Can we Quantify Domainhood?

Exploring Measures to Assess Domain-Specificity in Web Corpora

Marina Santini

marina.santini@ri.se

RISE Research Institutes of Sweden

RISE Research Institutes of Sweden, Division ICT - RISE SICS East, Sweden

TIR 2018, Regensburg, Germany, 4 Sept. 2018

Acknowledgement and Citation

Acknowledgement

This research was supported by **E-care@home**, a “SIDUS - Strong Distributed Research Environment” project, funded by the Swedish Knowledge Foundation. Project websit: <<http://ecareathome.se/>>.

Cite the paper as:

Santini M., Strandqvist W., Nyström M., Alirezai M., Jönsson A. (2018) Can We Quantify Domainhood? Exploring Measures to Assess Domain-Specificity in Web Corpora. In: Elloumi M. et al. (eds) Database and Expert Systems Applications. DEXA 2018. Communications in Computer and Information Science, vol 903. Springer, Cham

DOI https://doi.org/10.1007/978-3-319-99133-7_17

Presented at : TIR 2018: 15th International Workshop on Technologies for Information Retrieval. In conjunction with DEXA 2018. Regensburg, Germany, 4 Sept. 2018.

Outline

1. Research questions: evaluating specialized web corpora in terms of ”domainhood”
2. Case study: a web corpus for eCare
3. Methodology: how to measure domainhood
4. Conclusion & Future work

1. Evaluating specialized web corpora in terms of ”domainhood”

Introduction and Research Questions

Web Corpora

- Web corpora are important
- The evaluation of web corpora is important
- The evaluation of general-purpose web corpora is advanced
- The evaluation of specialized web corpora is less advanced

Quantitative Corpus Evaluation

“When will a grammar based on one corpus be valid for another? How much will it cost to port a Natural Language Processing (NLP) application from one domain, with one corpus, to another, with another?”

Adam Kilgarriff (2001) Comparing corpora. Int. J. Corpus Linguist. **6**(1), 97–133

Definition: *domainhood*

- ***Domainhood* is** the degree of domain representativeness or *domain specificity* of a web corpus.
  - Ex: a high frequency of medical terms is a sign that the corpus is a specialized medical corpus
- The importance of *domain granularity*
  - Coarse domains vs fine-grained domains
    - Lippincott et al. (2011) “while variation at a coarser domain level such as between newswire and biomedical text is well-studied and known to affect the portability of NLP systems, there is a need to develop an awareness of subdomain variation when considering the practical use of language processing applications […]”.

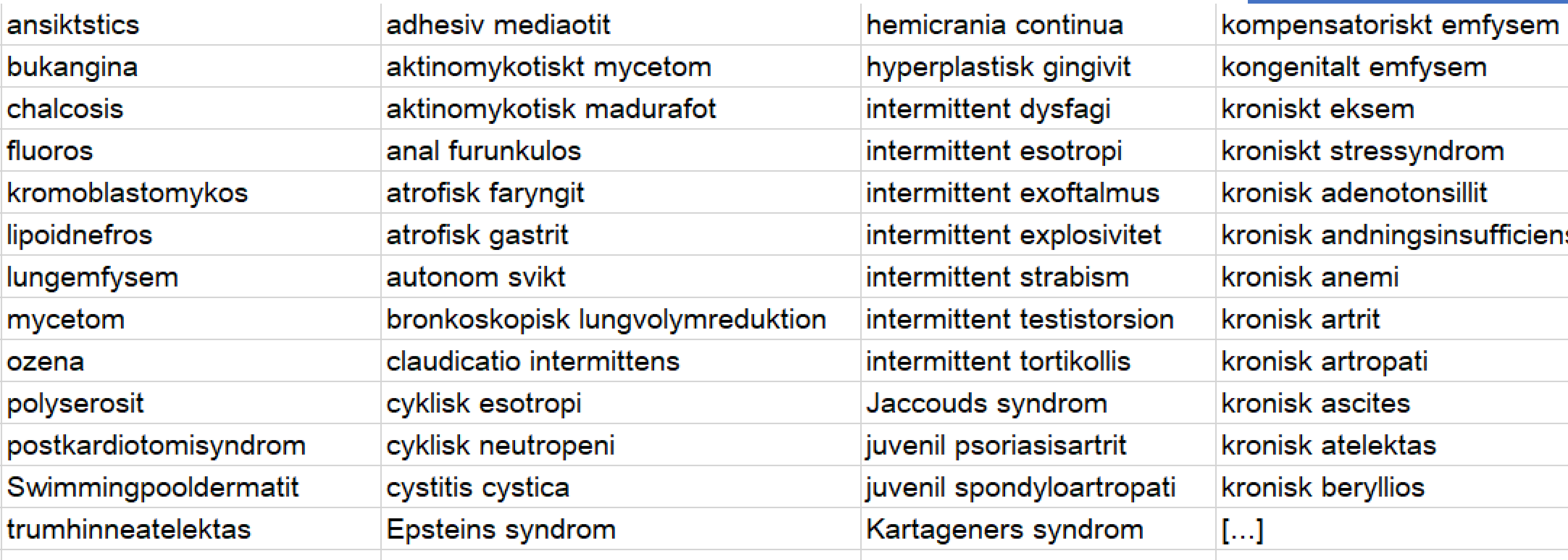
Research Questions: Quantifying *domainhood*

