Doing Text Analytics for Digital Humanities and Social Sciences with CLARIN

Pre-conference tutorial

LDK2017 Galway

Overview of methods and approaches in Digital Humanities

Antal van den Bosch @antalvdb

Radboud University
Nijmegen

Meertens Instituut
Amsterdam

Thanks to Martha van den Hoven, Matje van de Camp, Kalliopi Zervanou, Folgert Karsdorp, Darja Fiser, Franciska de Jong

Galway, Ireland, June 18, 2017

Tutorial: Day program

09.00 - 10.30	Overview of methods and approaches in Digital Humanities (Antal van den Bosch)
10.30 - 11.00	Covfefe break
11.00 - 12.30	Overview of methods and approaches in Social Sciences (Dong Nguyen)
12.30 - 13.30	Lunch
13.30 - 15.30	Hands-on session by Folgert Karsdorp (1)
15.30 - 16.00	Covfefe break
16.00 - 17.00	Hands-on session by Folgert Karsdorp (2)
17.00 - 18.00	CLARIN Researcher Clinic

Methods and approaches in Digital Humanities

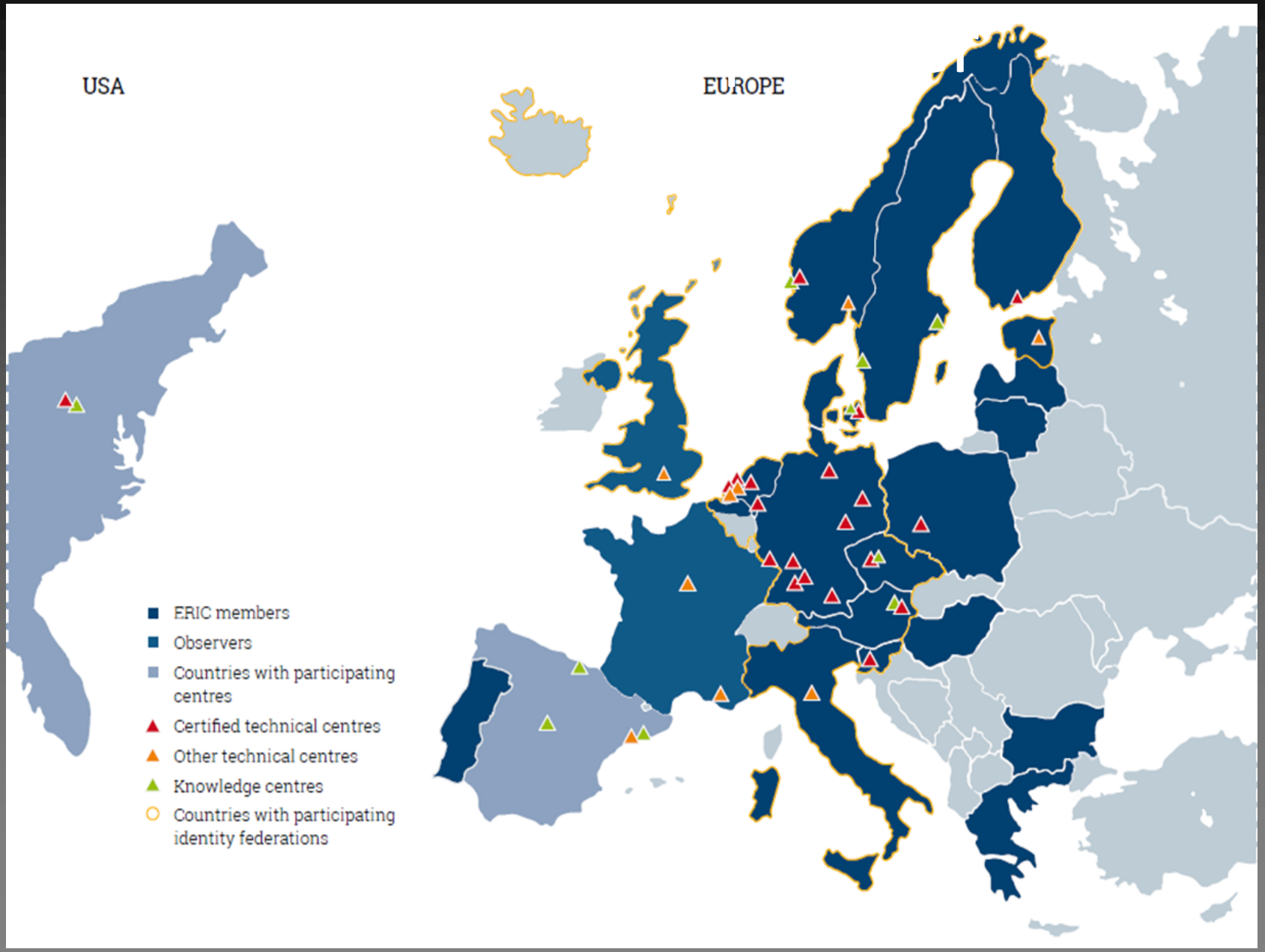
Main menu

- Context: CLARIN and Digital humanities
- The green button
- Language hides information
- Three self-portraits of the researcher (mixed techniques)

CLARIN in five bullets

- **CLARIN** is the Common Language Resources and Technology Infrastructure
- that provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
- to **digital language data** (in written, spoken, video or multimodal form),
- and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
- through a **single sign-on** online environment.

CLARIN ERIC: 19 members, 2 observers,



CLARIN in data types

- Literary texts
- Social Media data
- Historical letters
- Oral History data
- Disciplinary libraries
- Institutional archival data
- Broadcast archives
- Newspaper archives
- Parliamentary records
- …

CLARIN:

Infrastructural support

for the study and use of

language as social and cultural data

Digital humanities

or computational humanities, computing & the humanities, e-humanities, …

assumes:

- Some digitization having occurred
- Familiarizing with IT basics

does not offer:

- Big Green Button solutions
- One-fit-all procedures

Digital humanities

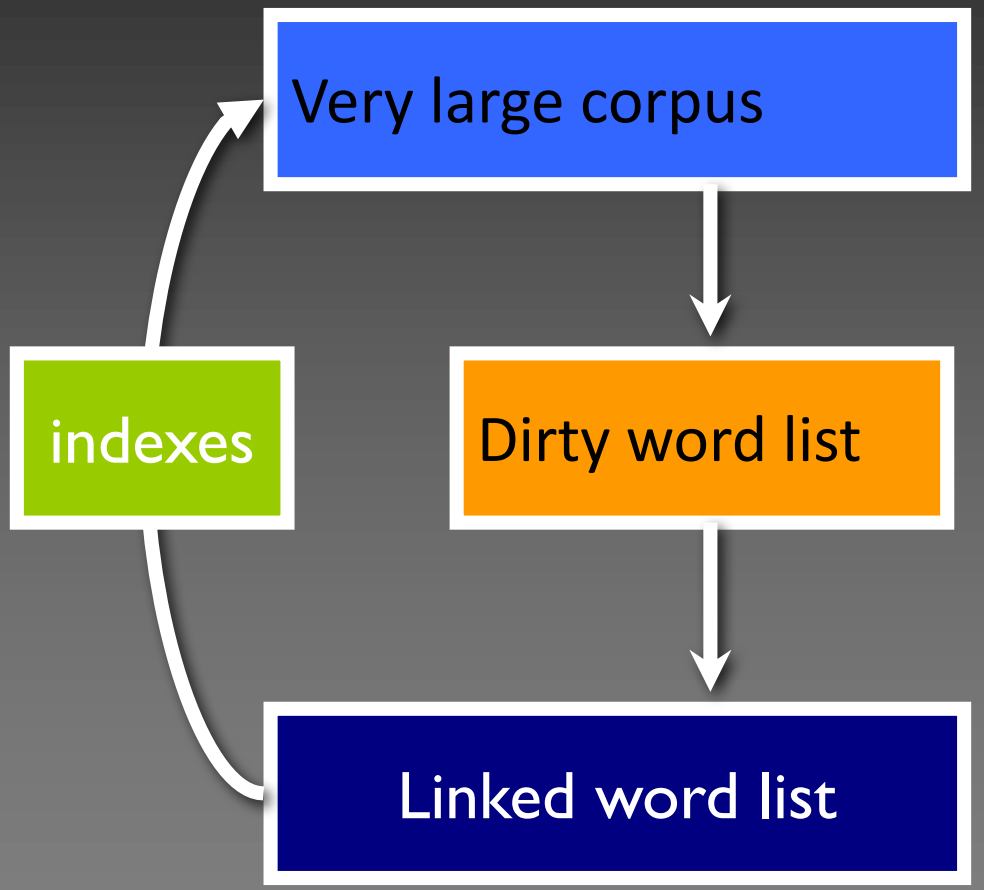
- Introducing simple models that only a computer can work out
  - e.g. random baselines
- Introducing quantitative evaluation metrics developed elsewhere
  - the ultimate ‘e-science’ idea
- Starting point of a conversation with theorists; test bed for different theories

The green button

- ”The NLP Pipeline”
  - and the text mining pipeline
- Everyone’s favorite module
  - Named Entity Recognition
- The plumbing
  - OSes, web services, servers, virtual machines
  - databases and search engines
  - annotation tools and viewers

Corpus cleanup and text normalization

- Text-induced corpus cleanup
  - Martin Reynaert
- Robust, scalable method for finding wordform variants
- Sensitive to morphology and context
- Knowledge-free



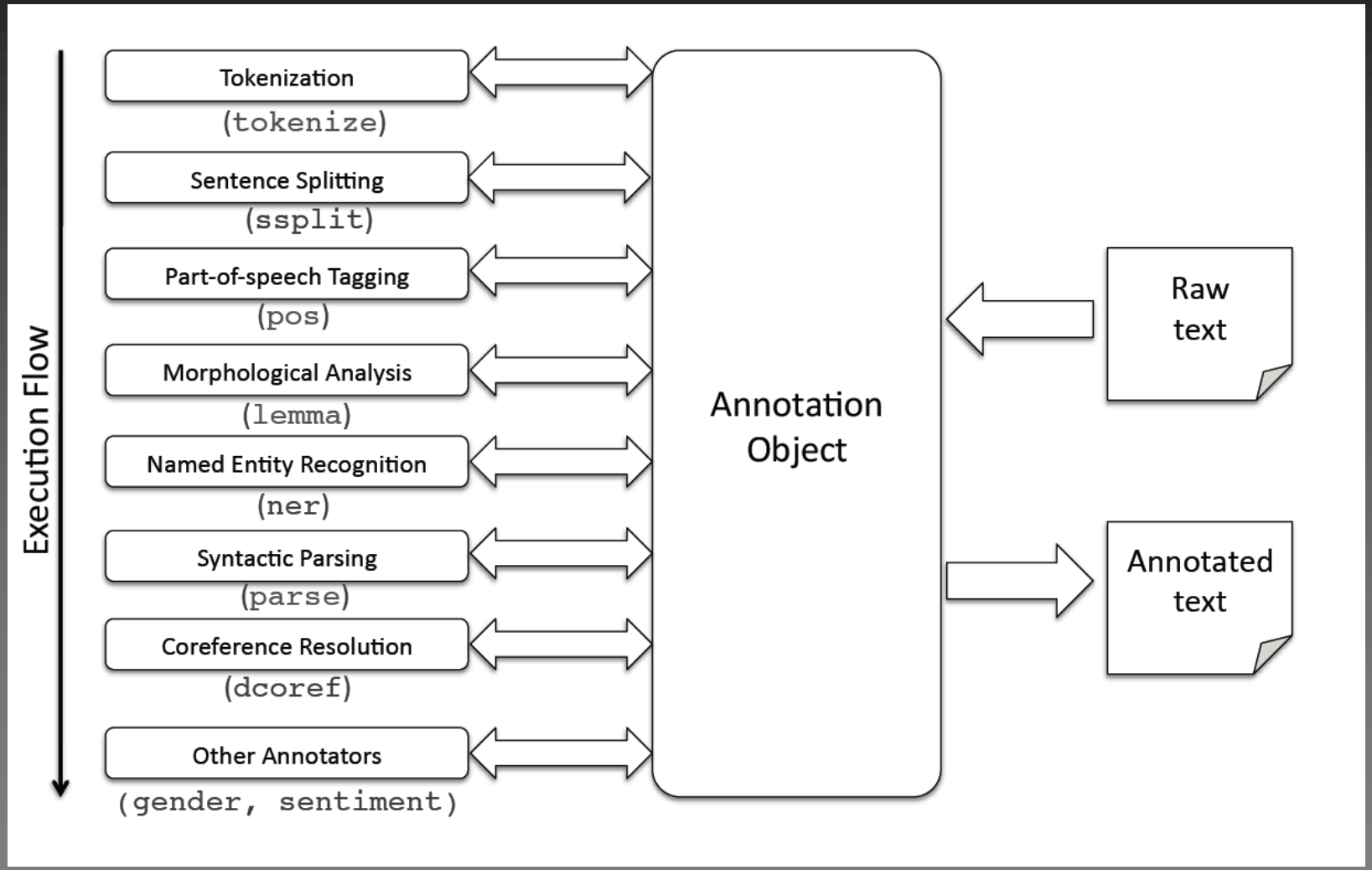
TICCL

- hartstochtelijk hartstochtelyk hartstochtelyke hartstochtlijk hartstochtlijke hartstochtlyk hartstogtelijk hartstogtelijke hartstogtelijks hartstogtelyk

- wenkbrauwen wenkbraauwen wenkbraeuwen wenkbrauwen winkbraauwen wynbraauwen wynbrauwen

- Nederland NEDERLANDEN Nederlan Nederland Nederlanden Nederlander Nederlandse Nederlandt Nederlandts Nederlandze Nederlansch Nederlanse Nederlant Nederlants Neederland Neerland Neerlands Neerlandts Neerlants Netherlands

Stanford CoreNLP pipeline (English)

