Analyzing and Improving the Quality of a Historical News Collection using Language Technology and Statistical Machine Learning Methods

Kimmo Kettunen[1], Timo Honkela[1,2], Krister Lindén[2], Pekka Kauppinen[2], Tuula Pääkkönen[1] & Jukka Kervinen[1]

1640 THE NATIONAL LIBRARY OF FINLAND

1

Presented by
Timo Honkela

in

IFLA Pre-Conference
Geneva, Switzerland,
13th of August, 2014

2

UNIVERSITY OF HELSINKI

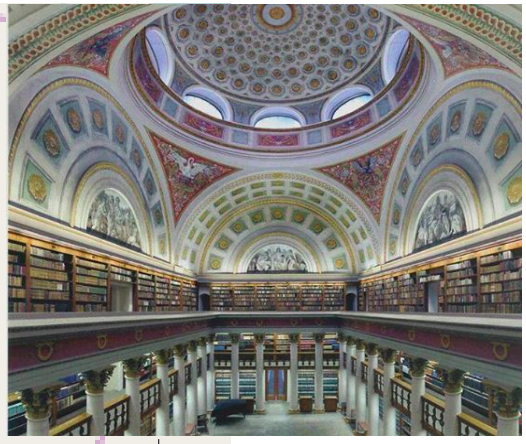
European Union
European Regional Development Fund