- ”is it possible to automatically quantify the *domainhood* of a web corpus regardless its *domain granularity*? If so, how?”

2. Case Study: a Web Corpus for eCare

eCare_sv_01

- eCare_sv_01*
    - 155 SNOMED CT terms (chronic diseases)



ansiktstics	adhesiv mediaotit	hemicrania continua	kompensatoriskt emfysem
bukangina	aktinomykotiskt mycetom	hyperplastisk gingivit	kongenitalt emfysem
chalcosis	aktinomykotisk madurafot	intermittent dysfagi	kroniskt eksem
fluoros	anal furunkulos	intermittent esotropi	kroniskt stressyndrom
kromoblastomykos	atrofisk faryngit	intermittent exoftalmus	kronisk adenotonsillit
lipoidnefros	atrofisk gastrit	intermittent explosivitet	kronisk andningsinsufficien
lungemfysem	autonom svikt	intermittent strabism	kronisk anemi
mycetom	bronkoskopisk lungvolymreduktion	intermittent testistorsion	kronisk artrit
ozena	claudicatio intermittens	intermittent tortikollis	kronisk artropati
polyserosit	cyklisk esotropi	Jaccouds syndrom	kronisk ascites
postkardiotomisyndrom	cyklisk neutropeni	juvenil psoriasisartrit	kronisk atelektas
Swimmingpooldermatit	cystitis cystica	juvenil spondyloartropati	kronisk beryllios
trumhinneatelektas	Epsteins syndrom	Kartageners syndrom	[...]

* Santini M., Jönsson A., Nystrom M. and Alirezai M. (2017) "A Web Corpus for eCare: Collection, Lay Annotation and Learning. First Results". Proceedings of LTA'17, FedCSIS 2017, Prague.

3. Methodology: How to Measure Domainhood

Which measures?

SUC & eCare_sv_01

Stockholm-Umeå Corpus (SUC) -> reference corpus (1 million words)

eCare_sv_01: domain-specific corpus (approx. 700 000 words)