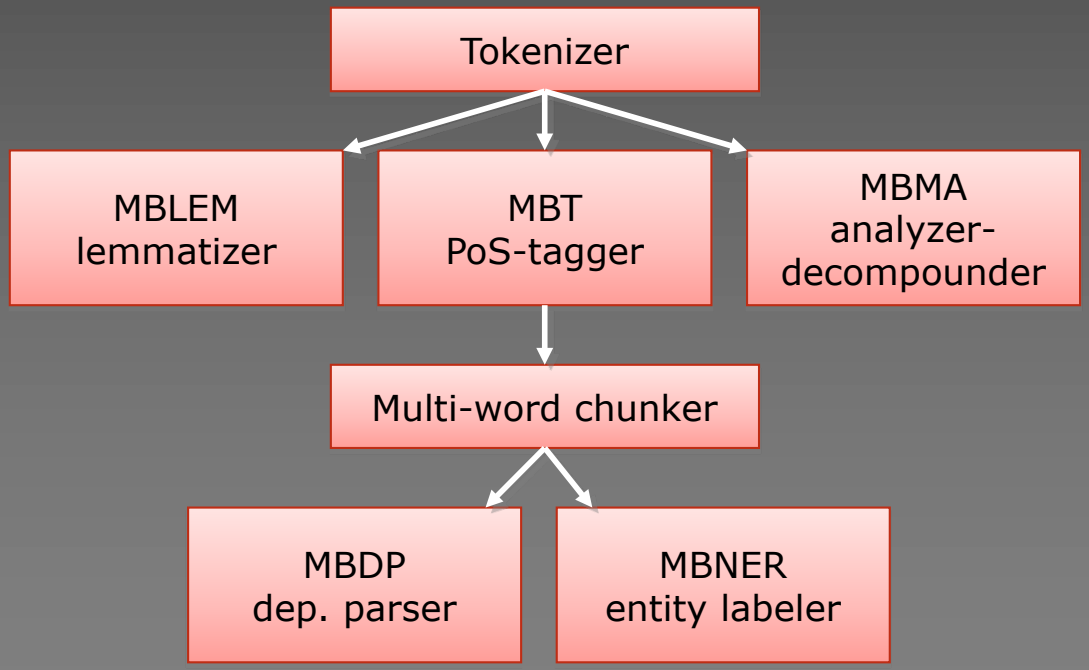


https://stanfordnlp.github.io/CoreNLP/

Frog pipeline (Dutch)

https://github.com/LanguageMachines/frog

```
1       Marie   Marie   [Marie] SPEC(deeleigen)                 1.000000        B-PER   B-NP    2       su
2       vroeg   vragen  [vraag] WW(pv,verl,ev)                  0.532544        O       B-VP    0       ROOT
3       zich    zich    [zich]  VNW(refl,pron,obl,red,3,getal)  0.999740        O       B-NP    2       se
4       af      af      [af]    VZ(fin)                         0.996853        O       O       2       svp
5       of      of      [of]    VG(onder)                       0.733333        O       B-SBAR  4       vc
6       hij     hij     [hij]   VNW(pers,pron,nomin,vol,3,ev,masc)      0.999659        O       B-NP    8       su
7       nog     nog     [nog]   BW()                            0.999930        O       B-ADVP  8       mod
8       zou     zullen  [zal]   WW(pv,verl,ev)                 0.999947        O       B-VP    5       body
9       komen   komen   [kom][en]       WW(inf,vrij,zonder)     0.861549        O       I-VP    8       vc
10      .       .       [.]     LET()                           0.999956        O       O       9       punct
```



Module	What	% correct
MBT	Known words	96.8 (98.7)
	Unknown words	76.4 (84.3)
	All words	96.5 (98.6)
MBLEM	Unknown words	73.9
MBMA	Unknown words	79.0
MBDP	Labeled relations	81.9
MBNER	Recognized entities	80.1

Other pipelines

- FreeLing http://nlp.lsi.upc.edu/freeling/
- GATE http://gate.ac.uk/
- NLTK http://nltk.sourceforge.net
- Illinois NLP Curator https://cogcomp.cs.illinois.edu/page/software_view/Curator
- Apache OpenNLP https://opennlp.apache.org/

And:

- TreeTagger http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Pipelines?

- Technologically positivistic idea that
  - modularization is good;
  - everyone needs one;
  - they will do the work for you.
- In reality,
  - modularization rarely works as intended;
  - most research questions require more than a standard pipeline
  - (if there is one for your language);
  - and often no pipeline at all.
- Pipelines tend to obfuscate the real question.

It starts with a question

- "What is the underlying research question?"
  - is really the first thing that should be on the table
- The researcher ("user") is in the lead
  - despite technology being pushed
- If in a collaboration
  - e.g. between humanities researchers and AI/IT/ML/NLP specialists
  - communicate, find common ground!

The rest of the plumbing

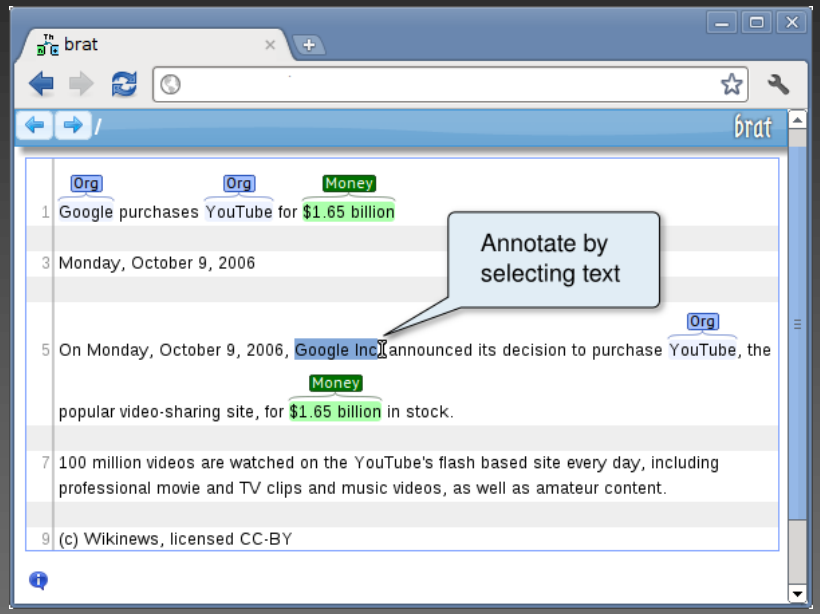
Formats (XML), Webservices, Workflows, Visualizers

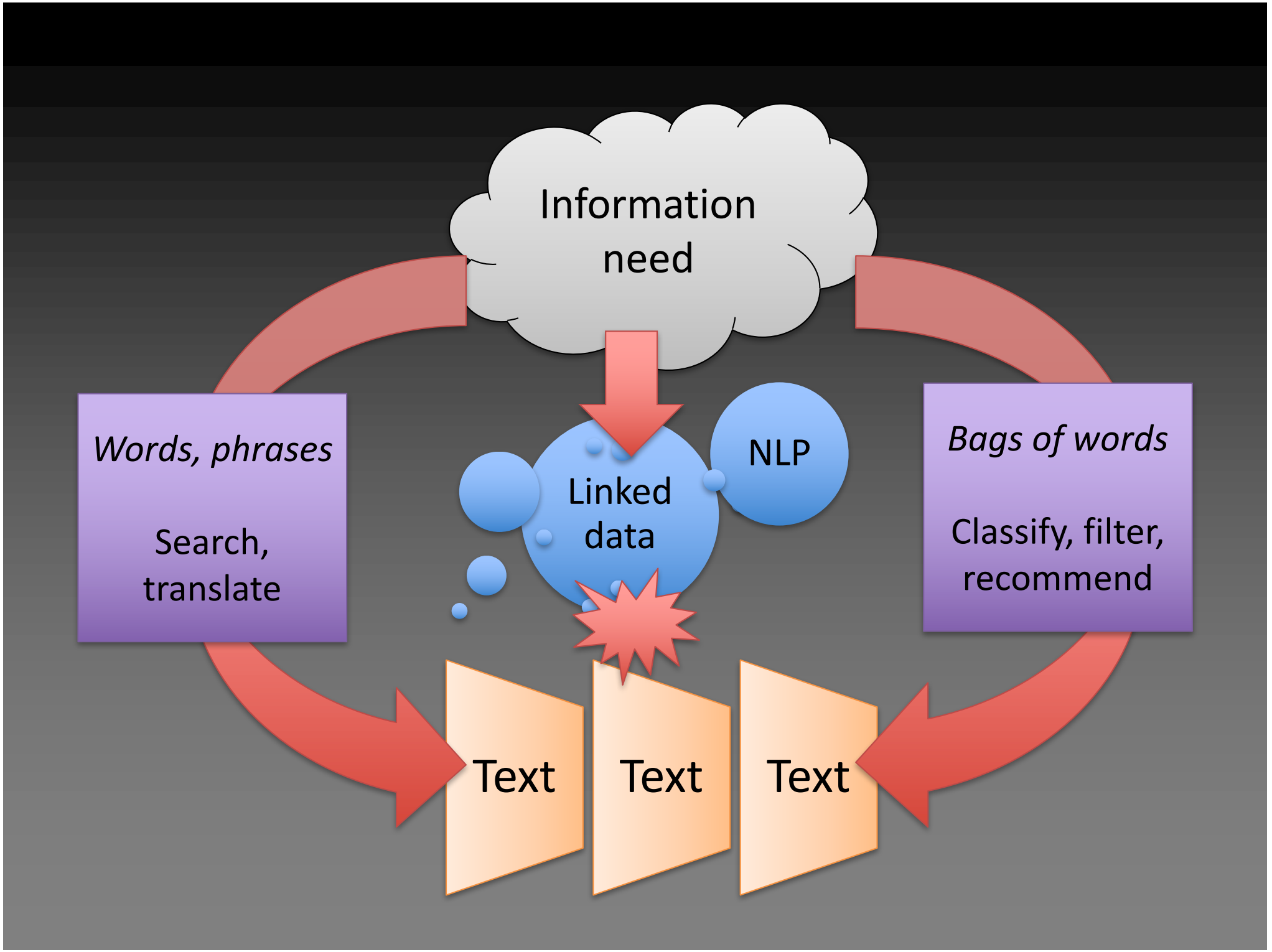
Our home-brew set of tools (demo!):

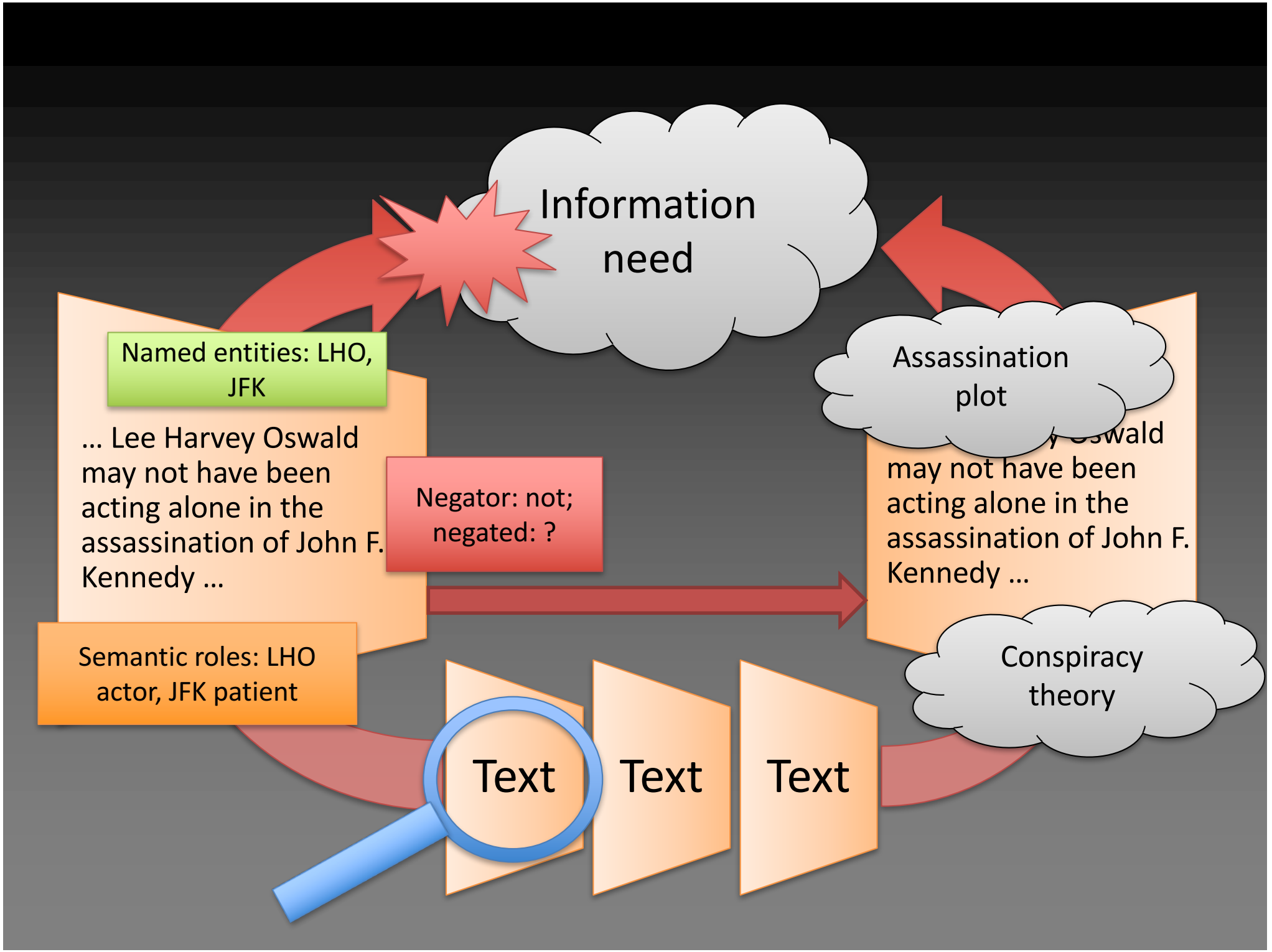
- CLAM (Computational Linguistics Applications Mediator)
  - http://proycon.github.io/clam/
- FLAT (FoLiA Linguistic Annotation Tool)
  - https://github.com/proycon/flat
- FoLiA (Format for Linguistic Annotation)
  - https://proycon.github.io/folia/