Leverage from the EU 2007–2013

UNIVERSITY OF HELSINKI

Department of
Modern Languages

Language Technology

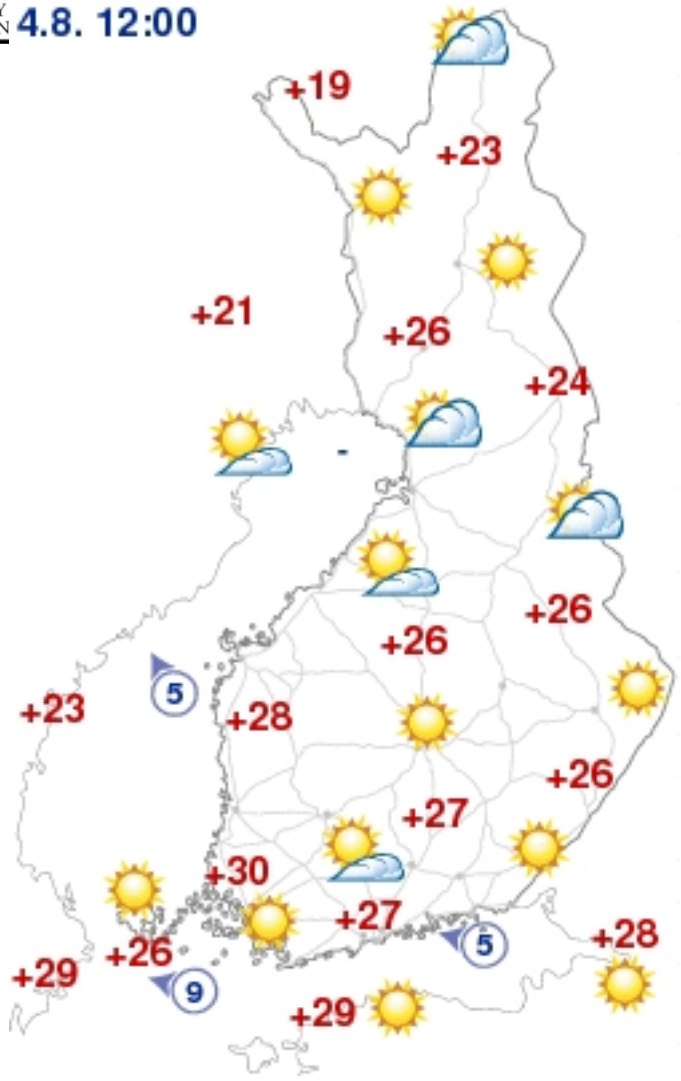


1640
THE NATIONAL LIBRARY OF FINLAND

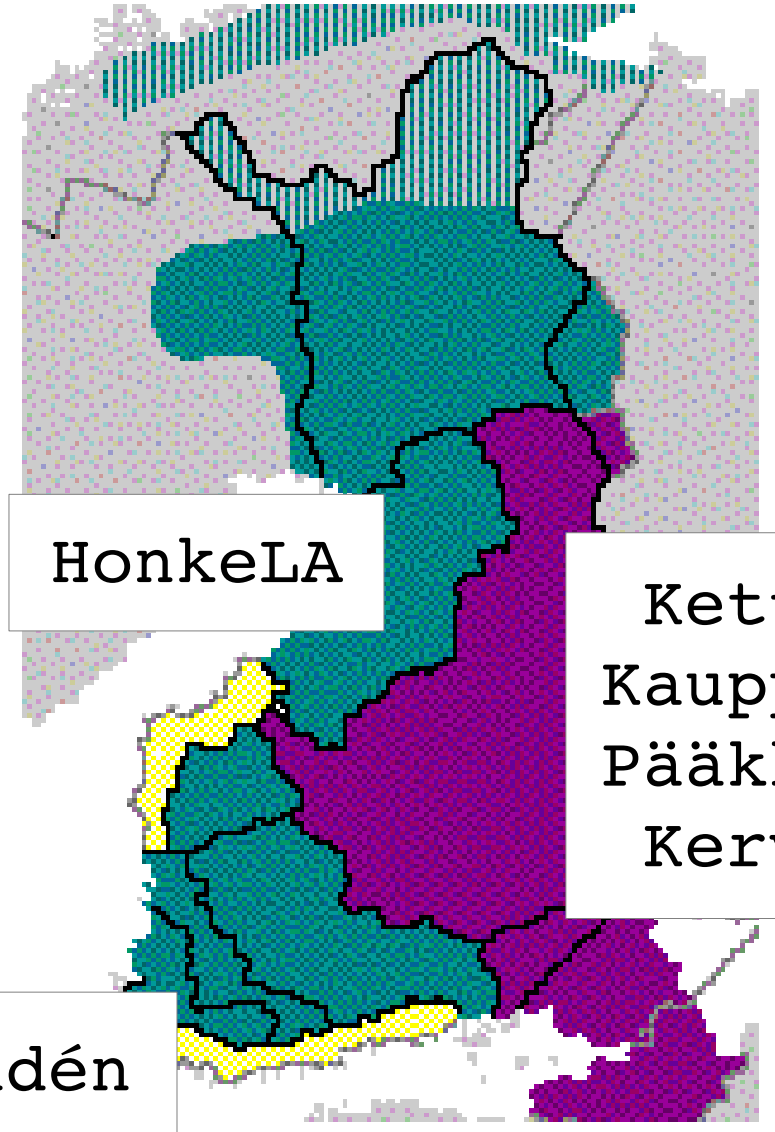
Center for Preservation
and Digitisation

HELSINKI

MIKKELI



www.fmi.fi



http://oppimateriaalit.internetix.fi

Structure of the presentation

- Some background on the digitalization process
- Introducing the paper content:
  analysis and correction of OCR results
- Discussion on future steps:
  In-depth analysis of newspaper contents to promote research in humanities and social sciences

Historical newspaper collection

- The National Library of Finland has digitized a large proportion of the historical newspapers published in Finland between 1771 and 1910 (Bremer-Laamanen 2001, 2005).
- This collection contains approximately 1.95 million pages in Finnish and Swedish
- According to Legal Deposit law, the National Library of Finland receives a copy of each newspaper and magazine published in Finland.

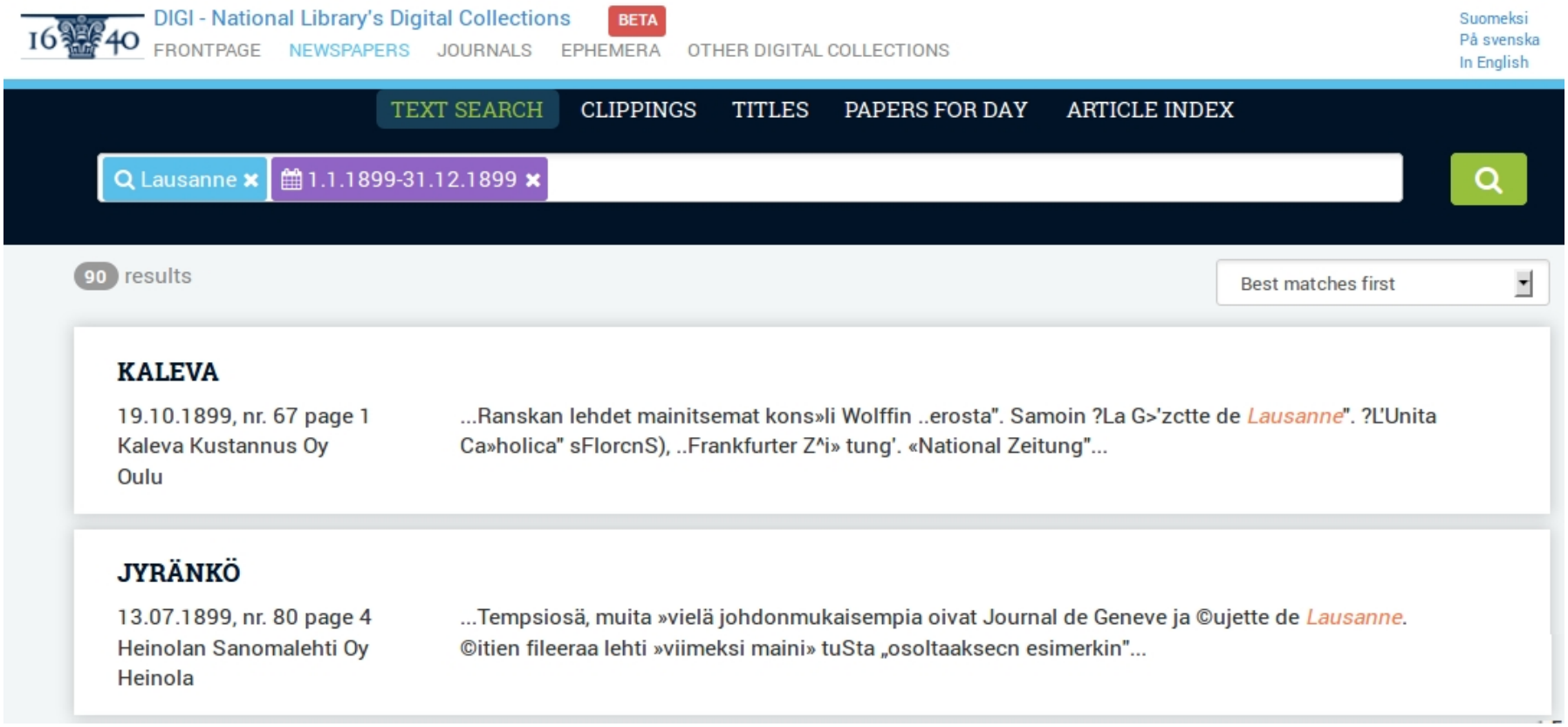
Digitisation of the historical newspaper collection