Metrics

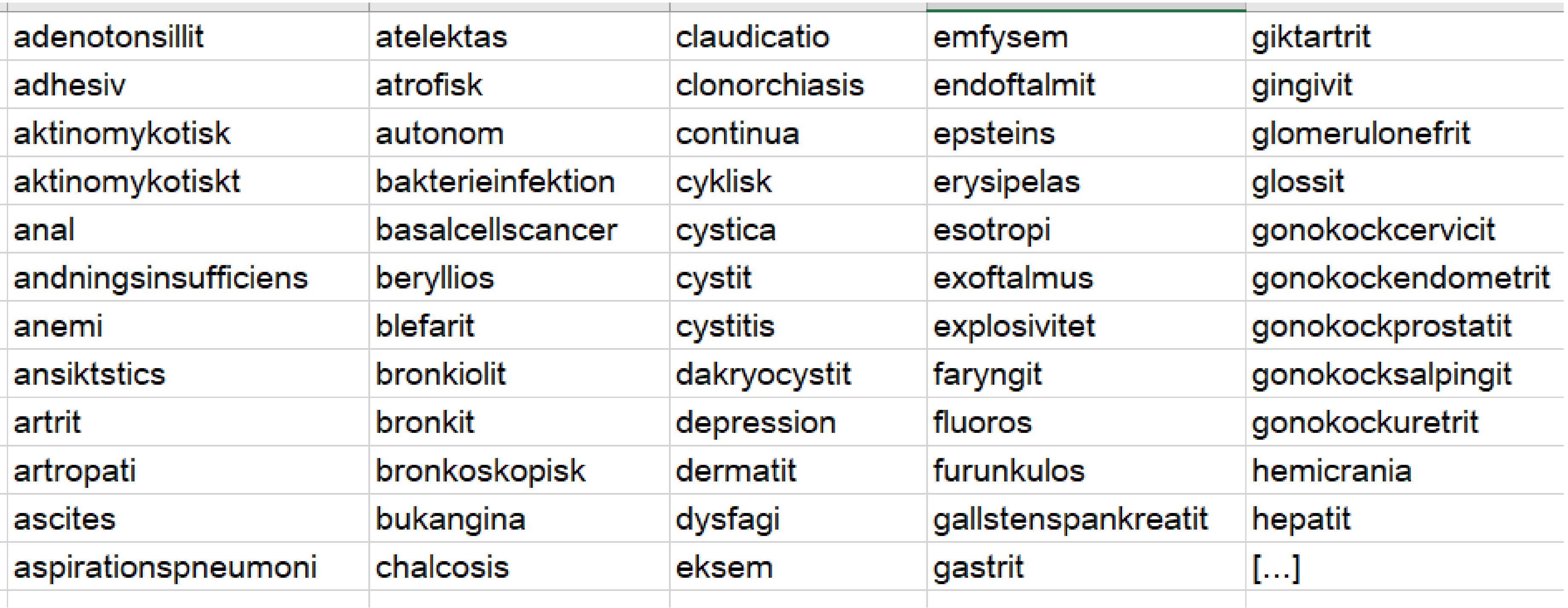
1.Mann-Withney-Wilcoxon Test
2.Kendall correlation coefficient (τ)
3.Kullback–Leibler (KL) divergence
4.Log-likelihood
5.Burstiness

Gold Standard

Seeds (example):
atrofisk faryngit
atrofisk gastrit

Gold Standard (example)
atrofisk
faryngit
gastrit

Tokenized gold standard (165 unigrams)

