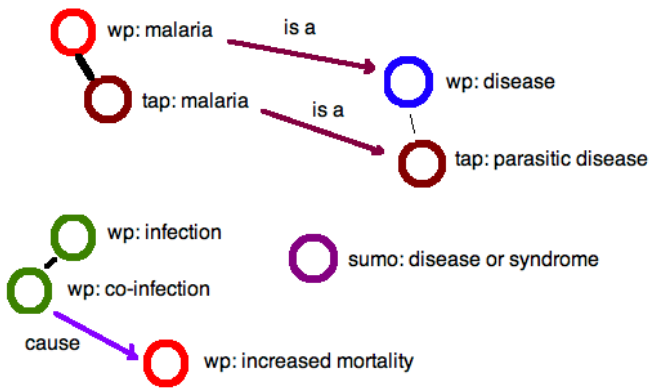
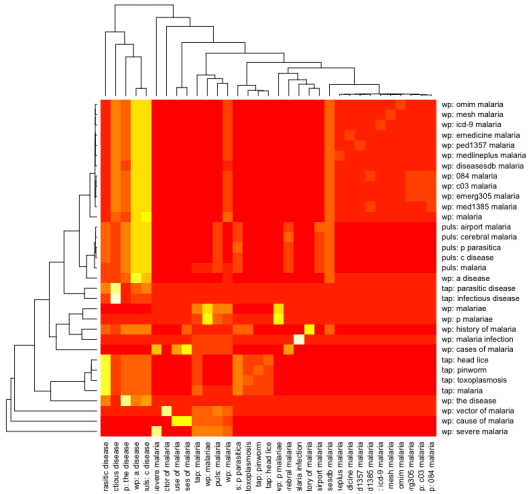




estimating term similarity and coverage

a statistical journey with syntactic evidence

Seppo Nyrkkö
PhD student, UHEL
MOLTO open day Sept. 2011

The Question

Many quasi-similar terms are used in multiple sources.
Can we compute when two terms share a common meaning?

Sources: Ontologies / Text corpora

How to divide a continuum of similar terms?

In which ways can a term be interpreted?

How to estimate term similarities across languages?

Applications

Term Harvesting
Domain Terminology Validation

Information extraction
Concept disambiguation

Ontology development
- harvesting
- systems adaptation
- alignment / merging / filling

The approach

- Each term is unique inside their context.
- Some terms are more similar than others.
- Terms acting similarly may be an evidence of similarity!

Goals:

- Compute Similarity / Coverage
- Find the necessary features for semantic space
- Analyze the methods for clustering / disambiguation

Computation

Distance initialization: "Similar appearance"

- string similarity, using edit distance

Evidence: "Similar behavior of terms"

- having a common frequent co-occurrence
- having a common frequent syntactic dependency
- having a common frequent syntactic role
- synonymy, using dictionaries
- having a common property in an ontology

Convergence:

- "Fuzzy mapping"
- "Overall term distance distribution" cost function

Related research

Likey
use document word (~ "morpheme") frequencies to "feature" topics, e.g. dictionary articles, clustering into ontologies

PuLS
use syntactical analysis to detect terms and patterns, towards information extraction / spotting