Other annotation tools

- BRAT
  - http://brat.nlplab.org/
- WebAnno (CLARIN-D)
  - https://webanno.github.io/webanno/
- Annis
  - http://corpus-tools.org/annis/
  - http://corpus-tools.org/home/




Information need
Words, phrases
Search, translate
Linked data
NLP
Bags of words
Classify, filter, recommend
Text
Text
Text


Information need
Named entities: LHO, JFK
… Lee Harvey Oswald may not have been acting alone in the assassination of John F. Kennedy …
Negator: not; negated: ?
Semantic roles: LHO actor, JFK patient
Assassination plot
may not have been acting alone in the assassination of John F. Kennedy …
Conspiracy theory
Text
Text
Text

Information vs. language

- Every little fact can be (and is) expressed in an endless variety of linguistic expressions
- Language is a great way of hiding information

Navajo Code Talkers Henry Bake and George Kirk, 12/1943 (ARC 593415)

マヨネーズ [編集]

従来、古くからある関西風お好み焼き店の多くはマヨネーズをかけたり、つけたりすることはなかった。広島や神戸では、現在でもマヨネーズの使用は少なく、同じ関西でも大阪と神戸ではマヨネーズに対する嗜好に違いがある。現在の大阪では多くの店でマヨネーズがかけられているのに対し、神戸ではより伝統的なお好み焼きにこだわり、マヨネーズを置かない店も少なからず存在し、また置いていても注文しないと出てこないことも多い。どこでいつから関西風お好み焼きにマヨネーズをつけるようになったかは諸説あり定かではなく、ぼてぢゅうの説、個人店が最初という説、またそれ以前から家庭内で使われていたと言う説もある。関西風お好み焼きを供する多くの店ではマヨネーズが使用される。また、店によっては溶き芥子を少量加えることもある。

モダン焼き [編集]

モダン焼き（「そばのせ」とも言う）は、関西風お好み焼きの一種で、具材に焼きそば用の茹でた（あるいは蒸した）中華麺を生地に混ぜ、または通常のお好み焼きに重ね、焼いたもの。一枚でお好み焼きと焼きそばを同時に賞味できるという関西的な合理的発想が根底にあり、根強い人気がある。中華麺の代わりにうどんを用いる場合もあり、「うどんモダン」と呼ばれる。また、店によっては、重ねるお好み焼きの生地に卵を加えない場合もある。ボリューム感あふれる外見とそれに違わない食感が特徴である。通常、具材としてはオプションとして用意されている。神戸・明石周辺では、焼きそばを生地とのつなぎとして固めたものが「モダン焼き」と言われている。薄く生地を焼き、その上に焼きそばを乗せ、その上から生地をかけてひっくり返して焼くものである。見た目は広島風お好み焼きに似ており、発祥は、1950年（昭和25年）に『志ば多』（神戸市）で考案されたという説が有力である。当初はそばではなく、うどんを使っていた。入れる具材によってバリエーションも少なからずあるが、玉子を上面にのせ焼いたものを特に「月見モダン」と称す。モダン焼きや広島風お好み焼きに似ているものとして、「にくてん」もある。こちらは大正時代にはあったと言われている。

神戸風モダン焼き

関西での文化 [編集]

関西風お好み焼き屋の業態として、オーダーごとに生の具材と生地を客に提供し、客が自分で調理し焼き上げる半セルフサービスの店がある。店側としては食材を用意するだけで良く省力化ともなるので、チェーン店などでこの方法をとる店も多く、関東一円でもこの形式の店は顕著に見られる。ホットプレートなどの普及で、お好み焼きが家庭でも広く一般化し、高度な調理技術を要求されないこともあり、店側の焼き方にとらわれず自由に焼き具合や調味加減ができる面白さも手伝って、カップルや学生、団体客などの需要に受けている。

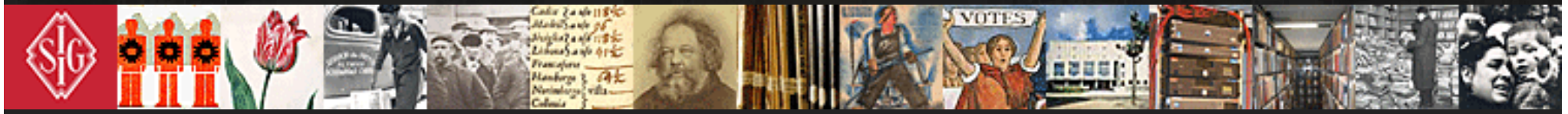
お好み焼きを米飯のおかずとする人が多いのが大阪の特徴[2]。また、関西のお好み焼き屋、定食屋には米飯を添える「お好み焼き定食」を出す店舗が存在する。

近年、関東でも関西風お好み焼き店が増えており、その客も関西出身に限らず関東や日本各地の出身者も多い。しかしその食べ方は、関西とそれ以外の出身の人たちでは大きく異なる。関西ではお好み焼きはテコでさいの目状に

Mixed techniques

- Rather than just using a pipeline,
  - which may offer some useful preprocessing,
- Most researchers resort to using a bag of methods
  - existing open source modules
  - scripting (Python, Perl, bash/sed/awk)
- requiring familiarizing themselves with the basics of programming
  - and Linux, github, …
- cf. Folgert Karsdorp's hands-on course this afternoon

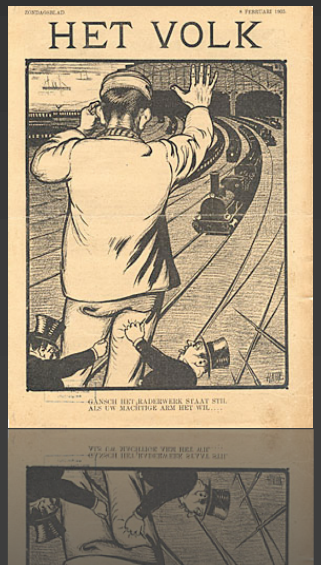
HiTiME

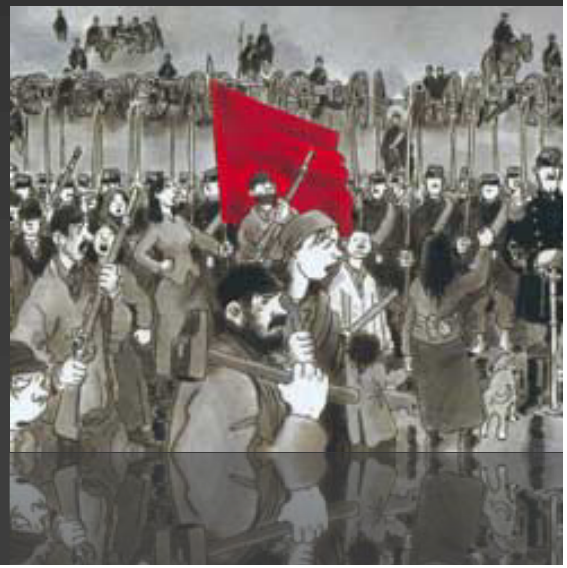


Historical Timeline Mining and Extraction

HiTiME aims

- History of the social movement in Europe, 1850-1940
- Automatic analysis of primary and secondary text sources
- Detecting persons, occupations, organisations, locations, time expressions, events
- Linking all of this information
- Visualising the links through timelines





Example project: Biographies

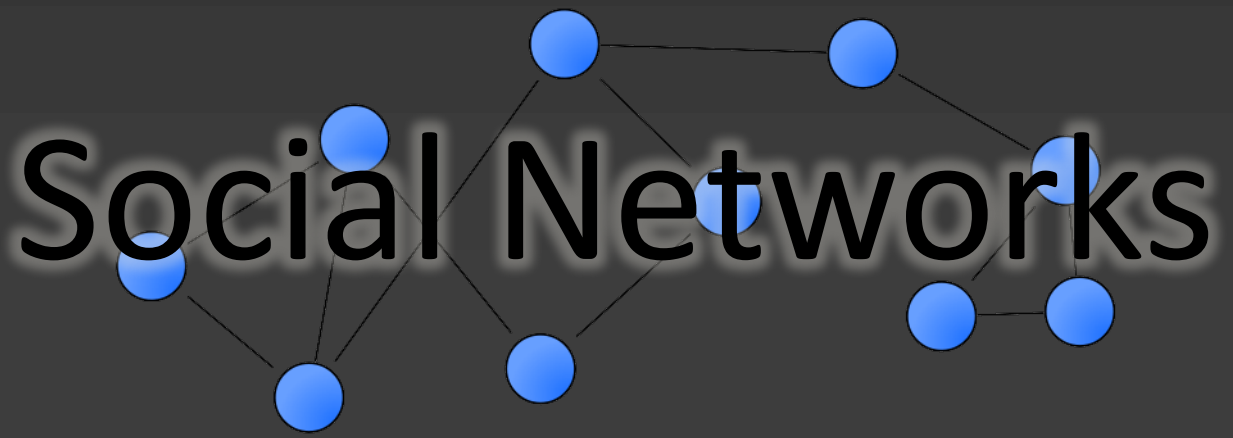
- BWSA: Biographical Dictionary of Socialism in the Netherlands
  - 575 biographies of politicians, activists, artists, writers, etc.
  - Improved search
  - Linking entities across sources
  - Facilitate discovery of new information

- Camp, M. van de (2015). *A link to the past: Constructing historical social networks from unstructured data*. Ph.D. thesis, Tilburg University.

BWSA Biografisch Woordenboek van het Socialisme en de Arbeidersbeweging in Nederland

Data

- BWSA: Biographical Dictionary of Socialism in the Netherlands
  - 575 biographies of politicians, activists, artists, writers, etc.
  - Improved search
  - Linking entities across sources
  - Facilitate discovery of new information



BWSA

BWSA

Biografisch Woordenboek van het Socialisme en de Arbeidersbeweging in Nederland

Ga naar:
Home BWSA
Alfabetische index
Zoeken
©

Email

NIEUWENHUIS, Ferdinand

bekend als Ferdinand Domela Nieuwenhuis, pionier van het socialisme, stichter van het blad *Recht voor Allen*, later sociaal-anarchist, is geboren te Amsterdam op 31 december 1846 en overleden te Hilversum op 18 november 1919. Hij was de zoon van Ferdinand Jacobus Domela Nieuwenhuis, luthers predikant en hoogleraar theologie, en Henriette Frances Berry. Op 24 maart 1870 trad hij in het huwelijk met Johanna Lulofs, met wie hij twee zoons kreeg. Na haar overlijden op 26 maart 1872 hertrouwde hij op 29 oktober 1874 met Johanna Adriana Verhagen, met wie hij twee dochters kreeg. Na haar overlijden op 1 augustus 1877 hertrouwde hij op 21 april 1880 met Johanna Frederika Schingen Hagen, met wie hij een zoon kreeg. Na haar overlijden op 27 februari 1884 hertrouwde hij op 27 mei 1891 met Johanna Egberta Godthelp, met wie hij een dochter en twee zoons kreeg. Bij Koninklijk Besluit van 10 juli 1859 werd de naam door toevoeging gewijzigd in Domela Nieuwenhuis.
Pseudoniemen: Criticus, Ex-Theoloog, Germanus, Philalethes, Dr. Sagittarius.

BWSA

BWSA

Biografisch Woordenboek van het Socialisme en de Arbeidersbeweging in Nederland

Ga naar:
Home BWSA
Alfabetische index
Zoeken
©

Email

NIEUWENHUIS, Ferdinand

bekend als Ferdinand Domela Nieuwenhuis, pionier van het socialisme, stichter van het blad *Recht voor Allen*, later sociaal-anarchist, is geboren te Amsterdam op 31 december 1846 en overleden te Hilversum op 18 november 1919. Hij was de zoon van Ferdinand Jacobus Domela Nieuwenhuis, luthers predikant en hoogleraar theologie, en Henriette Frances Berry. Op 24 maart 1870 trad hij in het huwelijk met Johanna Lulofs, met wie hij twee zoons kreeg. Na haar overlijden op 26 maart 1872 hertrouwde hij op 29 oktober 1874 met Johanna Adriana Verhagen, met wie hij twee dochters kreeg. Na haar overlijden op 1 augustus 1877 hertrouwde hij op 21 april 1880 met Johanna Frederika Schingen Hagen, met wie hij een zoon kreeg. Na haar overlijden op 27 februari 1884 hertrouwde hij op 27 mei 1891 met Johanna Egberta Godthelp, met wie hij een dochter en twee zoons kreeg. Bij Koninklijk Besluit van 10 juli 1859 werd de naam door toevoeging gewijzigd in Domela Nieuwenhuis.
Pseudoniemen: Criticus, Ex-Theoloog, Germanus, Philalethes, Dr. Sagittarius.

First name	Ferdinand
Last name	Nieuwenhuis
Year of birth	1846
Date of birth	31-12
Year of death	1919
Date of death	18-11
Extra info	(bekend als Ferdinand Domela Nieuwenhuis) pionier van het socialisme, stichter van het blad Recht voor Allen, later sociaal-anarchist

BWSA

BWSA

Biografisch Woordenboek van het Socialisme en de Arbeidersbeweging in Nederland

Ga naar:
Home BWSA
Alfabetische index
Zoeken
©

Email

NIEUWENHUIS, Ferdinand

bekend als Ferdinand Domela Nieuwenhuis, pionier van het socialisme, stichter van het blad *Recht voor Allen*, later sociaal-anarchist, is geboren te Amsterdam op 31 december 1846 en overleden te Hilversum op 18 november 1919. Hij was de zoon van Ferdinand Jacobus Domela Nieuwenhuis, luthers predikant en hoogleraar theologie, en Henriette Frances Berry. Op 24 maart 1870 trad hij in het huwelijk met Johanna Lulofs, met wie hij twee zoons kreeg. Na haar overlijden op 26 maart 1872 hertrouwde hij op 29 oktober 1874 met Johanna Adriana Verhagen, met wie hij twee dochters kreeg. Na haar overlijden op 1 augustus 1877 hertrouwde hij op 21 april 1880 met Johanna Frederika Schingen Hagen, met wie hij een zoon kreeg. Na haar overlijden op 27 februari 1884 hertrouwde hij op 27 mei 1891 met Johanna Egberta Godthelp, met wie hij een dochter en twee zoons kreeg. Bij Koninklijk Besluit van 10 juli 1859 werd de naam door toevoeging gewijzigd in Domela Nieuwenhuis.
Pseudoniemen: Criticus, Ex-Theoloog, Germanus, Philalethes, Dr. Sagittarius.