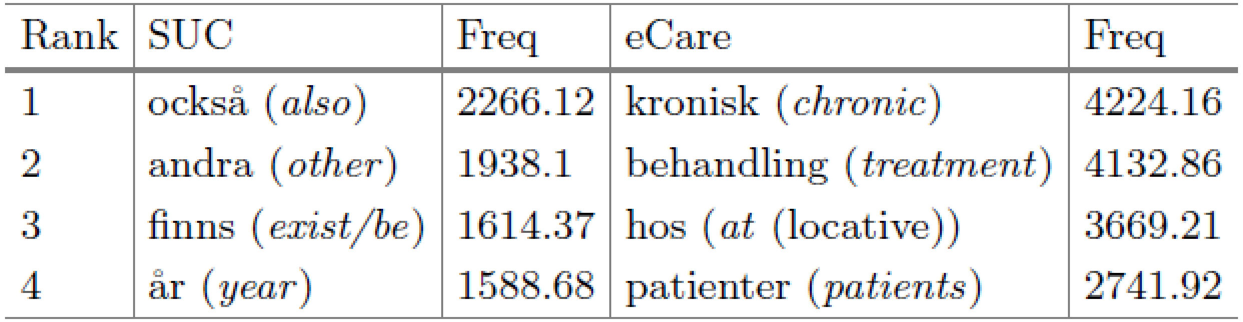


adenotonsillit
adhesiv
aktinomykotisk
aktinomykotiskt
anal
andningsinsufficiens
anemi
ansiktstics
artrit
artropati
ascites
aspirationspneumoni
atelektas
atrofisk
autonom
bakterieinfektion
basalcellscancer
beryllios
blefarit
bronkiolit
bronkit
bronkoskopisk
bukangina
chalcosis
claudicatio
clonorchiasis
continua
cyklisk
cystica
cystit
cystitis
dakryocystit
depression
dermatit
dysfagi
eksem
emfysem
endoftalmit
epsteins
erysipelas
esotropi
exoftalmus
explosivitet
faryngit
fluoros
furunkulos
gallstenspankreatit
gastrit
giktartrit
gingivit
glomerulonefrit
glossit
gonokockcervicit
gonokockendometrit
gonokockprostatit
gonokocksalpingit
gonokockuretrit
hemicrania
hepatit
[…]

Word Frequency Lists

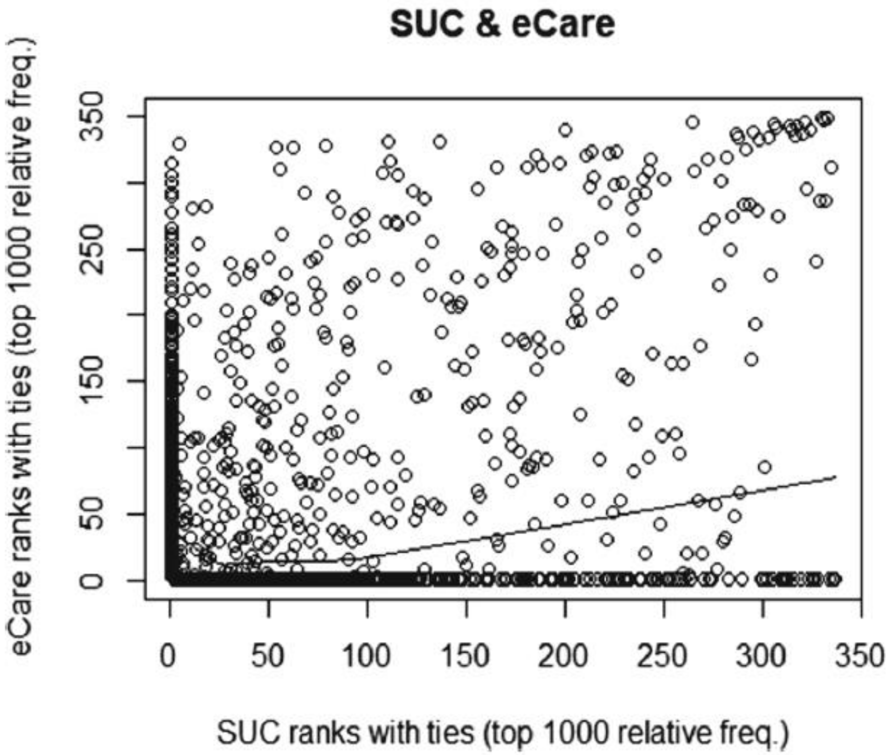
“A word frequency list is a “compact representation of a corpus, lacking much of the information in the corpus but small and easily tractable.”

Adam Kilgarriff (2010). Comparable corpora within and across languages, word frequency lists and the KELLY project. In: Proceedings of the 3rd Workshop on Building and Using Comparable Corpora.



Rank	SUC	Freq	eCare	Freq
1	också (*also*)	2266.12	kronisk (*chronic*)	4224.16
2	andra (*other*)	1938.1	behandling (*treatment*)	4132.86
3	finns (*exist/be*)	1614.37	hos (*at* (locative))	3669.21
4	år (*year*)	1588.68	patienter (*patients*)	2741.92

Ranked Word Frequencies: Scatter Plot



Mann-Withney-Wilcoxon Test: Theory

Non-parametric test:
Using the Mann-Whitney-Wilcoxon Test, we can decide whether the population distributions are identical without assuming them to follow the normal distribution.