Development example: Medical domain

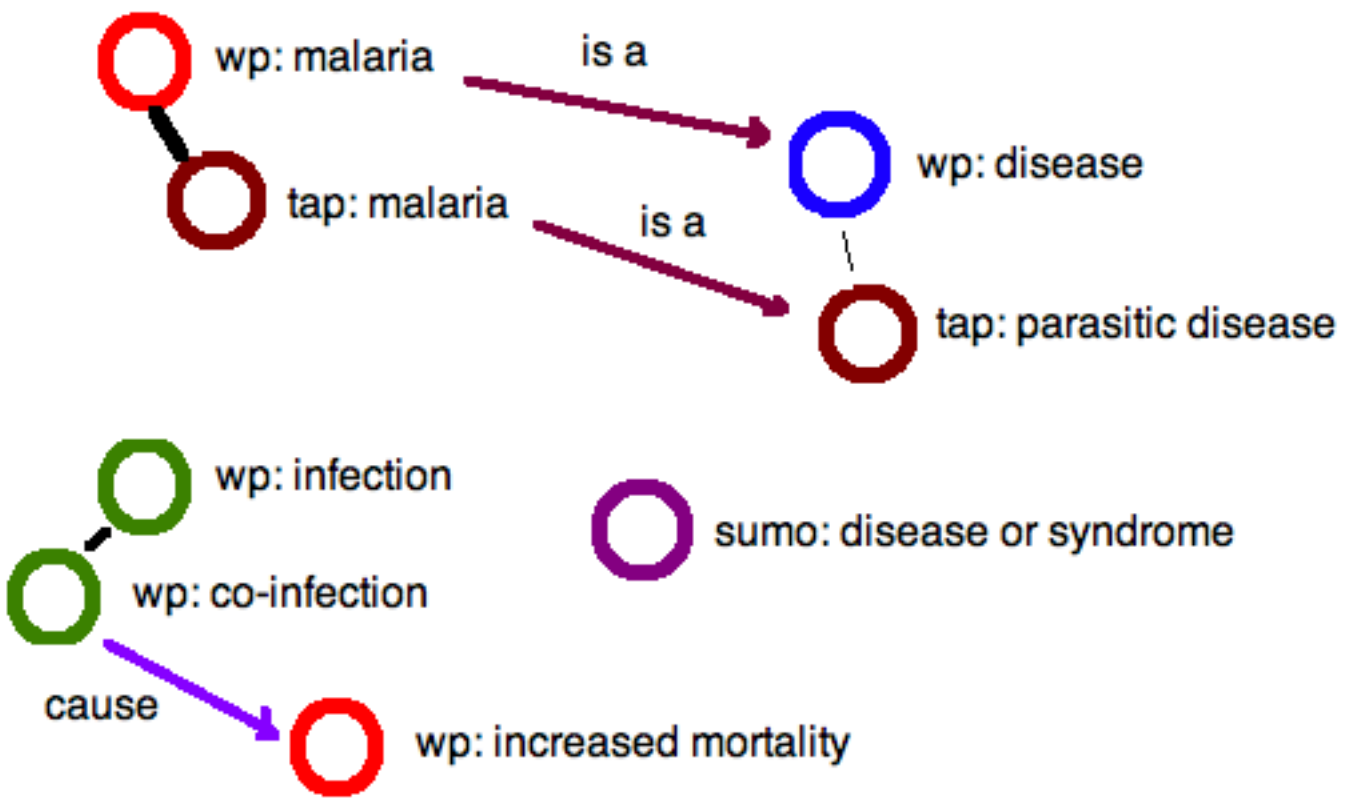
sources in example:

PuLS medical ontology
TAP ontology
SUMO ontology
Some wikipedia articles (incl. Malaria(en))

"semantic space"

Problem: find similarity weights (inverse distance)