First name	Ferdinand
Last name	Nieuwenhuis
Year of birth	1846
Date of birth	31-12
Year of death	1919
Date of death	18-11
Extra info	(bekend als Ferdinand Domela Nieuwenhuis) pionier van het socialisme, stichter van het blad Recht voor Allen, later sociaal-anarchist

BWSA

BWSA

Biografisch Woordenboek van het Socialisme en de Arbeidersbeweging in Nederland

Ga naar:
Home BWSA
Alfabetische index
Zoeken
©

Email

NIEUWENHUIS, Ferdinand

bekend als Ferdinand Domela Nieuwenhuis, pionier van het socialisme, stichter van het blad *Recht voor Allen*, later sociaal-anarchist, is geboren te Amsterdam op 31 december 1846 en overleden te Hilversum op 18 november 1919. Hij was de zoon van Ferdinand Jacobus Domela Nieuwenhuis, luthers predikant en hoogleraar theologie, en Henriette Frances Berry. Op 24 maart 1870 trad hij in het huwelijk met Johanna Lulofs, met wie hij twee zoons kreeg. Na haar overlijden op 26 maart 1872 hertrouwde hij op 29 oktober 1874 met Johanna Adriana Verhagen, met wie hij twee dochters kreeg. Na haar overlijden op 1 augustus 1877 hertrouwde hij op 21 april 1880 met Johanna Frederika Schingen Hagen, met wie hij een zoon kreeg. Na haar overlijden op 27 februari 1884 hertrouwde hij op 27 mei 1891 met Johanna Egberta Godthelp, met wie hij een dochter en twee zoons kreeg. Bij Koninklijk Besluit van 10 juli 1859 werd de naam door toevoeging gewijzigd in Domela Nieuwenhuis.

Pseudoniemen: Criticus, Ex-Theoloog, Germanus, Philalethes, Dr. Sagittarius.

First name	Ferdinand
Last name	Nieuwenhuis
Year of birth	1846
Date of birth	31-12
Year of death	1919
Date of death	18-11
Extra info	(bekend als Ferdinand Domela Nieuwenhuis) pionier van het socialisme, stichter van het blad Recht voor Allen, later sociaal-anarchist

BWSA

BWSA

Biografisch Woordenboek van het Socialisme en de Arbeidersbeweging in Nederland

Ga naar:
Home BWSA
Alfabetische index
Zoeken
©

Email

NIEUWENHUIS, Ferdinand

bekend als Ferdinand Domela Nieuwenhuis, pionier van het socialisme, stichter van het blad *Recht voor Allen*, later sociaal-anarchist, is geboren te Amsterdam op 31 december 1846 en overleden te Hilversum op 18 november 1919. Hij was de zoon van Ferdinand Jacobus Domela Nieuwenhuis, luthers predikant en hoogleraar theologie, en Henriette Frances Berry. Op 24 maart 1870 trad hij in het huwelijk met Johanna Lulofs, met wie hij twee zoons kreeg. Na haar overlijden op 26 maart 1872 hertrouwde hij op 29 oktober 1874 met Johanna Adriana Verhagen, met wie hij twee dochters kreeg. Na haar overlijden op 1 augustus 1877 hertrouwde hij op 21 april 1880 met Johanna Frederika Schingen Hagen, met wie hij een zoon kreeg. Na haar overlijden op 27 februari 1884 hertrouwde hij op 27 mei 1891 met Johanna Egberta Godthelp, met wie hij een dochter en twee zoons kreeg. Bij Koninklijk Besluit van 10 juli 1859 werd de naam door toevoeging gewijzigd in Domela Nieuwenhuis.

Pseudoniemen: Criticus, Ex-Theoloog, Germanus, Philalethes, Dr. Sagittarius.

First name	Ferdinand
Last name	Nieuwenhuis
Year of birth	1846
Date of birth	31-12
Year of death	1919
Date of death	18-11
Extra info	(bekend als Ferdinand Domela Nieuwenhuis) pionier van het socialisme, stichter van het blad Recht voor Allen, later sociaal-anarchist

BWSA

BWSA

Biografisch Woordenbo
en de Arbeidersbewegi

Ga naar:
Home BWSA
Alfabetische index
Zoeken
©

Email

NIEUWENHUIS, Ferdinand

bekend als Ferdinand Domela Nieuwer
pionier van het socialisme, stichter va
blad *Recht voor Allen*, later sociaal-an
is geboren te Amsterdam op 31 decem
1846 en overleden te Hilversum op 18
november 1919. Hij was de zoon van
Ferdinand Jacobus Domela Nieuwenh
luthers predikant en hoogleraar theolo
Henriette Frances Berry. Op 24 maart 1
hij in het huwelijk met Johanna Lulofs,
hij twee zoons kreeg. Na haar overlijde
maart 1872 hertrouwde hij op 29 oktob
met Johanna Adriana Verhagen, met w
twee dochters kreeg. Na haar overlijde
augustus 1877 hertrouwde hij op 21 ap
met Johanna Frederika Schingen Hage
wie hij een zoon kreeg. Na haar overlij
27 februari 1884 hertrouwde hij op 27 m
met Johanna Egberta Godthelp, met w
Koninklijk Besluit van 10 juli 1859 werd
Domela Nieuwenhuis.
Pseudoniemen: Criticus, Ex-Theoloog,

First name	Ferdinand
Last name	Nieuwenhuis
Year of birth	1846
Date of birth	31-12
Place of birth	Amsterdam
Year of death	1919
Date of death	18-11
Place of death	Hilversum
Extra info	(bekend als Ferdinand Domela Nieuwenhuis) pionier van het socialisme, stichter van het blad Recht voor Allen, later sociaal-anarchist
Son of	Ferdinand Jacobus Domela Nieuwenhuis, Henriette Frances Berry
Married to	Johanna Lulofs, Johanna Adriana Verhagen, Johanna Frederika Schingen Hagen, Johanna Egberta Godthelp
Pseudonyms	Criticus, Ex-Theoloog, Germanus, Philalethes, Dr. Sagittarius

Tools

- Named entity recognition
  - People, organizations, locations, …
- Named entity disambiguation
  - SDAP = Sociaal-Democratische Arbeiderspartij

In 1909 werd hij voor het eerst naar de Tweede Kamer afgevaardigd door het district Amsterdam IX. Eerder was hij reeds gekozen in de gemeenteraad van Amsterdam en in de Provinciale Staten van Noord-Holland. De 'parlementair van natuur' die hij volgens Domela Nieuwenhuis was, bepleitte in 1913 in zijn artikel 'Aanpakken' de aanvaarding van drie ministerportefeuilles die formateur dr. D. Bos aan de SDAP aanbood. In het oorlogsjaar 1914 werd hij naast Wibaut tot wethouder van Amsterdam gekozen, hetgeen hem noopte zijn taak aan *Het Volk* neer te leggen.

(biography: W.H. Vliegen)

Tools

- Relations based on (co-)occurrence:
  - PERSON to PERSON
  - PERSON to ORGANIZATION
  - PERSON to LOCATION

In 1909 werd hij voor het eerst naar de Tweede Kamer afgevaardigd door het district Amsterdam IX. Eerder was hij reeds gekozen in de gemeenteraad van Amsterdam en in de Provinciale Staten van Noord-Holland. De 'parlementair van natuur' die hij volgens Domela Nieuwenhuis was, bepleitte in 1913 in zijn artikel 'Aanpakken' de aanvaarding van drie ministerportefeuilles die formateur dr. D. Bos aan de SDAP aanbood. In het oorlogsjaar 1914 werd hij naast Wibaut tot wethouder van Amsterdam gekozen, hetgeen hem noopte zijn taak aan *Het Volk* neer te leggen.

(biography: W.H. Vliegen)

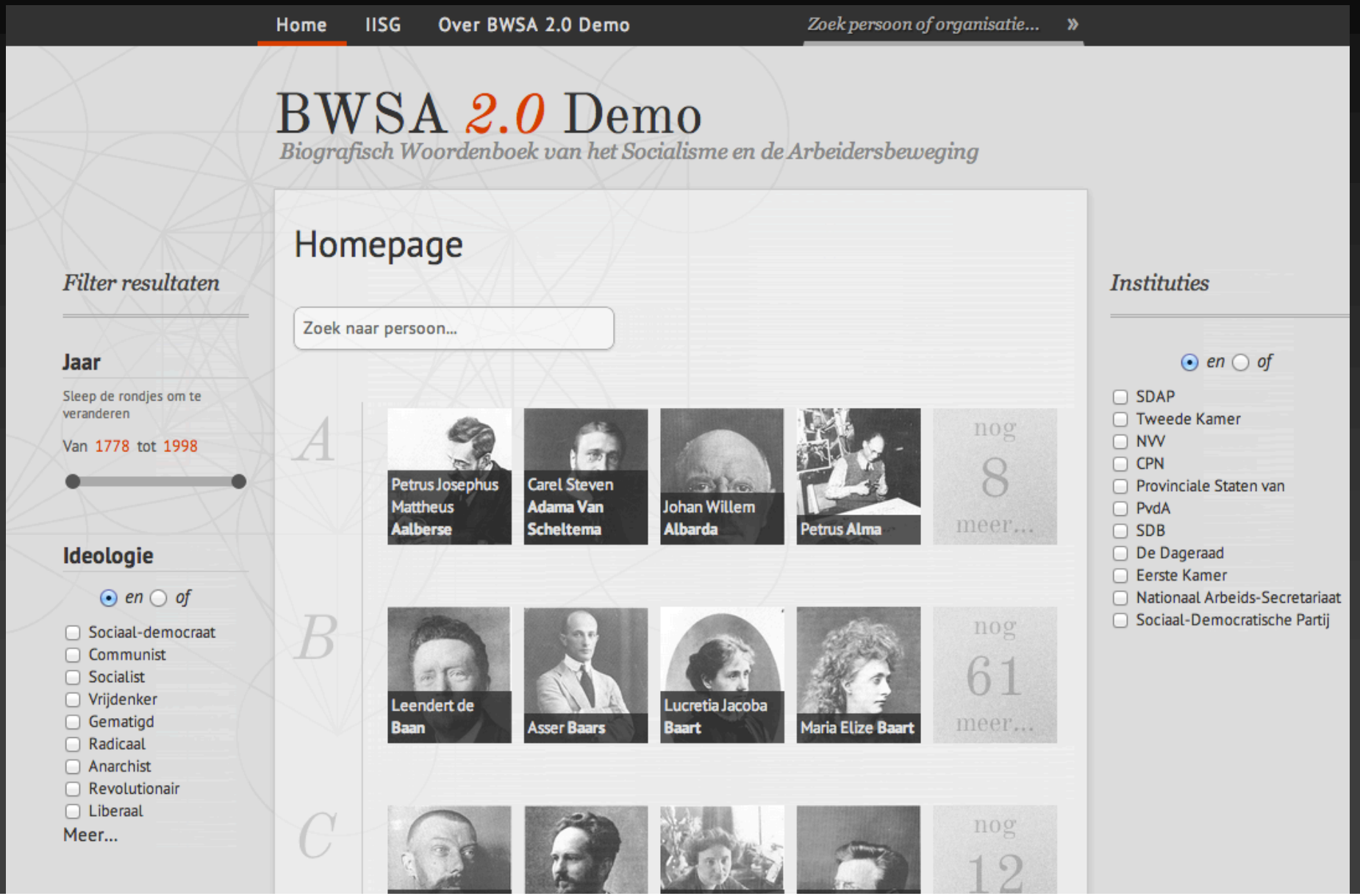
Tools

- Rule-based temporal tagging:
  - Date normalization
  - Event classification
  - Event-timex linking

In 1909 werd hij voor het eerst naar de Tweede Kamer afgevaardigd door het district Amsterdam IX. Eerder was hij reeds gekozen in de gemeenteraad van Amsterdam en in de Provinciale Staten van Noord-Holland. De 'parlementair van natuur' die hij volgens Domela Nieuwenhuis was, bepleitte in 1913 in zijn artikel 'Aanpakken' de aanvaarding van drie ministerportefeuilles die formateur dr. D. Bos aan de SDAP aanbood. In het oorlogsjaar 1914 werd hij naast Wibaut tot wethouder van Amsterdam gekozen, hetgeen hem noopte zijn taak aan *Het Volk* neer te leggen.

(biography: W.H. Vliegen)

The Socialist Network

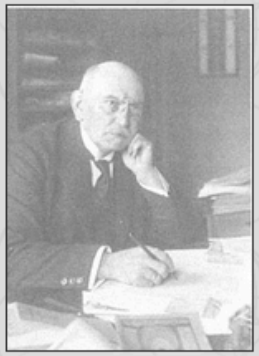


The Socialist Network

Home IISG Over BWSA 2.0 Demo *Zoek persoon of organisatie...* »

BWSA *2.0* Demo

Biografisch Woordenboek van het Socialisme en de Arbeidersbeweging



Connecties

- Franc van der Goes
- F. Domela Nieuwenhuis
- G.H. Pieters
- F.M. Wibaut
- dr. D. Bos
- J.H.A. Schaper
- W. Drees
- J.W. Albarda

Wilhelmus Hubertus VLIEGEN

Biografie | Extra informatie | Bronnen

(roepnaam: Willem), voorman van de Nederlandse socialistische beweging sinds de jaren tachtig van de vorige eeuw tot aan de Tweede Wereldoorlog en een van de twaalf oprichters van de SDAP, is geboren te Gulpen op 20 november 1862 en overleden te Bloemendaal op 29 juni 1947. Hij was de zoon van Jan Martinus Hubertus Vliegen, schrijnwerker, en Helena Jacquemin. Op 23 mei 1888 trad hij in het huwelijk met Maria Margaretha Hofman, met wie hij twee dochters en twee zoons kreeg.
Pseudoniem: Caesar T.