If the two distributions are dissimilar at .05 significance level, we can conclude that SUC and eCare come from different populations.

Mann-Withney-Wilcoxon Test: Results

- **The null hypothesis** is that SUC's word frequency list and eCare_sv_01 word frequency list come from identical populations.

- To test the hypothesis, we apply the *wilcox.test()* *[R* function] to compare the corpora.
- The p-value turns out to be **0.019**, and is less than the .05 significance level, we reject the null hypothesis.

- Conclusion: at .05 significance level, we conclude that SUC and eCare belong to non-identical populations.

Kendall correlation coefficient: Theory

Kendall correlation coefficient (tau) is a non-parametric measure of correlation between two rankings.

tau is a probability value which indicates the difference between 2 rankings.

(We used the R function "*cor.test()*" with method="kendall" to calculate the test).

Interpretation:

- -1 = strong negative correlation
- 0 = no association
- 1 strong positive correlation

Kendall correlation coefficient: Results

Null hypothesis: the two rankings are identical

(We used the “*cor.test()*” R function with method=“kendall”, "two.sided“ a to calculate the test.

tau -0.1093077;

the p-value of the test is 0.000000003122 (p-value in R: 3.122e−09) which is less than the significance level *p* = .05.

We reject the null hypothesis:
If the rankings of SUC and eCare’s word frequency lists are dissimilar at .05 significance level, we can conclude that the content of eCare is different from SUC.

Kullback–Leibler (KL) Divergence: Theory

(a.k.a. relative entropy)

- KL quantifies how "distant" an estimation of a distribution may be from the true distribution.

- Interpretation: *KL divergence is non-negative and equal to zero if the two distributions are identical*.

Kullback–Leibler (KL) Divergence: Results

- (We do not need a null hypothesis)

- (We used the R function "*KL.empirical()*", (log2), package "entropy" to compute KL divergence).

The KL divergence between SUC and eCare_Sv_01 is **5.80**

Up to now...

- ... the gold standard was not involved

- It is confirmed that two corpora were largely different, but we do not know whether eCare is representative of the target domain.

Log-Likelihood (LL): Theory

(a.k.a. G2)

A reference corpus is needed.

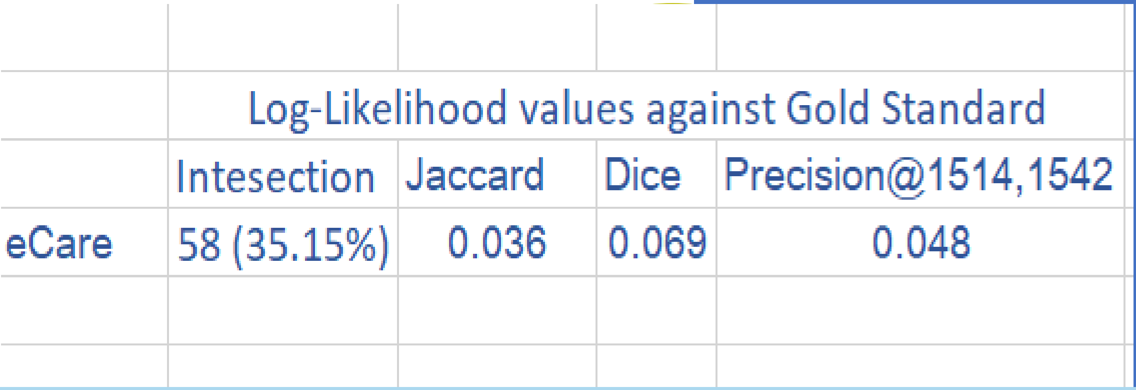
It is a measure based on a contingency table and compares the expected values in two corpora under observation.

Interpretation: *The larger the LL score of a word, the more different its distribution in the two corpora*.