- In the post-processing phase, the material is processed so that it can be shared to the library sector, researchers, and the wide public.
- The scanned images are enhanced and run through background software and processes which create METS/ALTO metadata (CCS Docworks)
- The optical character recognition (OCR) is conducted at the same time in order to get the text content from the materials.

Two channels

- Search and exploration interface (“Digi”)
  - Approximate search, focusing based on time/place, indexed contents,
    index creation using morphological analysis, etc.
  - Digitalkoot: enables the public to collectively mark and collect articles (crowdsourcing)
- Corpus (FIN-CLARIN)
  - Mainly used by linguists
  - Includes keyword-in-context (n-gram) view
  - Morphological and syntactical analysis results

Search interface



http://digi.kansalliskirjasto.fi

FIN-CLARIN corpus

— « (iazette de **Lausanne** " painatti t. f:n 15 p:nä ju »
Gazette de **Lausanne** " >! e! jbe3fs tehtiin t. tn 21 p:nd sello, lur
de **Lausanne**, Journal d'Alsac, L? Aurore ja Revue Chrét
nal des Debats, Le Tenips, Gazette de **Lausanne**, Soleil, MatiD, Journal de peuple, Daily N
ja Journal « Freiburg) y. m. Tribune (**Lausanne**) » n saanut Tukholmista kirjeen lähetys »
Samoin La Gazette de **Lausanne** ", L'Unita Catuoko » " Elorens),..
Gazette de **Lausanne** " mainitsee siitä omituisesta tapauksesta
Gazette de **Lausanne** " sisältää kolme palstaa pitkän pääkirjoitu
La Gazette de **Lausanne** " sisälsi t, kin 16 p:nä kirjoituksen siirtola
— » (Facette de **Lausanne** » paicattt t. l. 15 paä iu » Btu2Rijan.
l des Delats, la » l^mps, <3 » i. ette de **Lausanne**, Soleil, Matin, Journal tie peuple, Daily _
Gazette de **Lausanne** julkaisee kappalem Tukholman Aftonblac
owat journal de Geneve ja Oajette be **Lausanne**.
Gazette de **Lausanne** si< sältää pitkän suosiollisen kirjoituksen
Lontoo), La Gazette de Laussanne " (**Lausanne**) ja La Fronde'- (Paris).
nal des Debats, L>- Temps, Gazette de **Lausanne**, Soleil, Matin, Journal de penple, Daily N
hän sen johdosta, kertoo Gazette de **Lausanne**, kirjaston tirehtööriltä saanut käännökse
» Laitto de **Lausanne** " -nimi|en Sroettsin lehben l. l:n 11 p:n n:
onale, Gaulois, Petit Bleu, Gazette de **Lausanne**, Journal d'Alsac, L'Aurore ja Revue Chréti
lomaan! » n » maleh disjä. Ga; etle de **Lausanne** " sisällää l^m? palslaa pitkän pääkirjoituk
ssarosch seikkailuista), La gazette de **Lausanne** (toimenpiteistä, jotka tarkoittavat selitys
Freiburg, Geneve, Intnrlakeu, **Lausanne**.

Corpus

KLK suomi 1899

text attributes

numero: 87
julkaisu: Mikkelin Sanomat
numeron nimi: Mikkelin Sanomat
kieli: fi
päiväys: 03.08.1899
ISS: 1458-2481
digitoitu (pvm): 2008-10-14
parse state: parsed
sivunumero: 3
lehti: Mikkelin Sanomat no. 87 03.08.1899

word attributes

part-of-speech: noun
baseform: Lausanne
OCR: 0.95
perusmuoto (yhdyssanarajat): Lausanne
msd:
SUBCAT_Prop|NUM_Sg|CASE_Nom|CASECHANG
dependency relation: nominal subject

www.kielipankki.fi

OCR Challenges

- Regardless of recent development of the OCR software, there are still challenges with it, as some material is very old, with
  - varying paper and print quality,
  - varying number of columns and layout patterns,
  - different languages (mainly Finnish and Swedish but also French, German, etc.), and
  - and varying font types (fraktur and antiqua)

OCR Challenges

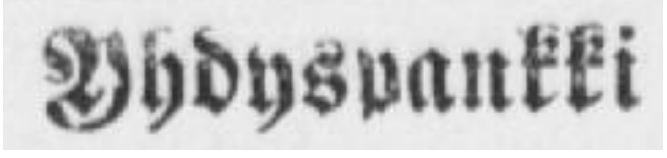
- The amount of material is such that human efforts – even crowdsourced – can only be a partial solution
- Fully or partially automated processes are needed

A very long tail of low frequency forms...