Initialize by string edit distance ... then analyze the data

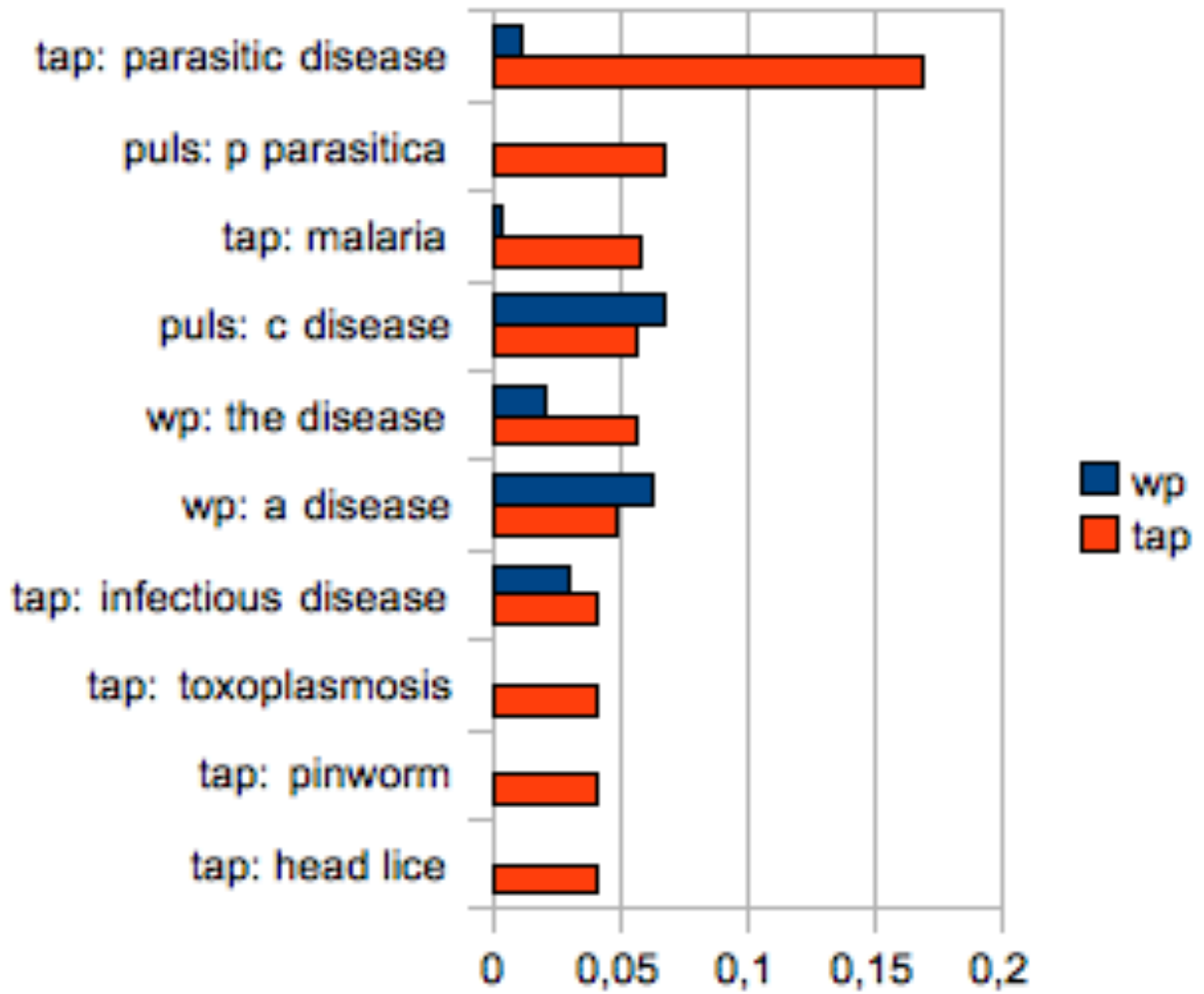


example: Behavior of Malaria terms

wp: malaria

tap: malaria

is a...

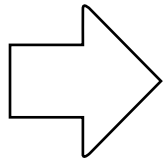


Experiment: Screenshots

corpus

Malaria is a mosquito-borne infectious disease of humans caused by eukaryotic protists of the genus Plasmodium. It is widespread in tropical and subtropical regions, including much of Subsaharan Africa, Asia and the Americas. The disease results from the multiplication of malaria parasites within red blood cells, causing symptoms that typically include fever and headache, in severe cases progressing to coma, and death.

