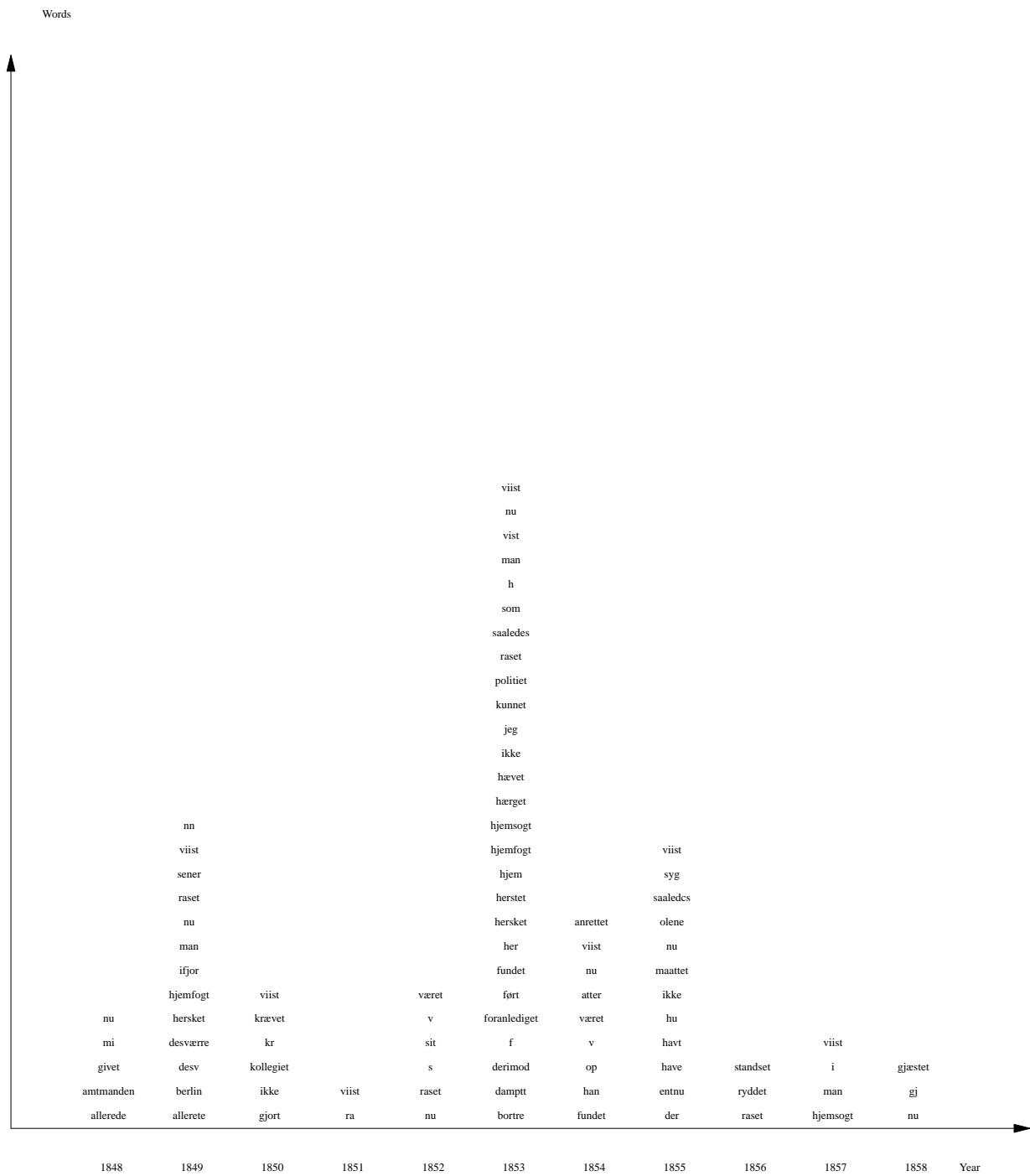
Words

	1848	1849	1850	1851	1852	1853	1854	1855	1856	1857	1858
					viist						
					nu						
					vist						
					man						
					h						
					som						
					saaledes						
					raset						
					politiet						
					kunnet						
					jeg						
					ikke						
					hævet						
					hærget						
	nn				hjemsogt						
	viist				hjemfogt		viist				
	sener				hjem		syg				
	raset				herstet		saaledcs				
	nu				hersket	anrettet	olene				
	man				her	viist	nu				
	ifjor				fundet	nu	maattet				
	hjemfogt	viist		været	ført	atter	ikke				
nu	hersket	krævet		v	foranlediget	været	hu				
mi	desværre	kr		sit	f	v	havt		viist		
givet	desv	kollegiet		s	derimod	op	have	standset	i	gjæstet	
amtmanden	berlin	ikke	viist	raset	damptt	han	entnu	ryddet	man	gj	
allerede	allerete	gjort	ra	nu	bortre	fundet	der	raset	hjemsogt	nu	

Year

Figure 6. Trigrams per year with the text begining with `cholera har`.

Words

til
riist   viist   vist
man   nogle   han   vist   viist

1848   1849   1850   1851   1852   1853   1854   1855   1856   1857   1858   Year

Figure 7. Trigrams per year with the text begining with `cholera havde`.

Lessons learned

1   My idea was that just inspection of the words in their context should give information about emotions or sentiments. I suppose that this was just due to my naivete when it comes to languages, since I cannot really find anything of that kind.

2   The tools coming with a standard Linux distribution isn't as unicode compliant one would wish. Some twenty years later, the nice utility `tr` is still not really UTF-8. Neither is the GNU implementation of Brian Kernighans graphics language `pic` which I used for the bigrams.

3   There is another big problem connected with the use of our newspaper text corpora for computational linguisics. The poor OCR quality.

---

https://en.wikipedia.org/wiki/Pic_language