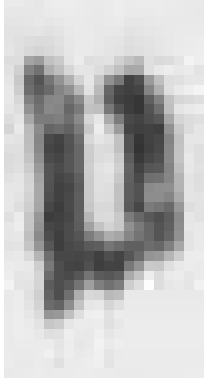
Freq.	OCR form	Correct form	Edit distance	Translation
1	zzhdysvautki	yhdyspankki	5	union bank
1	zzznuirypäleitä	wiinirypäleitä	4	grapes
1	wiljelystartaltutsessa	wiljelystarkoituksessa	4	in a cultivation purpose
1	urheilutarloinksiin	urheilutarkoituksiin	3	for sports purposes
1	uratkakupoissa	urakkakupoissa (urakkakaupoissa)	1	in contract jobs
1	taitanuiubcsta	taitawuudesta	4	of dexterity
1	taitamattomundestani	taitamattomuudestani	1	from my ineptitude
1	taiötelelutanteren	taistelutanteren	4	of the battlefield
1	taioafliftiutpn	tavallisuuden	8	of the usual
1	taimokkaisuudclllllln	tarmokkaisuudellaan	6	with his/her vigor

SUOMALAINEN WIRALLINEN LEHTI

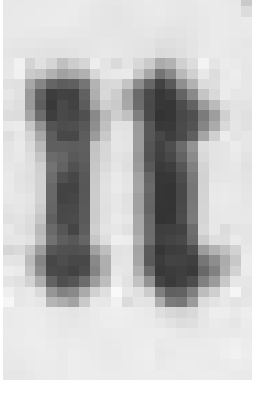
12.02.1876, nro. 17 sivu 4
Senaatin Kansliatoimituskunta
Helsinki
...§§ yhtiö» säännöissä mainilsemet. Wua< sassa, l p. helulilunta 187 tt, *zzhdysvautki*
Suomessa. (339) Hymälsyttyjen »valicn lautta on malittu tn» lemaan Mliotecn...



zzhdysvautki Yhdyspankki



v, u, p ?



u, n, ll ?

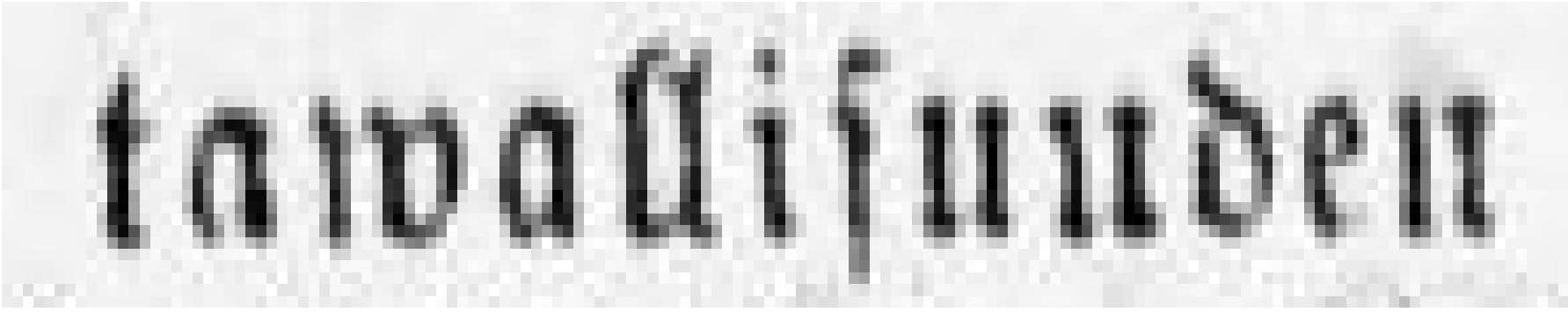
SATAKUNTA

25.04.1885, nro. 33 sivu 1

Oy Satakunta

Pori

...ojaa; liittoina tonfotuim .'> u:uii iuiibc« Iniipun» Toimitus, jota *taioafliftiutpn* mnfnait pi tietääit lnlon sataristosla, aljetaan ffUo !> aamulla. Povi...



taioafliftiutpn tavallisuuden

Sources of complexity

Word (lexeme)

→ Inflections

⟶ Historical differences

⟶ Typos

⟶ Recognition errors

⇨ “Recognized” surface word

Inflections:

Complexity of Finnish at the level of word forms

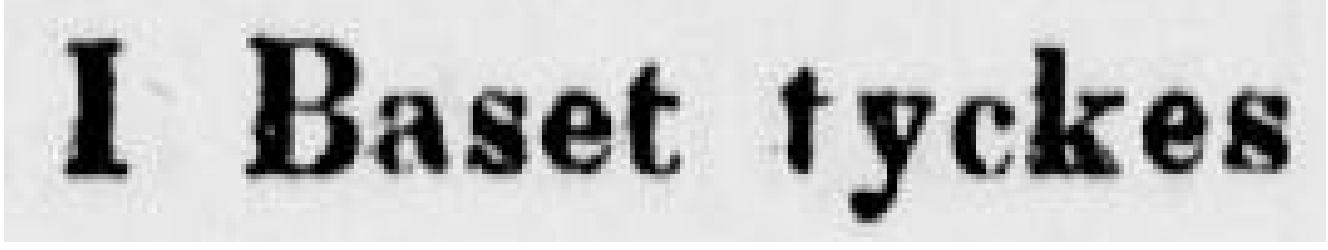
	kerroin	*yhteensä erilaisia muotoja*
perusmuoto: KATTO		1
yksikkö ja monikko: KATTO, KATOT	2	2
sijamuodot: N, A, NA, KSI, SSA, STA, VN, LLA, LTA, LLE, TTA, INE, IN	13	26
omistusliitteet: NI, SI, VN, NSA, MME, NNE	7	182
liitepartikkelit: KIN, HAN, PA, KO, …	11	2 002
yksiosaisille substantiiveille: AALTO, AAMU, …, KATTO, …	90 000	180 180 000
kaksiosaisille yhdyssanoille: AALTO-PELTI, …	180 000	32 432 400 000 000
kolmiosaisille yhdyssanoille: AALTO-PELTI-KATTO, …	180 000	5 837 832 000 000 miljoonaa
neliosaisille yhdyssanoille: JAUHE-LIHA-MAKARONI-LAATIKKO	180 000	1 050 809 760 000 miljoonaa miljoonaa

Kimmo Koskenniemi (2013): Johdatus kieliteknologiaan, sen merkitykseen ja sovelluksiin (Introduction to language technology, its significance and applications)

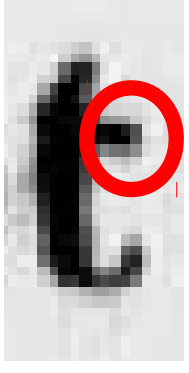
https://helda.helsinki.fi/bitstream/handle/10138/38503/kt-johd.pdf?sequence=1

Typos

Not a major source of problem but they do exist



Basel



Most likely not a stain

Historical differences

- All the time, new names and words are being introduced
- Even more static morphological aspects evolve over centuries

Net outcome

- A collection of millions of newspaper pages gives rise to a list of hundreds of millions of different word forms that have been found in the process
- A large proportion of these forms is not correct

Detection and correction

- Improving OCR quality – not considered here
- Improving the OCR output based on linguistic knowledge and statistical considerations
  - Detecting incorrect forms
  - Correcting the incorrect form

Introduction to the basic ideas

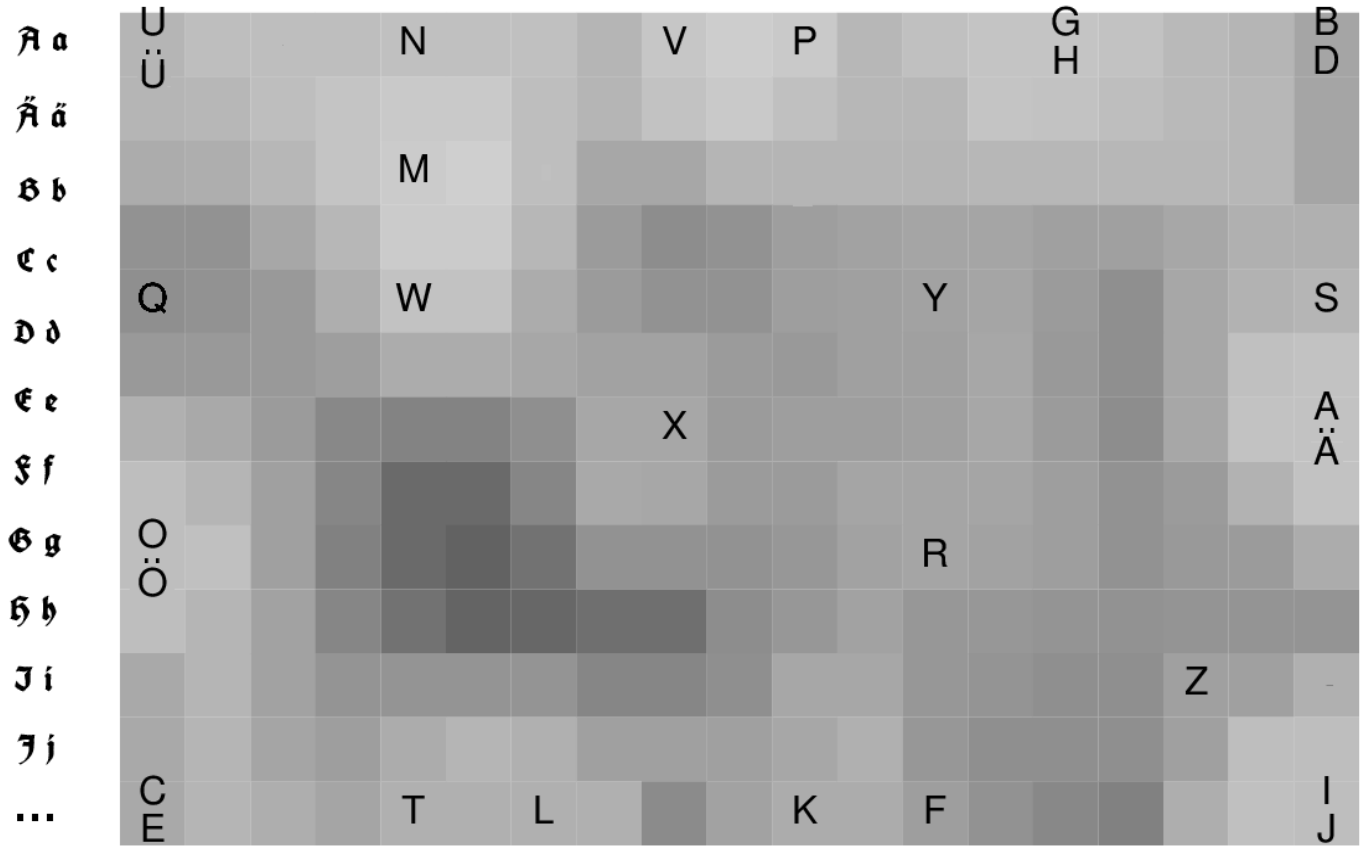
- Detection
  - Morphological analyzer
  - Special dictionaries (e.g. names)
  - N-grams
- Correction
  - Transformation rules created through a supervised learning scheme
  - Edit distance approach using corpus statistics
  - Weighted edit distance based on letter shapes
  - Future: context information (problem of sparsity)