Het was de toenmalige Tweede Kamervoorzitter P.J.M. Aalberse , die in 1937 de vijfenzeventigjarige nestor Vliegen uitluidde met de woorden dat deze lange jaren in 's lands vergaderzaal met zijn gaven had gewoekerd. In feite bestond de academische opleiding van de Limburger slechts uit enkele jaren lagere school, aangevuld met de ervaring als leerling-typograaf in zijn geboorteplaats. Hij verwierf zich een grondige kennis van het Duits bij zijn zetarbeid, weldra ook van het Frans, toen hij op negentienjarige leeftijd naar Luik ging. Te Amsterdam trad hij in 1883 toe tot de afdeling van de Sociaal-Democratische Bond (SDB) na het horen van een rede van de Belg Edward Anseele. Het jaar daarop ging hij naar Limburg terug, waar hij te Maastricht een afdeling van de Bond voor Algemeen Kies- en Stemrecht oprichtte. Aldra schreef hij ook bijdragen in landelijke organen als Recht voor Allen. Als typograaf ontslagen moest hij naar elders vertrekken. Te Den Haag ging hij een rol spelen als leidende figuur van de SDB (lid van de Centrale Raad). De sociaal-liberaal A. Kerdijk merkte zijn parlementaire welsprekendheid op bij een meeting tegen de arbeidswet van 1889. Vliegen

Geboren
20 november 1862
Gulpen
Overleden
29 juni 1947
Bloemendaal

1883
Te Amsterdam trad hij in 1883 toe tot de afdeling van de Sociaal-Democratische

1888
Op 23 mei 1888 trad hij in het huwelijk met Maria Margaretha Hofman, met wie

1889
De sociaal-liberaal A. Kerdijk merkte zijn parlementaire welsprekendheid op bij

1890
Van 1890 af verscheen te Maastricht De Volkstribuun, waarvan hij schrijver, zetter,

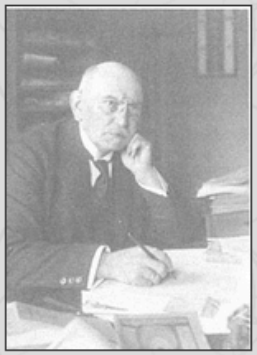
1893
Toen de SDB vanaf 1893 in anti-parlementair vaarwater raakte, schroomde Vliegen aanvankelijk nog om tegenover de door hem vereerde Domela

The Socialist Network

Home IISG Over BWSA 2.0 Demo *Zoek persoon of organisatie...* »

BWSA *2.0* Demo

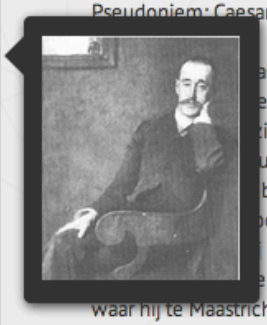
Biografisch Woordenboek van het Socialisme en de Arbeidersbeweging



Wilhelmus Hubertus VLIEGEN

Biografie | Extra informatie | Bronnen

(roepnaam: Willem), voorman van de Nederlandse socialistische beweging sinds de jaren tachtig van de vorige eeuw tot aan de Tweede Wereldoorlog en een van de twaalf oprichters van de SDAP, is geboren te Gulpen op 20 november 1862 en overleden te Bloemendaal op 29 juni 1947. Hij was de zoon van Jan Martinus Hubertus Vliegen, schrijnwerker, en Helena Jacquemin. Op 23 mei 1888 trad hij in het huwelijk met Maria Margaretha Hofman, met wie hij twee dochters en twee zoons kreeg.
Pseudoniem: Caesar T.



…alige Tweede Kamervoorzitter P.J.M. Aalberse, die in 1937 de …e nestor Vliegen uitluidde met de woorden dat deze lange jaren in 's lands …ijn gaven had gewoekerd. In feite bestond de academische opleiding van de …uit enkele jaren lagere school, aangevuld met de ervaring als leerling-…boorteplaats. Hij verwierf zich een grondige kennis van het Duits bij zijn …ook van het Frans, toen hij op negentienjarige leeftijd naar Luik ging. Te … in 1883 toe tot de afdeling van de Sociaal-Democratische Bond (SDB) na het …e van de Belg Edward Anseele. Het jaar daarop ging hij naar Limburg terug, waar hij te Maastricht een afdeling van de Bond voor Algemeen Kies- en Stemrecht oprichtte. Aldra schreef hij ook bijdragen in landelijke organen als Recht voor Allen. Als typograaf ontslagen moest hij naar elders vertrekken. Te Den Haag ging hij een rol spelen als leidende figuur van de SDB (lid van de Centrale Raad). De sociaal-liberaal A. Kerdijk merkte zijn parlementaire welsprekendheid op bij een meeting tegen de arbeidswet van 1889. Vliegen …

Connecties

- Franc van der Goes
- F. Domela Nieuwenhuis
- G.H. Pieters
- F.M. Wibaut
- dr. D. Bos
- J.H.A. Schaper
- W. Drees
- J.W. Albarda

Geboren
20 november 1862
Gulpen
Overleden
29 juni 1947
Bloemendaal

1883
Te Amsterdam trad hij in 1883 toe tot de afdeling van de Sociaal-Democratische

1888
Op 23 mei 1888 trad hij in het huwelijk met Maria Margaretha Hofman, met wie

1889
De sociaal-liberaal A. Kerdijk merkte zijn parlementaire welsprekendheid op bij

1890
Van 1890 af verscheen te Maastricht De Volkstribuun, waarvan hij schrijver, zetter,

1893
Toen de SDB vanaf 1893 in anti-parlementair vaarwater raakte, schroomde

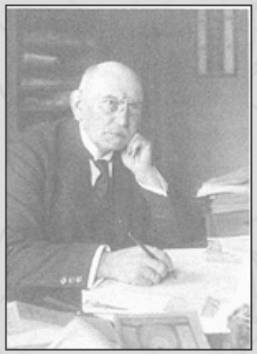
1894

The Socialist Network

Home IISG Over BWSA 2.0 Demo Zoek persoon of organisatie... »

BWSA 2.0 Demo

Biografisch Woordenboek van het Socialisme en de Arbeidersbeweging



Wilhelmus Hubertus VLIEGEN

Biografie | Extra informatie | Bronnen

Informatie

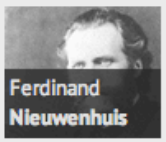
Ideologieën: Sociaal-democraat, Revolutionair

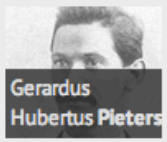
Vakgebied: Politiek, Schrijver

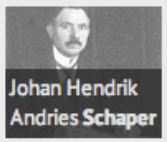
Kringen

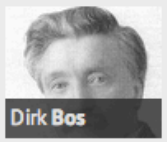
SDB en Sociaal-Democratische Bond


Franc van der Goes


Ferdinand Nieuwenhuis


Gerardus Hubertus Pieters


Johan Hendrik Andries Schaper


Dirk Bos

Connecties

- Franc van der Goes
- F. Domela Nieuwenhuis
- G.H. Pieters
- F.M. Wibaut
- dr. D. Bos
- J.H.A. Schaper
- W. Drees
- J.W. Albarda

Geboren

20 november 1862

Gulpen

Overleden

29 juni 1947

Bloemendaal

1883

Te Amsterdam trad hij in 1883 toe tot de afdeling van de Sociaal-Democratische

1888

Op 23 mei 1888 trad hij in het huwelijk met Maria Margaretha Hofman, met wie

1889

De sociaal-liberaal A. Kerdijk merkte zijn parlementaire welsprekendheid op bij

1890

Van 1890 af verscheen te Maastricht De Volkstribuun, waarvan hij schrijver, zetter,

1893

Toen de SDB vanaf 1893 in anti-parlementair vaarwater raakte, schroomde

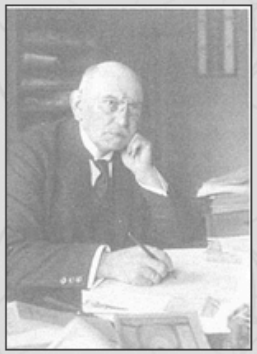
1894

The Socialist Network

Home IISG Over BWSA 2.0 Demo Zoek persoon of organisatie... »

BWSA *2.0* Demo

Biografisch Woordenboek van het Socialisme en de Arbeidersbeweging



Wilhelmus Hubertus VLIEGEN

Biografie | Extra informatie | Bronnen

Genoemde plaatsen

Geboren

20 november 1862

Gulpen

Overleden

29 juni 1947

Bloemendaal

1883

Te Amsterdam trad hij in 1883 toe tot de afdeling van de Sociaal-Democratische

1888

Op 23 mei 1888 trad hij in het huwelijk met Maria Margaretha Hofman, met wie

1889

De sociaal-liberaal A. Kerdijk merkte zijn parlementaire welsprekendheid op bij

1890

Van 1890 af verscheen te Maastricht De Volkstribuun, waarvan hij schrijver, zetter,

1893

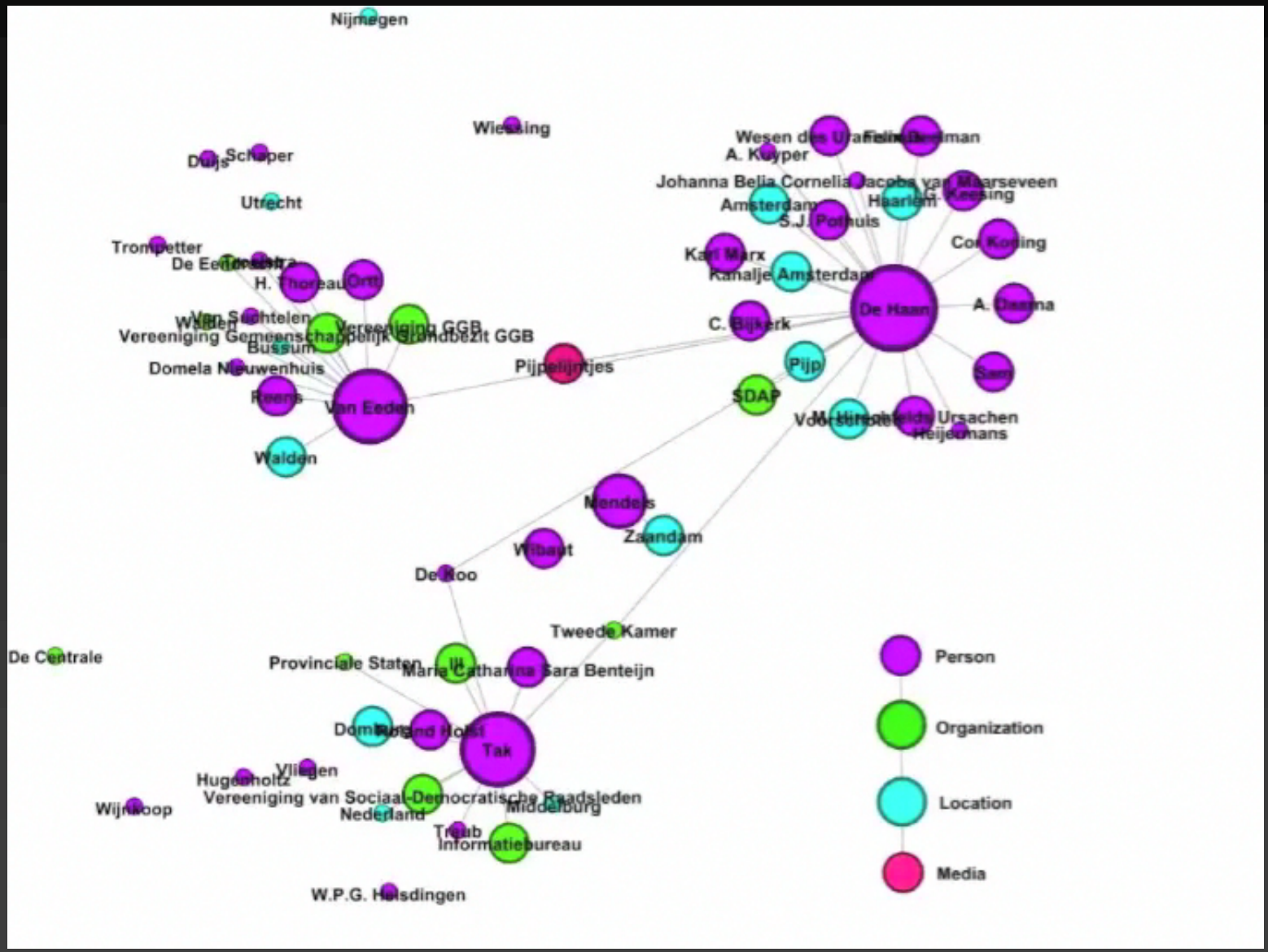
Toen de SDB vanaf 1893 in anti-parlementair vaarwater raakte, schroomde