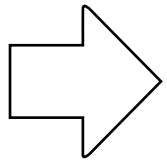
syntactic dependency analysis



```
nsubj(disease-21, Malaria-16)
cop(disease-21, is-17)
det(disease-21, a-18)
amod(disease-21, mosquito-borne-19)
amod(disease-21, i
prep_of(disease-21
partmod(humans-23,
amod(protists-27,
agent(caused-24, p
det(Plasmodium-31,
amod(Plasmodium-31
prep_of(protists-2
```

```
amod(w_1950s-41,w_late-40).
amod(w_congo-37,w_belgian-36).
amod(w_daughter-22,w_youngest-21).
amod(w_missionary-27,w_american-25).
amod(w_missionary-27,w_christian-26).
amod(w_tablets-47,w_quinine-46).
appos(w_bible-12,w_1998-14).
appos(w_bible-12,w_may-18).
aux(w_live-33,w_to-32).
conj_and(w_tablets-47,w_malaria-50).
cop(w_novel-9,w_s-8).
dep(w_bible-12,w_daughter-22).
dep(w_stops-43,w_-4).
dep(w_stops-43,w_139-2).
det(w_1950s-41,w_the-39).
det(w_congo-37,w_the-35).
det(w_daughter-22,w_the-20).
det(w_missionary-27,w_an-24).
dobj(w_brings-29,w_family-31).
dobj(w_taking-44,w_malaria-50).
dobj(w_taking-44,w_tablets-47).
infmod(w_family-31,w_live-33).
nn(w_bible-12,w_poisonwood-11).
nn(w_kingsolver-7,w_barbara-6).
nn(w_malaria-50,w_contracts-49).
nn(w_may-18,w_ruth-17).
nsubj(w_brings-29,w_missionary-27).
nsubj(w_novel-9,w_kingsolver-7).
nsubj(w_stops-43,w_bible-12).
poss(w_family-31,w_his-30).
poss(w_tablets-47,w_her-45).
prep_in(w_live-33,w_congo-37).
prep_of(w_congo-37,w_1950s-41).
prep_of(w_daughter-22,w_missionary-27).
prepc_in(w_stops-43,w_novel-9).
rcmod(w_missionary-27,w_brings-29).
stem(s-41,s).
stem(w_-4,w_).
stem(w_american-25,w_american).
stem(w_an-24,w_an).
stem(w_barbara-6,w_barbara).
stem(w_belgian-36,w_belgian).
stem(w_bible-12,w_bible).
stem(w_brings-29,w_brings).
stem(w_christian-26,w_christian).
stem(w_congo-37,w_congo).
stem(w_contracts-49,w_contracts).
stem(w_daughter-22,w_daughter).
stem(w_family-31,w_family).
```

semantic pattern matching



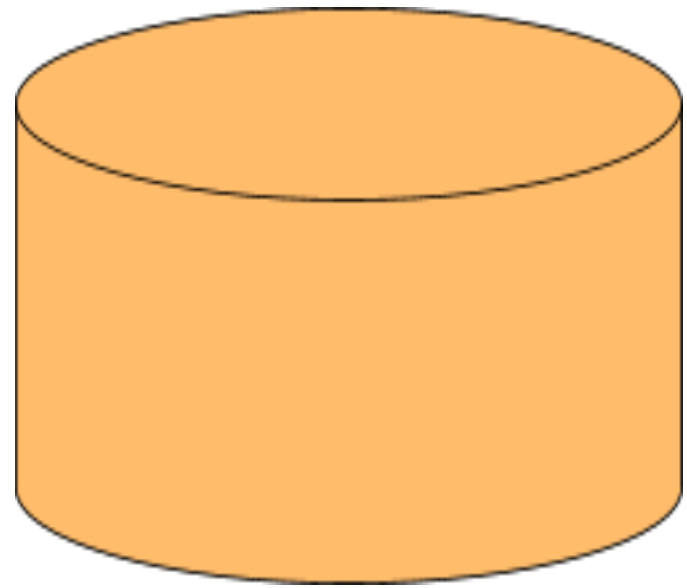
```
hasname(Token,Name) :- =([Token],Name).
hasname(Token,Name) :- append([NNToken],[Token],Name), nn(Token,NNToken).
hasname(Token,Name) :- append([NNToken],[Token],Name), amod(Token,NNToken).
hasname(Token,Name) :- append([NNToken],[Token],Name), det(Token,NNToken).
hasname(Token,Name) :- append([Token],[w_of,NNToken],Name), prep_of(Token,NNToken).


cmember(Member,Class) :- nsubj(Class,Member), cop(Node,_), det(Node,_).
memberof(Mname,Cname) :- cmember(Member,Class), hasname(Member,Mname), hasname(Class,Cname).

cinvolve(Subject,Target) :- nsubj(V,Subject), dobj(V,Target), stem(V,w_involve).
involves(Sname,Tname) :- cinvolve(Stoken,Ttoken),
                         hasname(Stoken,Sname),
                         hasname(Ttoken,Tname).

ccause(Subject,Target) :- nsubj(Node,Subject),dobj(Node,Target),stem(Node,w_cause).
causes(Sname,Tname) :- ccause(Stoken,Ttoken),hasname(Stoken,Sname),hasname(Ttoken,Tname).
```