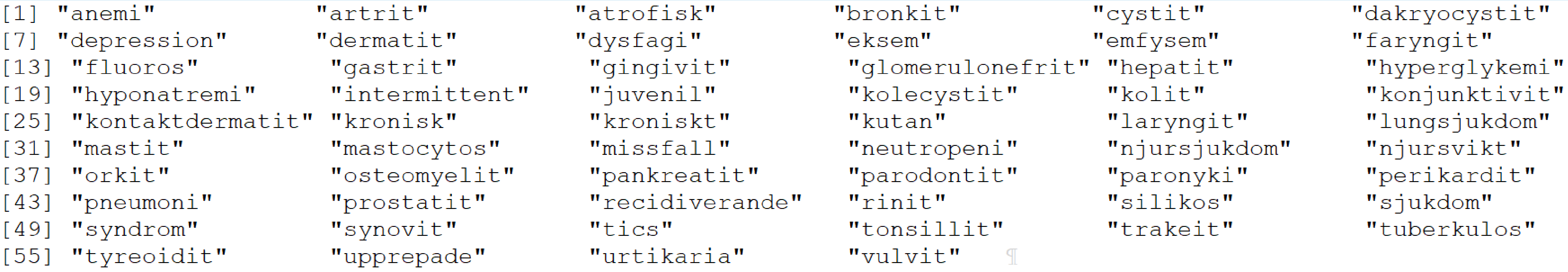
A LL score of 3.8415 or higher is significant at the level of <0.05 and a LL score of 10.8276 is significant at the level of <0.001 (Desagulier, 2017).

Log-Likelihood (LL): Results

The intersection between LL scores and the gold standard is 58, i.e. 35.15%.



	Log-Likelihood values against Gold Standard			
	Intesection	Jaccard	Dice	Precision@1514,1542
eCare	58 (35.15%)	0.036	0.069	0.048



```
[1]  "anemi"             "artrit"            "atrofisk"          "bronkit"           "cystit"            "dakryocystit"
[7]  "depression"        "dermatit"          "dysfagi"           "eksem"             "emfysem"           "faryngit"
[13]  "fluoros"           "gastrit"           "gingivit"          "glomerulonefrit"  "hepatit"           "hyperglykemi"
[19]  "hyponatremi"       "intermittent"      "juvenil"           "kolecystit"        "kolit"             "konjunktivit"
[25]  "kontaktdermatit"  "kronisk"           "kroniskt"          "kutan"             "laryngit"          "lungsjukdom"
[31]  "mastit"            "mastocytos"        "missfall"          "neutropeni"        "njursjukdom"       "njursvikt"
[37]  "orkit"             "osteomyelit"       "pankreatit"        "parodontit"        "paronyki"          "perikardit"
[43]  "pneumoni"          "prostatit"         "recidiverande"     "rinit"             "silikos"           "sjukdom"
[49]  "syndrom"           "synovit"           "tics"              "tonsillit"         "trakeit"           "tuberkulos"
[55]  "tyreoidit"         "upprepade"         "urtikaria"         "vulvit"   ¶
```

Burstiness: Theory

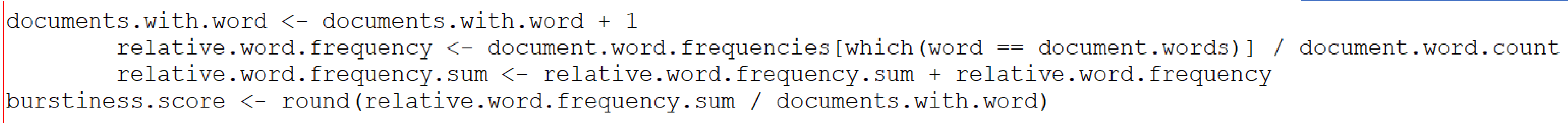
Burstiness helps identify words that are frequent in certain documents, but that are unevenly distributed in the corpus as a whole.

”Burstiness is like the mean but it ignores documents with no intances” (Church and Gale, 1995)

$$B_w = \frac{\sum_{d_i \in D} rf_{w_{d_i}}}{df_w}$$

Irvine, A., & Callison-Burch, C. A (2017) Comprehensive Analysis of Bilingual Lexicon Induction. *Computational Linguistics*, *43*(2).