Please see the paper for methodological details and analysis results

Frequency	OCR form	Translation
23818195	ja	and
12473329	on	is
5737985	että	that
4003638	oli	was
3224891	ei	no
2501891	niin	so
2465708	hän	he, she
2457173	se	it
2447894	joka	that, which
2188373	sen	its

Freq.	OCR form	Correct form	Edit distance	Translation
100	ytsimicliscsti	yksimielisesti	3	unanimously
100	ysläm**ä**llisesti	ystäwällisesti	2	kindly
100	todistuskappaleilla	todistuskappaleilla	0	with the pieces of evidence
100	peltikattovernissaa	peltikattovernissaa	0	tin roof varnish
100	**m**astaaminen	wastaaminen	1	answering
100	**l**y**f**ymy**l**feen	kysymykseen	3	into the question
100	k**n**stann**n**ksella	kustannuksella	2	at the expense of
100	glasgo**m**ista	glasgowista	1	from glasgow
100	anni**s**le**l**uosa**l**eyhtiön	anniskeluosakeyhtiön	2	of the licensed limited liability company
100	a**m**ioliitoista	awioliitoista	1	of marriages

Similarity diagram of Fraktur letter shapes (a self-organizing map)



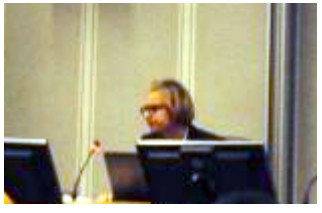
Research direction

Socio-Historical Text Mining of Newspaper Collections

Areas of analysis

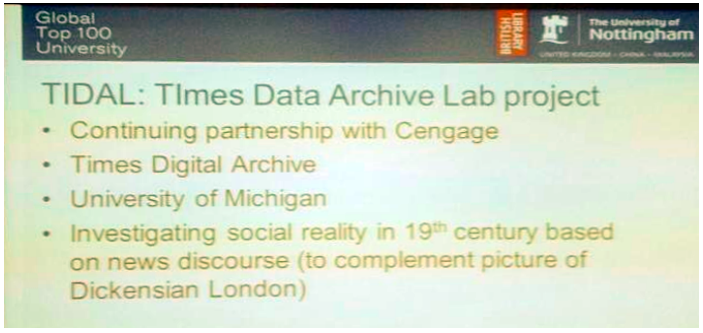
- Named entity recognition (people, organizations, places, events)
- Time series analysis
- Social network analysis
- Topic modeling

cf. Virginie Fortun's presentation



Areas of analysis

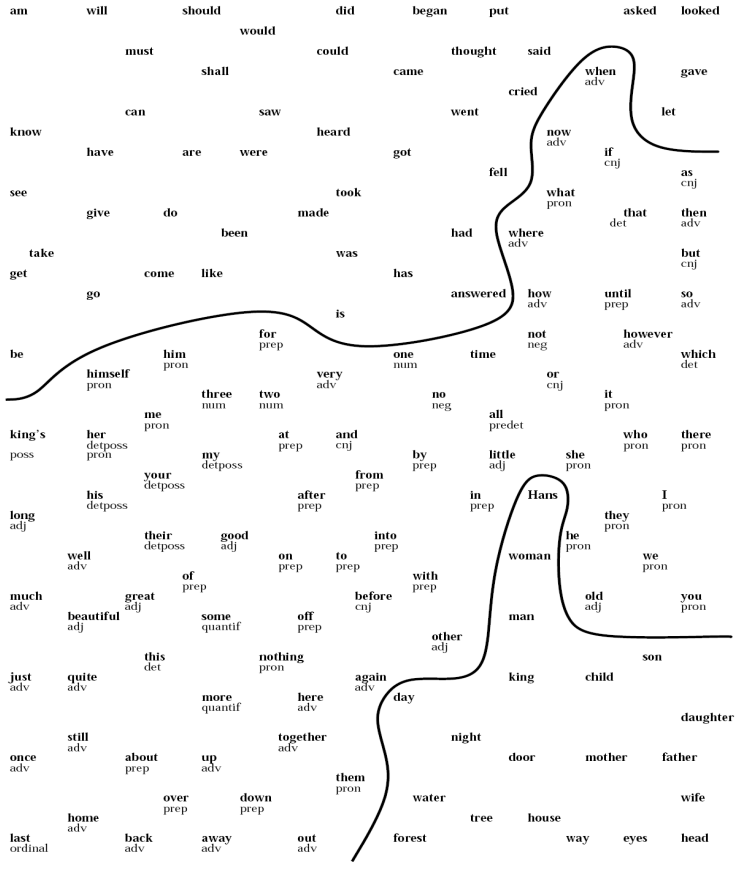
- Multidimensional sentiment analysis
- Analysis of social and historical context
- Intercultural and multilingual analysis
- Analysis of point of view
- Analysis of subjective understanding