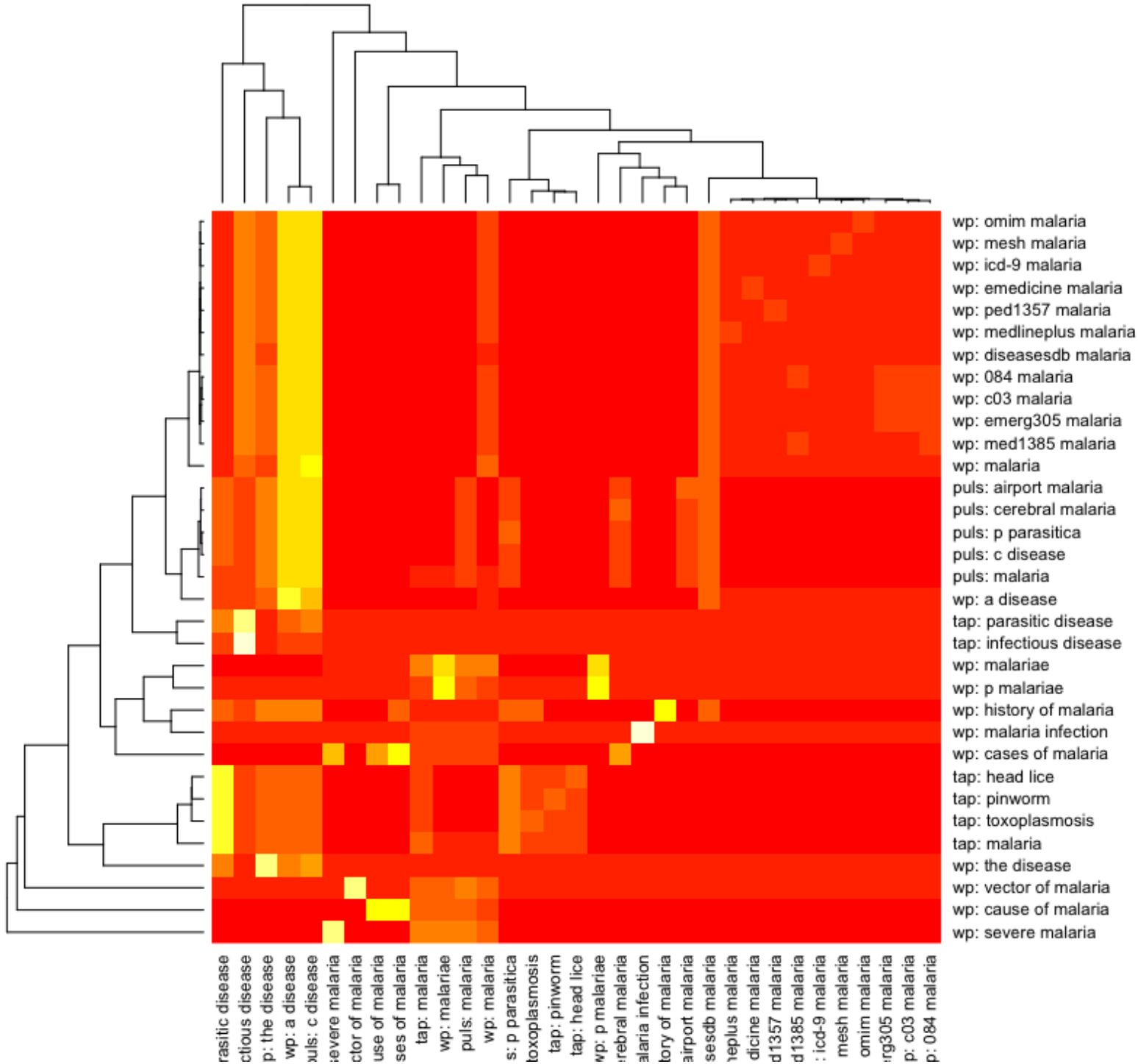
ontology based data

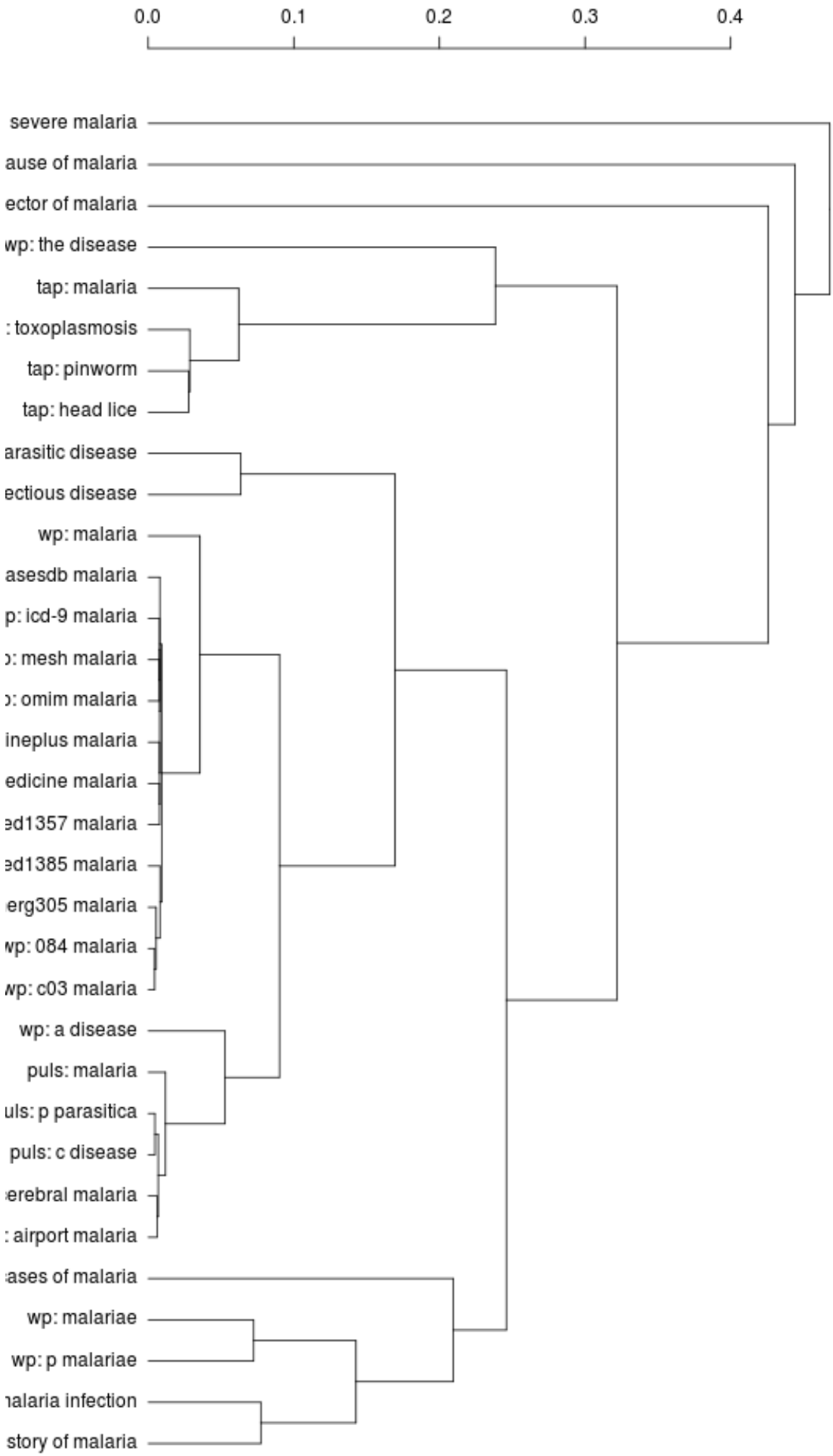
```
reltype source target conf
member "tap: bacterial disease" "tap: infectious disease" 1
member "tap: infectious disease" "tap: health disorder" 1
member "tap: fungal disease" "tap: infectious disease" 1
member "tap: toxoplasmosis" "tap: parasitic disease" 1
member "tap: sexually transmitted disease" "tap: infectious disease" 1
member "tap: parasitic disease" "tap: infectious disease" 1
member "tap: pinworm" "tap: parasitic disease" 1
member "tap: whooping cough" "tap: infectious disease" 1
member "tap: malaria" "tap: parasitic disease" 1
member "tap: viral illness" "tap: infectious disease" 1
member "tap: head lice" "tap: parasitic disease" 1
member "tap: lyme disease" "tap: infectious disease" 1
member "tap: encephalitis" "tap: infectious disease" 1
member "tap: scabies" "tap: parasitic disease" 1
member "puls: vaginal yeast infection" "puls: c disease" 1
member "puls: ttv infection" "puls: c disease" 1
member "puls: histoplasmosis" "puls: c disease" 1
member "puls: new york 1 virus infection" "puls: c disease" 1
member "puls: rsv infection" "puls: c disease" 1
member "puls: pork tapeworm infection" "puls: c disease" 1
member "puls: inflammatory bowel disease" "puls: c non infectious disease" 1
member "puls: p parasitica" "puls: c disease" 1
member "puls: malaccensis" "puls: c disease" 1
member "puls: non polio enterovirus infection" "puls: c disease" 1
member "puls: cancer" "puls: c non infectious disease" 1
member "puls: equine infectious anemia" "puls: c disease" 1
member "puls: fasciolopsiasis infection" "puls: c disease" 1
member "puls: angina pectoris" "puls: c non infectious disease" 1
member "puls: diabetes" "puls: c non infectious disease" 1
member "puls: mac infection" "puls: c disease" 1
member "puls: equine piroplasmosis" "puls: c disease" 1
member "puls: delusional parasitosis" "puls: c disease" 1
member "puls: acute intestinal infection" "puls: c disease" 1
member "puls: malaria" "puls: c disease" 1
```