1894

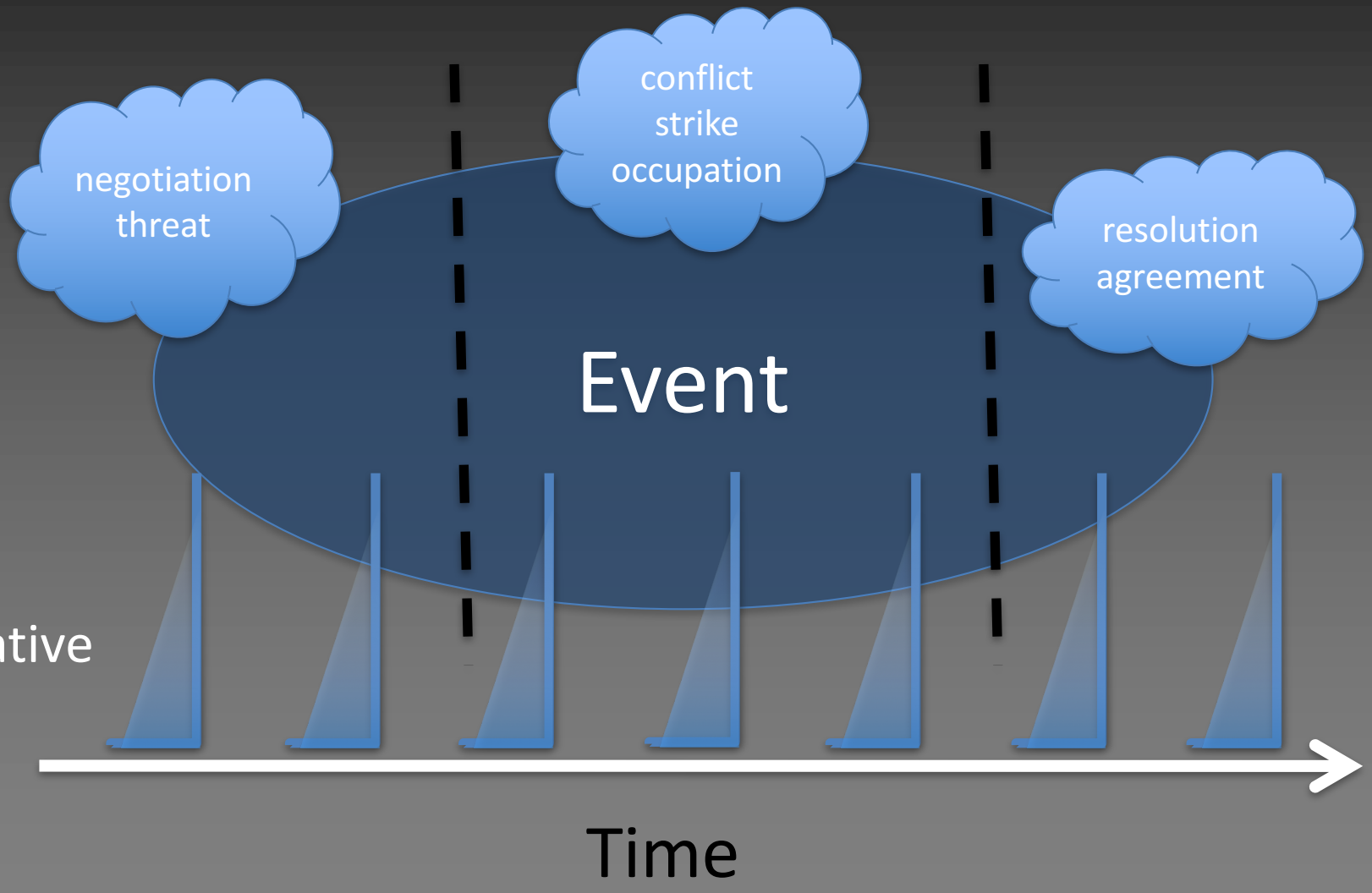
Connecties

- Franc van der Goes
- F. Domela Nieuwenhuis
- G.H. Pieters
- F.M. Wibaut
- dr. D. Bos
- J.H.A. Schaper
- W. Drees
- J.W. Albarda

The Socialist Network



Events in time



Strike database

- Designed and created by Sjaak van der Velden
- Contains over 16,000 strikes in the Netherlands

Bedrijf	Leidsche Katoenmaatschappij
Plaatsen	Leiden (Zuid-Holland)
Beroep	katoenwever
Eis	tegen loonsverlaging
Sector	Industrie/bouw
Soort actie	Staking
Type staking	Klassiek
Karakter	Vakbond
Resultaat	Verlies
Datum	22 februari 1922
Duur	133 dagen
Verslag	De bonden wilden eind mei een eind aan de strijd, maar de stakers weigerden dit. Waarnemend burgemeester Van der Lip deed een poging tot bemiddeling. Half augustus was vrijwel iedereen weer aan het werk. Veel sympathie van de bevolking. Van de stakers waren er 213 bondslid. In de vergadering van 27 juni van de Arbeidsraad bedankte de directie de bazen voor hun hulp tijdens de staking.
Aantallen	Stakers (hoogste aantal stakers): 265 Gestaakte dagen (totaal niet-gewerkte mensdagen): 30210 Onvrijwillige stakers (werknemers die door de actie niet verder kunnen werken): 0 Verloren dagen onvrijwillige stakers (arbeidsdagen dat de niet stakers niet konden werken): 0 Aantal betrokken bedrijven (aantal betrokken bedrijven): 1

Newspaper archives

- National Library, The Hague
  - Newspapers from 1618 to 1995
  - Over 11 million fully OCRd pages (delpher.nl)
  - Searchable on title, date, words
- Cf. *Forces of labor*, Silver (2003)

Example query

```
SELECT all articles
CONTAINING THE WORDS
stak?n* AND
(blokband OR Amsterdam OR
  taxichauffeur)
AND artikeldatum BETWEEN 6 apr 1937 – 7
                         AND 9 apr 1937 + 3
```

De staking der Amsterdamsche taxichauffeurs

Mededeelingen van werkgeverszijde

DE RIJKSBEMIDDELAAR VRAAGT INLICHTINGEN.

Naar wij vernemen, heeft de rijksbemiddelaar, mr dr S. de Vries Czn, de werkgevers in het taxibedrijf te Amsterdam verzocht hedenmiddag een afgevaardigde naar Den Haag te zenden om hem in te lichten omtrent de staking in het blokbandtaxi-bedrijf te Amsterdam.

De voorzitter van het bestuur der taxi-centr
de heer C. J. van Leusden, heeft in verba
met de staking medegedeeld, dat tot Woensd
avond noch het bestuur der Taxi-centrale, n
de daarbij aangesloten ondernemers, van
chauffeurs eenig verzoek hebben ontvangen
in verband met de verhooging der tarieven
loonen te herzien. De staking is Dinsdagmidd
uitgebroken zonder dat te voren overleg is
pleegd en ondanks het feit, dat er tusschen
werkgevers in het blokband-taxibedrijf en
vier groote organisaties van transportarbeid
waarbij de chauffeurs zijn aangesloten, een d
het gemeentebestuur goedgekeurde collecti
arbeidsovereenkomst bestaat, waarvan art
bepaalt, dat beide partijen zich verbinden
durende den duur der overeenkomst g
werkstakingen of uitsluitingen tegen elkan
te proclameren of uit te voeren of te doen
voeren.

Deze staking moet dus, volgens het oord
der werkgevers, als een wilde staking wor
beschouwd.

De staking der Amsterdamsche taxichauffeurs

Rijksbemiddelaar verklaart zich onbevoegd verder van het geschil kennis te nemen

DE GELDENDE ARBEIDSOVEREENKOMST VAN KRACHT.

De Rijksbemiddelaar in het 2e district, mr S. de Vries Czn, heeft gisteren in het Departement van Sociale Zaken een conferentie gehad met de werkgevers- en werknemersorga
trokken bij het conflict in het st
taxi-bedrijf te Amsterdam.

Allereerst is daarbij komen vast
de collectieve arbeids-overeenkoms
op 1 Augustus 1936, thans nog geld

Artikel 3 dier C.A.O. bepaalt, dat
tuerende partijen en de leden der
organisaties zich verbinden, gedurer
van de overeenkomst geen werksta
sluiting tegen elkander te proclam
te voeren of te doen uitvoeren.

De bedoeling der staking is, om
duur der C.A.O. daarin wijzinge
brengen, op grond dat gevreesd w
tariefsverhooging een nadeeligen
hebben op de loonen.

Naar de meening van den Rijks
is door de stakers niet den juiste
wandeld. Immers in de C.A.O. is g
gelegd tusschen den ritprijs en he

DE AMSTERDAMSCHE TAXI-STAKING GEëINDIGD.

Tariefsverhooging blijft gehand haafd.

Concessies aan de chauffeurs.

De stakende Amsterdamsche taxi-chauffeurs hebben op een gistermiddag in „Krasnapolsky" gehouden vergadering besloten de staking op te heffen, op basis van het bemiddelingsvoorstel, dat een concessie aan de werknemers genoemd kan worden. Des avonds heeft de avondploeg zich bij de garages gemeld en de chauffeurs reden met hun wagens naar de standplaatsen. Amsterdam was weer uit den taxi-nood!

De getroffen regeling komt in het kort

dhaafd
ats van

wordt
en voor

onder
menko-
t stand

oor het

„Sit Down" als Reclame!

In een slagerij in de Kinkerstraat te Amsterdam is door het personeel, vermoedelijk met medewerking van den patroon, een sit-down-staking in de winkel geënscèneerd. Deze „staking" was, naar A.N.I.P.-Aneta uit Amsterdam seint, klaarblijkelijk bedoeld als reclame!

De politie moest charges uitvoeren om de straat voor het verkeer vrij te houden.

De „staking" werd inmiddels opgeheven.

Strikes that never happened

- Many strikes have a lead-in “prelude” period (often associated with social unrest)
- Prelude does not necessarily lead to strike
- By finding prelude periods typical for strikes, can we find **strikes that never happened**?

Word frequency

Discriminatory power

Did we find strikes that did not happen?

Check against database: e.g.

- 1912: policemen in Amsterdam, tram conductors The Hague (2 articles), and teachers (3 articles)
- 1926: butchery in Boxmeer; transport workers of Steenkolen Handelsvereeniging
- 1938: metal factory in Delft

Mean average precision: 0.43

Van den Hoven, M., Van den Bosch, A., and Zervanou, K. (2010). Beyond reported history: Strikes that never happened. In S. Daranyi and P. Lendvai (Eds.), *Proceedings of the First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts*, Vienna, Austria, pp. 20-28.

Computational Counterfactual History

A simple mixed techniques study

"Text Analytics for Detecting Alternative Perspectives on Events at the End of the Second World War in the Arnhem-Nijmegen Region" (with Marten Düring)

- Operation Market Garden
  - 17-25 September 1944
  - about 34,000 parachuters and *a bridge too far*
- 7 month stand-still and evacuations
- Operation Veritable: 1,000,000 soldiers in the region

Sources

Airborne Museum Hartenstein (Oosterbeek)

- About 60,000 digitized items such as memoirs, diaries, photos, reports, historical narratives…

  Most of which have been OCRed or contain metadata

- Dutch, English, German

- Military, civilian, police, academic discourses

Overall goal

Help historians and curators with the laborious task of finding relevant information by (semi-) automatically linking related sources. Create a basis for interpretation of

1. Events - **Link primary sources through text analysis**
2. Motifs – **Study the changing(?) building blocks of war narratives**

Aggregate depictions of events based on location, time and type of event

We are looking for:

interconnected actions which are identifiable by mentioned points in time and/or a specific location in the Arnhem/ Nijmegen area between September 1944 and May 1945.

Practical applications:

- Highlight contradictions and ambiguities of sources and their interpretations. Patterns?
- Who depicts what? Patterns?
- Changes over time? Which events have (not) been documented?
- Identify new, potentially relevant sources
- Raise new questions:
  - Why doesn't SourceX mention EventY?

Case study

- 85 texts, digitally transcribed or born digital

	Civilian narratives	Soldier narratives	Historical monographs	Other
Dutch	53	-	-	16
English	1	10	3	1

- Starting point: street names
  - Regexp
  - NER

Named Entity Recognition

- Textcat language identification
- Frog: open source NLP tool for Dutch
  - Memory-based, http://ilk.uvt.nl/frog
- Stanford NER
  - CRF-based, http://nlp.stanford.edu/software/CRF-NER.shtml

```
1     Marie    Marie    [Marie] SPEC(deeleigen)                  1.000000   B-PER   B-NP    2   su
2     vroeg    vragen   [vraag] WW(pv,verl,ev)                   0.532544   0       B-VP    0   ROOT
3     zich     zich     [zich]  VNW(refl,pron,obl,red,3,getal)   0.999740   0       B-NP    2   se
4     af       af       [af]    VZ(fin)                          0.996853   0       0       2   svp
5     of       of       [of]    VG(onder)                        0.733333   0       B-SBAR  4   vc
6     hij      hij      [hij]   VNW(pers,pron,nomin,vol,3,ev,masc) 0.999659  0       B-NP    8   su
7     nog      nog      [nog]   BW()                             0.999930   0       B-ADVP  8   mod
8     zou      zullen   [zal]   WW(pv,verl,ev)                   0.999947   0       B-VP    5   body
9     komen    komen    [kom][en]       WW(inf,vrij,zonder)     0.861549   0       I-VP    8   vc
10    .        .        [.]     LET()                            0.999956   0       0       9   punct
```

Named Entity Recognition

- Frog finds 6,927 locations in 71 Dutch texts
  - 201 Arnhem; 195 Rhenen; 140 Oosterbeek; 106 Amersfoort; 98 Spakenburg; 97 Duitsland; 79 Scherpenzeel; 68 Rijn; 64 Ede; 61 Nijkerk
- Stanford NER finds 17,221 locations in 13 English texts
  - Without work on Canadian army: 8,170 locations
  - 395 Aachen; 182 Antwerp; 129 Nijmegen; 123 Arnhem; 122 Germany; 107 Holland; 105 London; 92 Normandy; 92 Netherlands; 90 Geilenkirchen