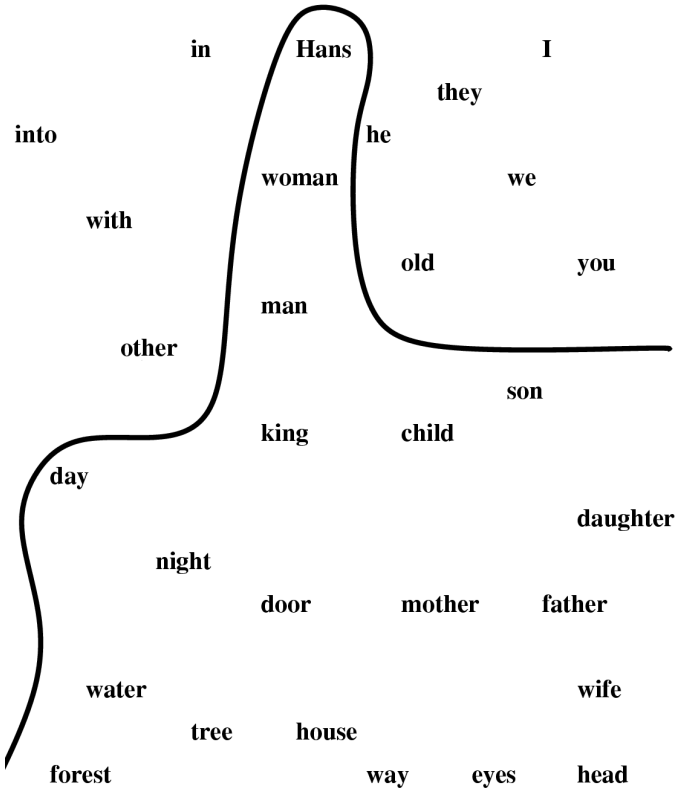


Stella Wisdom & Neil Smyth

Earlier related results

Learning meaning from context:
Maps of words in Grimm fairy tales





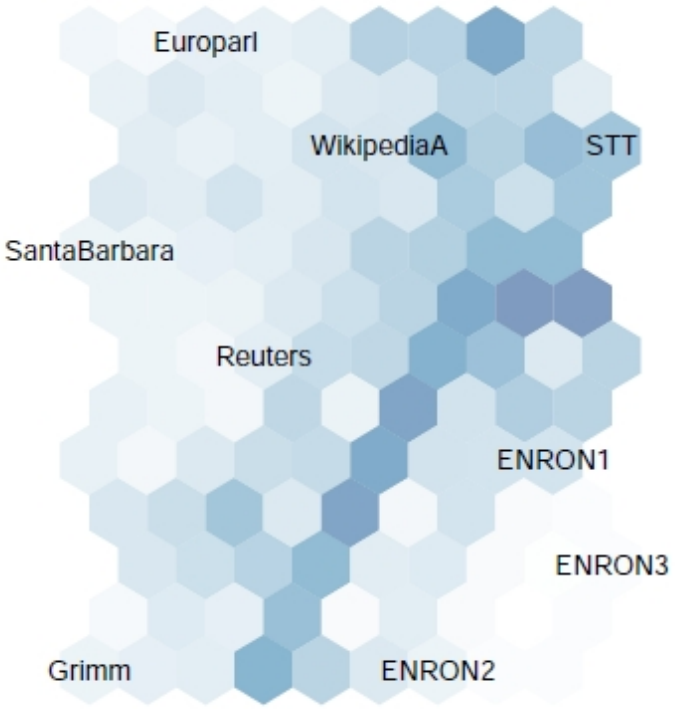
Honkela, Pulkki & Kohonen 1995

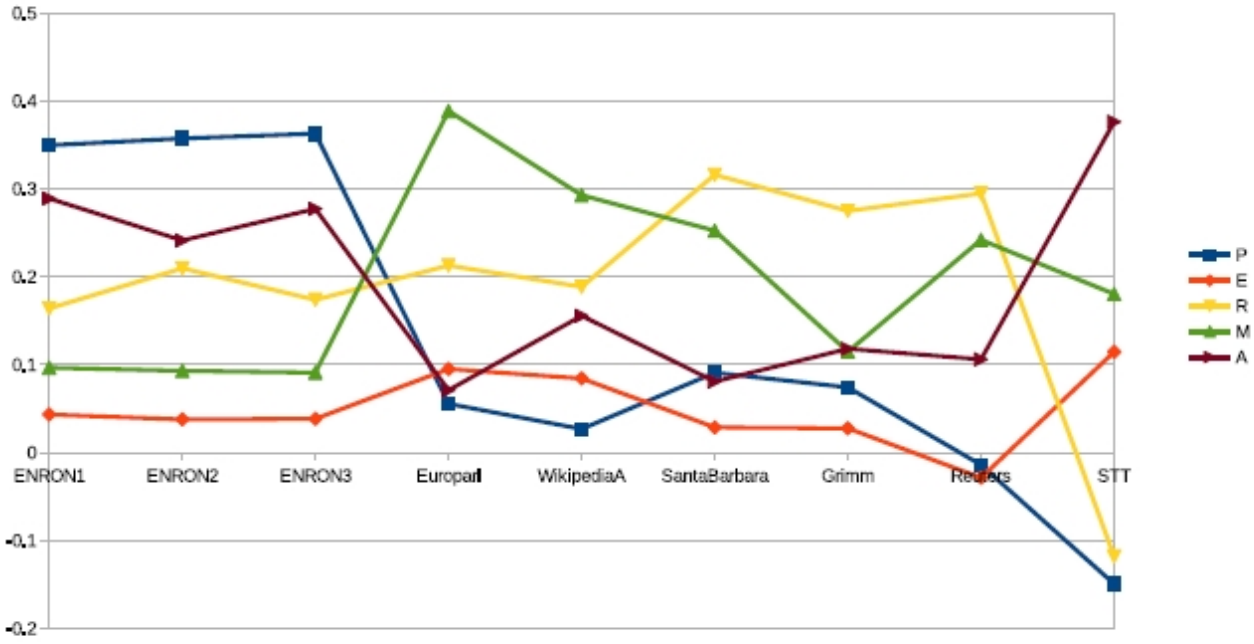
Multidimensional sentiment using the PERMA model

- Seligman and his colleagues has developed the PERMA model that addresses different aspects of wellbeing.
- The model includes five components related to subjective well-being:
  - Positive emotion (P),
  - Engagement (E),
  - Relationships (R),
  - Meaning (M) and
  - Achievement (A)

Honkela, Korhonen, Lagus & Saarinen 2014

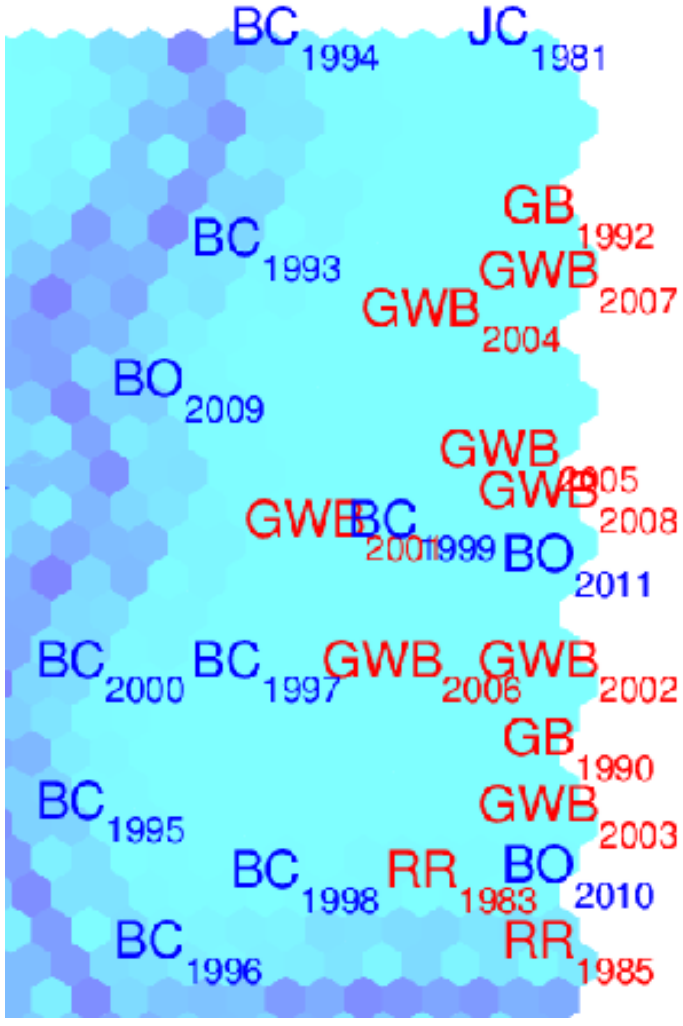
PERMA profiles of different corpora





Honkela, Korhonen, Lagus & Saarinen 2014

Analysis of the subjective meaning: word 'health'



Analysis of the State of the Union Adresses

JC	Jimmy Carter
RR	Ronald Reagan
GB	George Bush
BC	Bill Clinton
GWB	George W. Bush
BO	Barack Obama

Timo Honkela, Juha Raitio, Krista Lagus, Ilari T. Nieminen, Nina Honkela, and Mika Pantzar: **Subjects on objects in contexts: Using GICA method to quantify epistemological subjectivity** (IJCNN 2012)

Socio-Historical Text Mining of Newspaper Collections

A call for interdisciplinary international collaboration

Libraries, researchers within journalism, corpus linguistics, history, sociology, political science, psychology, computer science, machine learning, etc.

Merci!
Danke schön!
Grazie!
Multumesc!
¡Gracias!
Thank you!
Kiitos!
Tack!
謝謝！
Σας ευχαριστούμε!