corpus based data

```
member "wp: malaria" "wp: disease" 0.5
member "wp: malaria" "wp: milder disease" 0.5
member "wp: malaria" "wp: a disease" 0.5
member "wp: disease" "wp: fatal" 0.5
member "wp: milder disease" "wp: fatal" 0.5
member "wp: a disease" "wp: fatal" 0.5
member "wp: zoonosis" "wp: causes" 0.5
member "wp: a zoonosis" "wp: causes" 0.5
member "wp: zoonosis" "wp: infect" 0.5
member "wp: a zoonosis" "wp: infect" 0.5
member "wp: species" "wp: zoonosis" 0.5
member "wp: species" "wp: a zoonosis" 0.5
```




wp: omim malaria
wp: mesh malaria
wp: icd-9 malaria
wp: emedicine malaria
wp: ped1357 malaria
wp: medlineplus malaria
wp: diseasesdb malaria
wp: 084 malaria
wp: c03 malaria
wp: emerg305 malaria
wp: med1385 malaria
wp: malaria
puls: airport malaria
puls: cerebral malaria
puls: p parasitica
puls: c disease
puls: malaria
wp: a disease
tap: parasitic disease
tap: infectious disease
wp: malariae
wp: p malariae
wp: history of malaria
wp: malaria infection
wp: cases of malaria
tap: head lice
tap: pinworm
tap: toxoplasmosis
tap: malaria
wp: the disease
wp: vector of malaria
wp: cause of malaria
wp: severe malaria
tap: malaria
wp: malariae
puls: malaria
wp: malaria
tap: pinworm
tap: head lice
wp: p malariae
wp: a disease
wp: the disease
p: c03 malaria
p: 084 malaria


0.0
0.1
0.2
0.3
0.4
severe malaria
ause of malaria
ector of malaria
wp: the disease
tap: malaria
: toxoplasmosis
tap: pinworm
tap: head lice
arasitic disease
ectious disease
wp: malaria
asesdb malaria
p: icd-9 malaria
: mesh malaria
: omim malaria
ineplus malaria
edicine malaria
ed1357 malaria
ed1385 malaria
erg305 malaria
wp: 084 malaria
wp: c03 malaria
wp: a disease
puls: malaria
uls: p parasitica
puls: c disease
erebral malaria
: airport malaria
ases of malaria
wp: malariae
wp: p malariae
alaria infection
story of malaria

The set-up

The R statistical environment

... also:

Stanford parser (robust dependency parser)

Java/Jena for ontology I/O

GNU Prolog for graph pattern matching

GF for verbalization of results

For everything else, there is Perl

Embedding Stanford Parser in R

Work to do

- workbench is ready, but

- detection patterns
- controls: nuts and bolts
- normalization of data
- more data

- user interaction
- comprehension of output

- verbalization of results using Grammatical Framework

- analyzing Finnish input

Thank you

Seppo.Nyrkko (at) helsinki.fi