Identifying street names

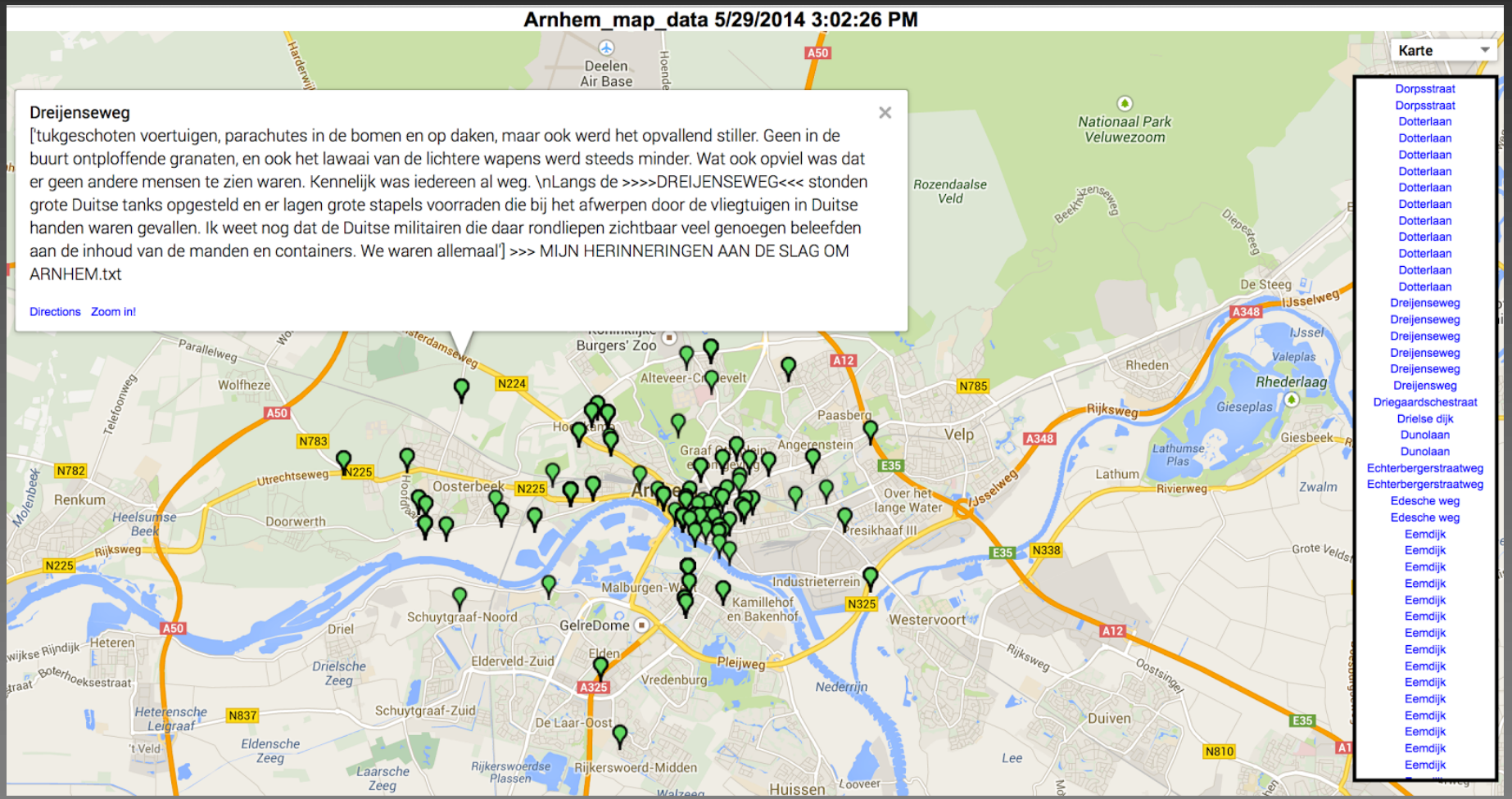
Regular expression

- *road, *street, *boulevard, *avenue, *path, *square, *straat, *laan, *weg, *plein, *dijk, *pad, *brug, *park, *singel

Evaluation

- Precision: 182/205 = 88.8%
- Recall: 182/203 = 89.7%
- F-score: 89.2%

Mapalist + Google Maps



Automatically found: Dreijenseweg

Civilian reports seeing German tanks along the road (18 Sept 1944)

- *… Langs de DREIJENSEWEG stonden grote Duitse tanks opgesteld en er lagen grote stapels voorraden die bij het afwerpen door de vliegtuigen in Duitse handen waren gevallen. Ik weet nog dat de Duitse militairen die daar rondliepen zichtbaar veel genoegen beleefden aan de inhoud van de manden en containers. …*

Allied soldier describing the skirmish in his memoirs (19 Sept 1944)

- *… Advanced through woods in area between Johannahoeve and DREIJENSEWEG, with Sgt Shepley, and three others of my company, I cannot recall their names; we were held up trying to cross a track.. …*

Police Report: Mrs. Jansen notifies authorities about a body (20 August 1945)

- *Mejuffrouw Jansen, wonende AMSTERDAMSEWEG No 244 (waterleiding) geeft kennis dat een Engelse soldaat begraven ligt in de berm van de DREIJENSEWEG juist in de bocht voorbij het huis van Buuren. …*

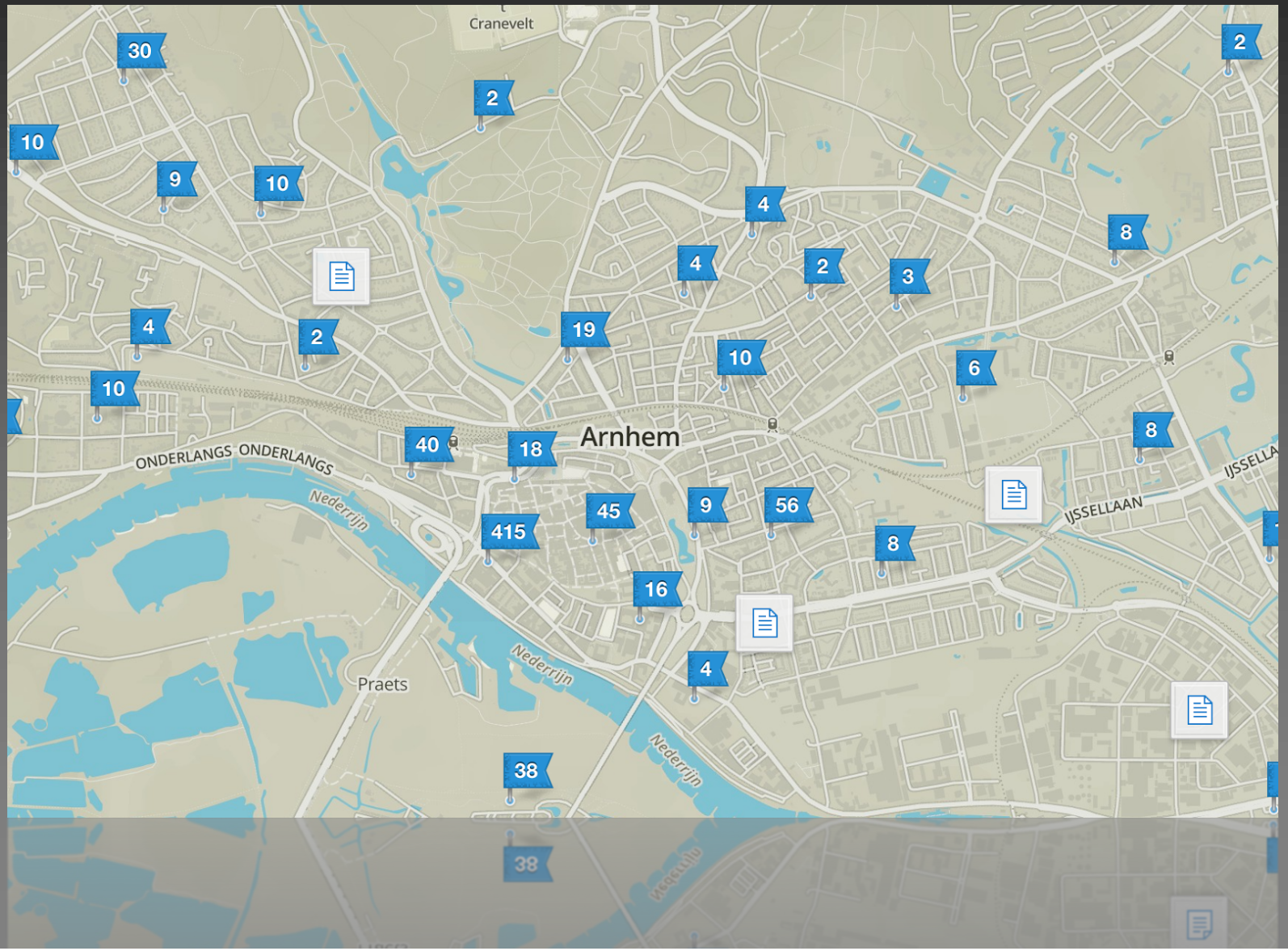
Police Report: Human bones and British gear found (30 October 1948)

- *In de tuin van het pand DREIJENSEWEG, bewoond door de Gruijter, is door opgraving verschillende beenderen van een menselijk lichaam en resten van uitrustingsstukken gevonden. Ook zit er (gezien de verpakking) nog een lijk in de grond. De uitrustingsstukken zijn afkomstig van Engelse …*

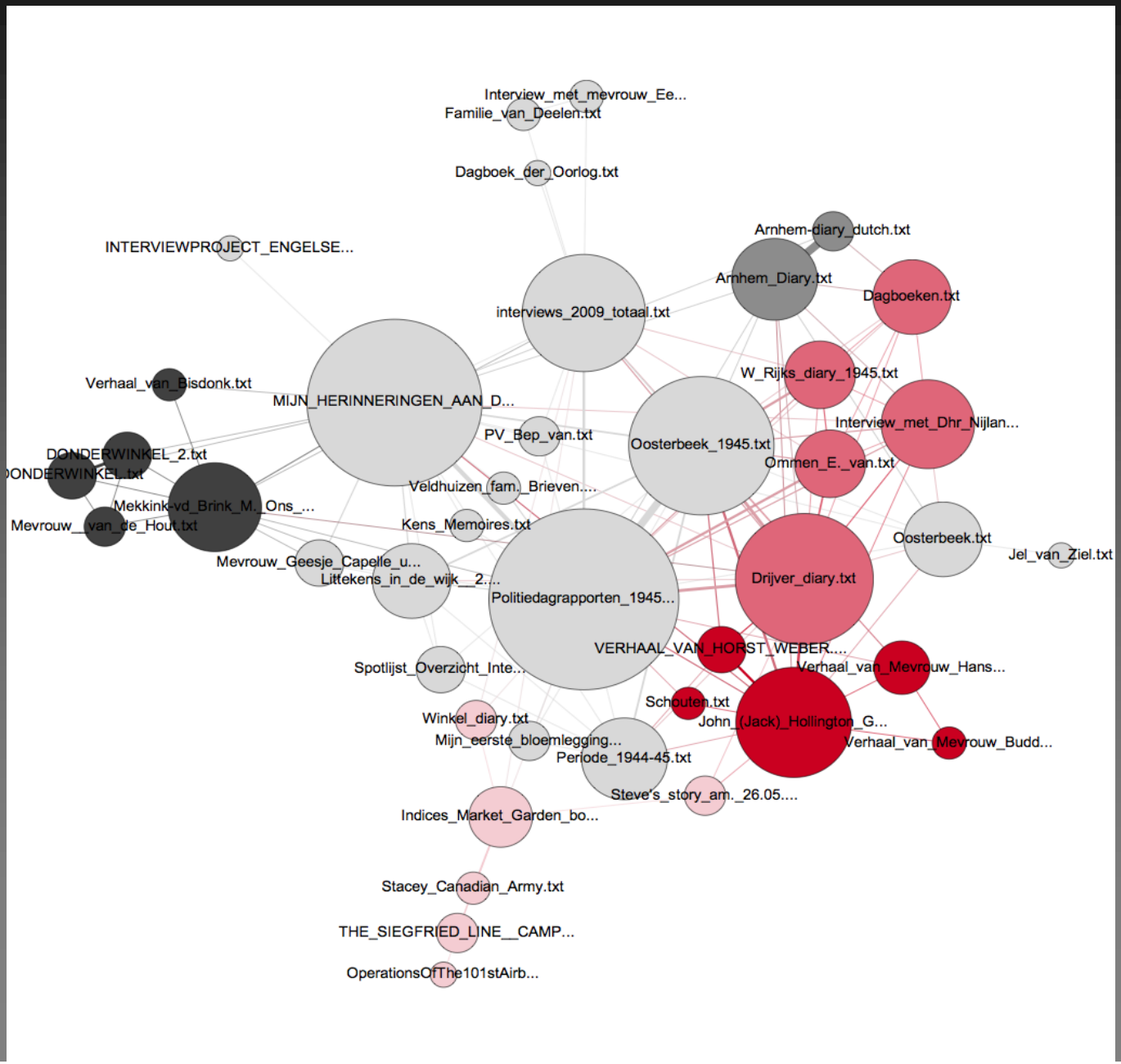
Police Report: Bodies of three British soldiers have been found (4 November 1948)

- *Door de Nederlandse en Engelse lijkendienst zijn in de tuin van de Gruijter, DREIJENSEWEG, alhier, drie lijken van Engelse militairen opgegraven en overgebracht naar de Engelse begraafplaats.*

Evernote’s Atlas function



Texts connected by streets (Gephi)



I had the most wonderful dream

Language Machines group project

- explicitly intended as a mixed bag exercise
- question provided by experts (dream psychology)
  - What distinguishes dream descriptions from other personal narratives?
- team:
  - Antal van den Bosch, Maarten van Gompel, Iris Hendrickx, Ali Hürriyetoglu, Folgert Karsdorp, Florian Kunneman, Louis Onrust, Martin Reynaert and Wessel Stoop
- (PDF slides)

A big thank you to:

- Language Machines
  - Maarten van Gompel, Iris Hendrickx, Ali Hürriyetoglu, Florian Kunneman, Louis Onrust, Martin Reynaert and Wessel Stoop
- HiTiME, ISHER, MERIT
  - Kalliopi Zervanou, Matje van de Camp, Steve Hunt, Martha van den Hoven, Marten Düring
- Tunes and Tales
  - Folgert Karsdorp, Theo Meder

antal.van.den.bosch@meertens.knaw.nl

What makes dream text dreamy?

A text analytics exploration of reported dreams
DHBenelux 2015

LANGUAGE MACHINES

CLS | Centre for Language Studies
Radboud University

Dreaming Language Machines

Antal van den Bosch, Maarten van Gompel, Iris Hendrickx, Ali Hürriyetoğlu, Folgert Karsdorp, Florian Kunneman, Louis Onrust, Martin Reynaert and Wessel Stoop

http://cls.ru.nl/languagemachines/

Overview

Motivation

Data
- Dreambank
- Comparable corpora of diaries, real stories, and dreams

Experiments
- Text Classification
- LDA topic modeling on the Dreambank
- *n*-gram comparison
- Coherence experiments

Conclusions

Why do we dream?