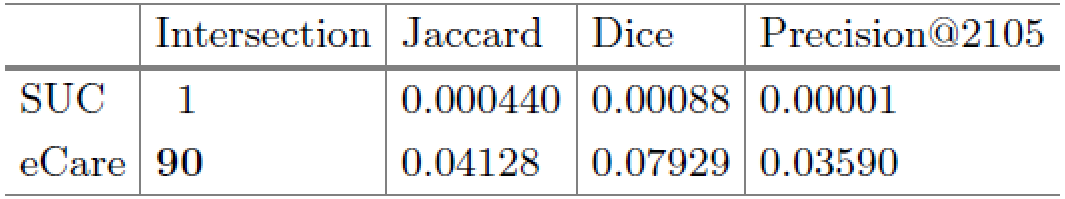
Implementation in R:



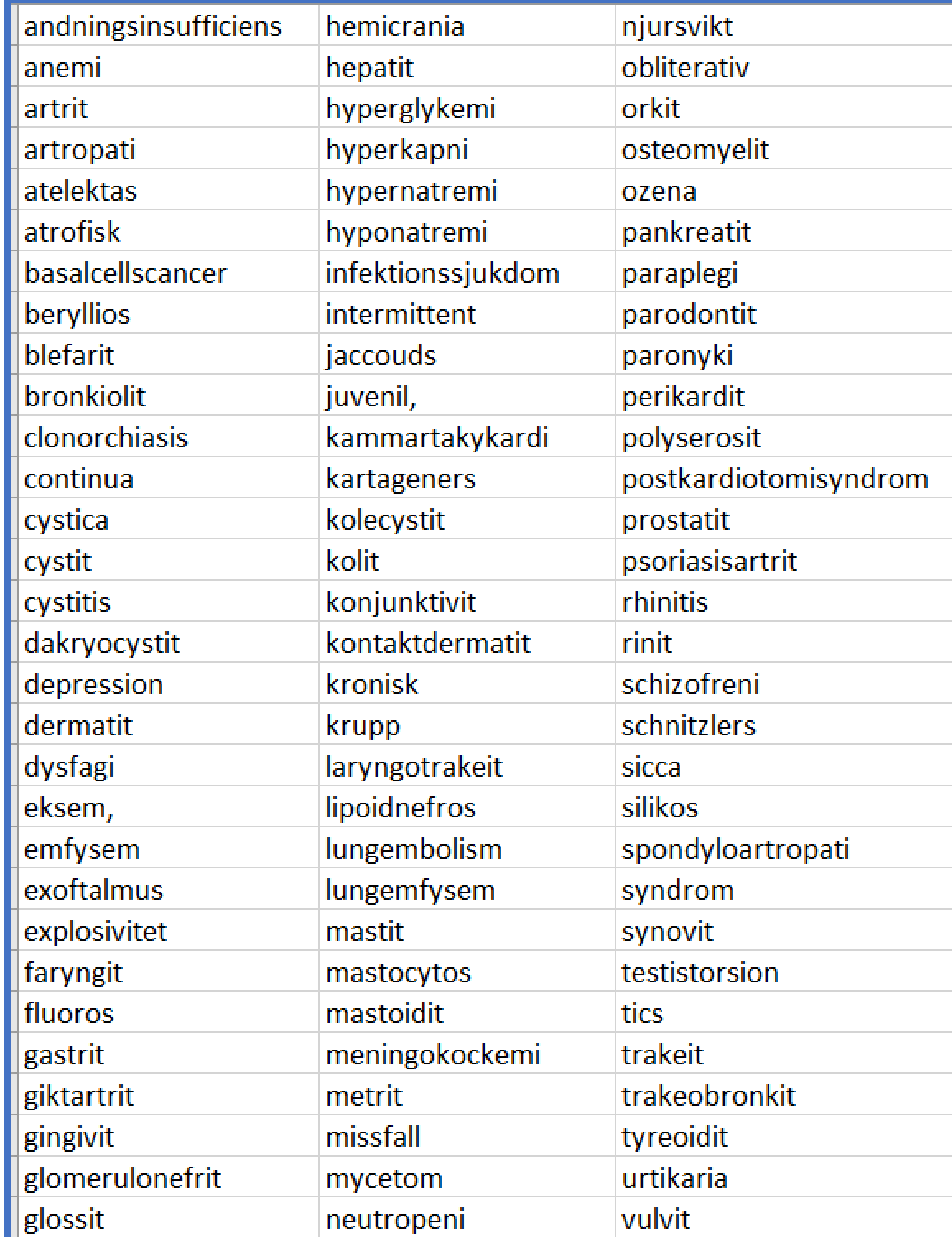
```
documents.with.word <- documents.with.word + 1
        relative.word.frequency <- document.word.frequencies[which(word == document.words)] / document.word.count
        relative.word.frequency.sum <- relative.word.frequency.sum + relative.word.frequency
burstiness.score <- round(relative.word.frequency.sum / documents.with.word)
```