Dream analysis in its historical context

The analysis of dreams has a long history. One of the earliest recorded dream analyses was written on clay tablets in Mesopotamia, 5000 years ago (Black and Green, 1992). In ancient Greek and Egyptian times, dreams were seen as messages from the gods. Nowadays, many different fields study the meaning and purpose of dreams such as psychiatry, psychology, neuroscience and religious studies, but a definite explanation of the purpose of dreams is still far from being found.

What it is that makes a dream text different from other texts?

This question came up in our correspondence with dream expert Kelly Bulkeley.

Studies have shown [Domhoff, 2003, Bulkeley and Domhoff, 2010]

- content of dreams is closely related to daily life and personal concerns
- around 75-80% content: everyday settings, characters, and activities
- dream content is stable over time
- certain shared unusual topics: flying, teeth falling out, appearing without clothes in public

Our approach

Comparative study between reported dreams and those texts that are most similar to dreams text, namely personal stories like diaries, true stories and confessions. (if we compare reported dreams to newspaper text, it will be very simple to distinguish the two)

Method: "shoot with many guns"

Exploratory study to try out different angles to measure differences between dreams and personal stories

We present the following methods:

- text classication: can we automatically predict whether a personal narrative is a dream or not?
- *n*-grams: what type of patterns are typical for dreams?
- topics: do dreams contain special topics?
- coherence: are dreams less coherent than other personal narratives?

Data sources

Comparable corpora of diaries, real stories, and dreams

Dreambank Collection of 30K reported dreams gathered from personal diaries and scientific studies

Reddit Online community, where subcommunities can be formed to discuss a particular interest

Prosebox Online writing community, where members can read and write diary entries

Preprocessing for all three datasets:

- language identification (`langid.py`)
- tokenization (`twokenize`)
- random sample of 1.3 Mw

What is the Dreambank and who uses it?

Reported dreams

Dreambank is an archive of personal dream descriptions collected in different studies. In total the Dreambank contains 68 collections ($\pm$ 30K dreams)

143	Blind dreamers (M)	**42**	Joan: a lesbian
100	Chris: a transvestite	**234**	The Natural Scientist
681	College women, late 1940s	**1235**	Norman: a child molester
899	Dorothea: 53 years of dreams	**384**	Peruvian men
		106	Phil 1: teens
490	Hall/VdC Norms: Female	**219**	Phil 2: late 20s
		180	Phil 3: retirement
491	Hall/VdC Norms: Male	**33**	Toby: A friendly party animal

Dreambank studies

Previous research

- manual analysis: Hall/van de Castle codes
- Hall van de Castle provides annotations for e.g. emotions, roles, activities, settings, misfortune versus good fortune, objects
- psychologists form hypotheses based on the dreams; these are discussed with the patient (dreamer). For example, the relationship with the parents

Lama version of the Dreambank

- only reported dreams in English
- 22,323 dream descriptions
- converted to fixed format for automatic processing
- preprocessed with Stanford tools

Balancing the set

Since many subjects are followed for a longitudinal study, their dreams are overrepresented compared to other subjects for which only few dreams are available. We limited this skewness by limiting the number of reports per author to 100.

Comparable corpora of diaries, real stories, and dreams

Reddit

An online community, where subcommunities (subreddits) can be formed to discuss a particular interest.

- Anxiety
- Diaries
- depression
- Diary
- loseit
- mommit
- nosleep
- offmychest
- pettyrevenge
- ProRevenge
- psychonaut
- relationships
- relationship_advice
- self
- shortscarystories
- twoxchromosomes
- **dreams**
- badpeoplestories
- **storiesofdreams**
- **thisdreamihad**
- LetsNotMeet
- lifeinapost
- **TodayIdreamed**

Example post

In the first half, my dad brought me to a house where he, my brother, and I would stay for vacation. He then left with my brother so I was alone in the house.

I went around and I kept hearing someone else. Everywhere I went, I felt someone was in the house with me. Then the house warped into the house of a previous dream. [...]

Comparable corpora of diaries, real stories, and dreams

Prosebox

An online writing community, where members can read and write diary entries.

Example post

[. . .] You don't give me much reason to trust you anymore, but it's me that's untrustworthy. It's a ridiculous invasion of privacy to do what I did, what I continue to do. But otherwise I wouldn't have known, wouldn't have been able to prepare myself.
Firstly, ages ago now, the kiss with the girl when I was in Costa Rica. We weren't in a relationship but it was practically so. Then, more recently, the photograph of a you-know-what on your phone, sent to you by a friend. [. . .]

Text classification

Set-up

- Dream: 7000 Dreambank documents
- No dream: 2739 Prosebox documents
- Features: word uni-, bi- and trigrams (lowercase, stripped of punctuation)
- Weighting: Infogain
- Classifier: Balanced Winnow (Linguistic Classification System)
- 10-fold Cross Validation

Text classification

Results

	Precision	Recall	F1	TPR	FPR	AUC
dream	0.92	0.97	0.95	0.97	0.23	0.87
no dream	0.91	0.77	0.84	0.77	0.03	0.87

Text classification

Dreamy *n*-grams	
or	a
mary	table
girl	hall
my	bus
woman	of
tables	i_tell_him
window	on_the
the_ground	seem_to
white	was_there
asks	the_back

Non-dreamy *n*-grams	
it's	can't
'm	i_'m
2015	will
weight	thanks
day	truly
free	pain
ok	it's
still	5
work	thank_you
on_saturday	your

LDA topic modeling on the Dreambank

Purpose: unsupervised content analysis

- Mallet toolkit [McCallum, 2002]: LDA with 2000 iterations with Gibbs sampling and 50 topics.
- only look at topics with proportion >0.1 per document (avg. of 3 topics per document)
- Look for significant differences in topic distributions for balanced subsets of the dream bank: men vs. women

LDA topics confirm observation that dreams relate to daily events and concerns

Some topic examples

0 store money buy get pay bill man grocery lot counter bank shopping machine give tickets shop put dollars change bought

3 bathroom toilet hair shower water go room bath floor clean wash see naked sink tub pee bathtub face cut towel

8 book paper read books find picture write writing reading letter pictures written name looking look something library letters office computer

Comparison on subset Hall vd Castle men vs. women (loglikelihood test)

Topics most dreamed by women

10 house mother father baby old boy brother family girl home little sister children years parents child see aunt son kids

13 church wedding people married front sit seats seat aisle back table sitting ceremony group place priest getting side room chair

43 wearing white dress black blue red clothes hair dressed shirt put shoes wear pair pants hat green suit see old

15 pool swimming water swim go board end burt dive diving bottom little bathing side deep suit lake watching shallow underwater

19 play dance stage music audience group playing song people dancing singing show sing part piano good band performance do concert

Comparison on subset Hall vd Castle men vs. women (loglikelihood test)

Most dreamed topics by men

24 car driving drive road truck get going go seat side back cars stop front parked turn driver street station parking

29 game ball playing team play basketball football field high baseball coach good balls hit players tennis school man other player

31 gun men man shoot shot people fire kill shooting guns police war run killed enemy escape being soldiers fight get

35 building stairs floor go get elevator people going door top hall room roof wall high office find steps way ladder

LDA-based topical cosine similarities

Cosine similarity in 200-dimensional space

- POS filter: only NN and VBD
- gensim LDA: 2500 iterations, 200 topics, $\alpha \propto |\text{tokens}|$

	dR	aD	D	aR	nR	P
(dreamreddit) dR	1.00	**0.84**	**0.79**	0.60	0.59	0.59
(alldreambank) aD	**0.84**	1.00	**0.95**	0.70	0.71	0.70
(dreambank) D	**0.79**	**0.95**	1.00	0.69	0.68	0.70
(allreddit) aR	0.60	0.70	0.69	1.00	**0.99**	**0.87**
(nondreamreddit) nR	0.59	0.71	0.68	**0.99**	1.00	**0.87**
(prosebox) P	0.59	0.70	0.70	**0.87**	**0.87**	1.00

Findings

- dreamreddit looks more like dreambank than posts from other subreddits (0.8 vs 0.6)
- diary entries from prosebox are very similar to the non-dream posts from reddit

n-gram Loglikelihood

Experiment

Which *n*-grams are indicative of dreams?

1. Extract *n*-grams from dream data
2. Extract *n*-grams from control data
3. Compute log-likelihood comparison

n-gram Loglikelihood

Findings
Indicative of dreams:

- **The obvious** *dream*, *dreamt*, *the dream*, etc.. . .
- **Past tense of "to be"** *was*, *were* (in top 3)
- **Setting the scene** *was in*, *there was*, *room*, *house*, *street*, *car*, *I was with*, *was trying to*, *sitting*, *driving*, *scene*
- **Feelings and qualities** *feel*, *love*, *strange*, *big*, *large*
- **Recollection** *remember*, *I can't remember*, *I recall*
- **Relating, vagueness** *seemed (to/like)*, *feel like*, *(seemed) as if*, *some sort*, *or something*
- **Sequential narrating** *Then*, *Then I*, *Suddenly*
- **Sleeping and waking** *sleep*, *I woke*

Coherence Experiments

Hypothesis: dreams are less coherent in their structure than personal stories

We try out two methods:

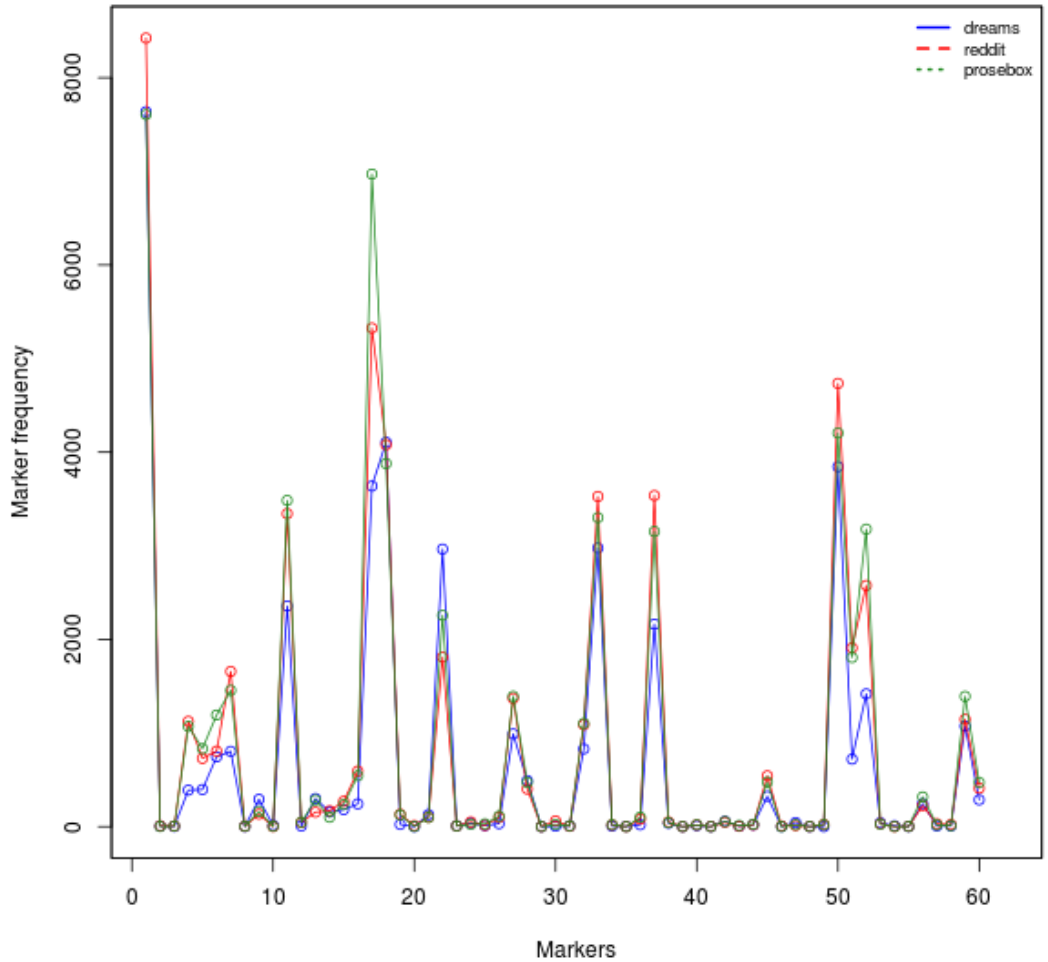
- simply study the amount and type of discourse markers in the texts
- look at entity-based coherence

Discourse marker frequency

Simple count of discourse markers. List of 61 markers based on annotations from the Penn Discourse Treebank used in this year's ConLL shared task.
Examples: *but*, *meanwhile*, *only*, *when*, *since*, *until*, *though*, *after*, *because*, *if*, *for example*, *finally*, etc.

Results
Dreams contain fewer discourse markers. The only one that occurs much more often in dreams is **then** (2965 vs 1810 in Reddit and 2257 times in Prosebox).

Discourse marker frequency



Coherence Experiments: Binary Discrimination Test

- Brown Coherence Toolkit v1.0 [Elsner and Charniak, 2007]
- The binary discrimination test tests the model's ability to distinguish between a human-authored document in its original order, and a random permutation of that document.
- The test reads any number of documents and performs the test on each one, using 20 random permutations.

	Accuracy	F-score
Dreambank	0.23	0.32
Prosebox	0.37	0.42
Reddit	0.37	0.43

What makes a dream text dreamy?

- Sequential narration (. . . *then* . . .), like fairytales
- Scene descriptions
- Trouble recalling
- Meta language (*sleep*, *dream*, *woke*)
- Somewhat lower coherence than other personal narratives

Reported dreams can be distinguished from other personal narratives to a fairly high degree (dream class: .95 F-score, .87 AUC)

Back to the dream experts and the drawing board.

References


Bulkeley, K. and Domhoff, G. W. (2010).
Detecting meaning in dream reports: An extension of a word search approach.
Dreaming, 20(2):77.


Domhoff, W. G. (2003).
The Scientific Study of Dreams: Neural Networks, Cognitive Development, and Content Analysis.
American Psychological Association (APA), 1 edition.


Elsner, M. and Charniak, E. (2007).
A generative discourse-new model for text coherence.
Technical report, Technical Report CS-07-04, Brown University.


McCallum, A. K. (2002).
Mallet: A machine learning for language toolkit.
http://mallet.cs.umass.edu.