Burstiness: Results

Comparison between *bursty* words and the chronic diseases' gold standard



	Intersection	Jaccard	Dice	Precision@2105
SUC	1	0.000440	0.00088	0.00001
eCare	**90**	0.04128	0.07929	0.03590



andningsinsufficiens
anemi
artrit
artropati
atelektas
atrofisk
basalcellscancer
beryllios
blefarit
bronkiolit
clonorchiasis
continua
cystica
cystit
cystitis
dakryocystit
depression
dermatit
dysfagi
eksem,
emfysem
exoftalmus
explosivitet
faryngit
fluoros
gastrit
giktartrit
gingivit
glomerulonefrit
glossit

hemicrania
hepatit
hyperglykemi
hyperkapni
hypernatremi
hyponatremi
infektionssjukdom
intermittent
jaccouds
juvenil,
kammartakykardi
kartageners
kolecystit
kolit
konjunktivit
kontaktdermatit
kronisk
krupp
laryngotrakeit
lipoidnefros
lungembolism
lungemfysem
mastit
mastocytos
mastoidit
meningokockemi
metrit
missfall
mycetom
neutropeni

njursvikt
obliterativ
orkit
osteomyelit
ozena
pankreatit
paraplegi
parodontit
paronyki
perikardit
polyserosit
postkardiotomisyndrom
prostatit
psoriasisartrit
rhinitis
rinit
schizofreni
schnitzlers
sicca
silikos
spondyloartropati
syndrom
synovit
testistorsion
tics
trakeit
trakeobronkit
tyreoidit
urtikaria
vulvit

Discussion

- Both statistical tests confirm that the two corpora are weakly correlated. No gold standard involved, but based on a Null Hypothesis

- KL divergence returns a large value that indicate that the two corpora are distant from each other. No gold standard involved.

- LL scores needs a reference corpus. They single out words with different distributions in two corpora, results are compared against a gold standard, but it is not clear to which corpus the the words that are singled out belong to.

- Burstiness can be computed without a reference corpus. Results can be measured against a gold standard. Provides promising results.

Profiling Bursty Words: Open Issues

- Less empirical cut-off points.

- Is burstiness affected by the size of corpus?

- Evaluation metrics (overlap coefficients and precision@) are not so indicative. Intersection gives a better idea of the quantification.

- The best way to test the design of gold standards (=target domains) for this kind of experiments.

4. Conclusion and Future Work

What next?

Conclusion

Mann-Withney-Wilcoxon Test: hypothesis testing on distributions

Kendall correlation coefficient: hypothesis testing on rank correlation

Kullback–Leibler (KL) divergence: requires a reference corpus, cannot be tested on a gold standard

Log-likelihood: requires a reference corpus, can be tested on a gold standard

Burstiness: does not require a reference corpus and can be tested on a gold standard

Future Work

- Implementation of additional burstiness formulas
- Inclusion of multi-words in the frequency lists
- Application of burstiness for domainhood on larger corpora and other languages
- Investigating the ideal design of a gold standard for domainhood detection

Thanks for your attention !

RI.
SE