Kielellisten merkkijonojen käsittelyn erityispiirteistä

Timo Honkela

Teknillinen korkeakoulu
Tietojenkäsittelytieteen laitos
Adaptiivisen informatiikan tutkimusyksikkö

Puhujasta

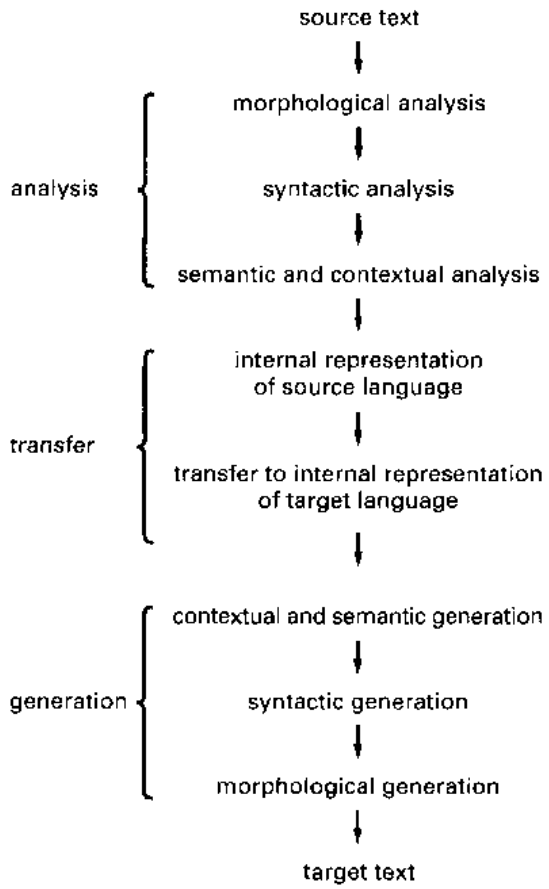
- FT (TKK 1998), dosentti (TKK, HY, TaiK)
- Kielikoneprojektin tutkija 1985-89
- VTT:n tutkija 1990-95
- TKK:lla 1995-1998, 2003-
- TaiK/Media Lab: 1998-2000, Gurusoft Oy 2000-02
- Johtava tutkija, vetää kognitiivisten järjestelmien tutkimusryhmää

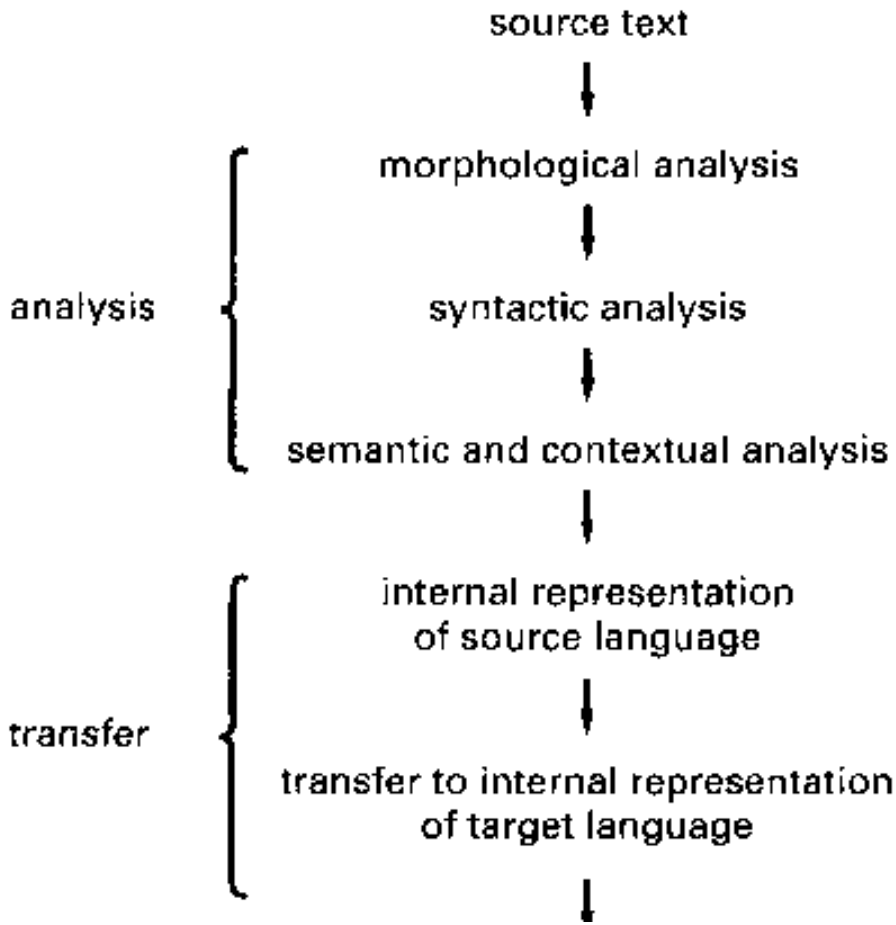

i n k o r e k ä p j ä i v o m
e o i k ä i ä p k v r j n o m
i r ä o o i e ä p k m k j v n
e k o i r j i m v o p n ä k ä
o n i ä o j k i k v m r e p ä
e o p m k i n v i ä r o j ä k
m k v k e i j o p o i ä n ä r
o e ä k o i ä n i p k r m j v
r v o ä e i o k j m ä n i k p
m j ä i k p k i ä n v e o r o

Vertailua

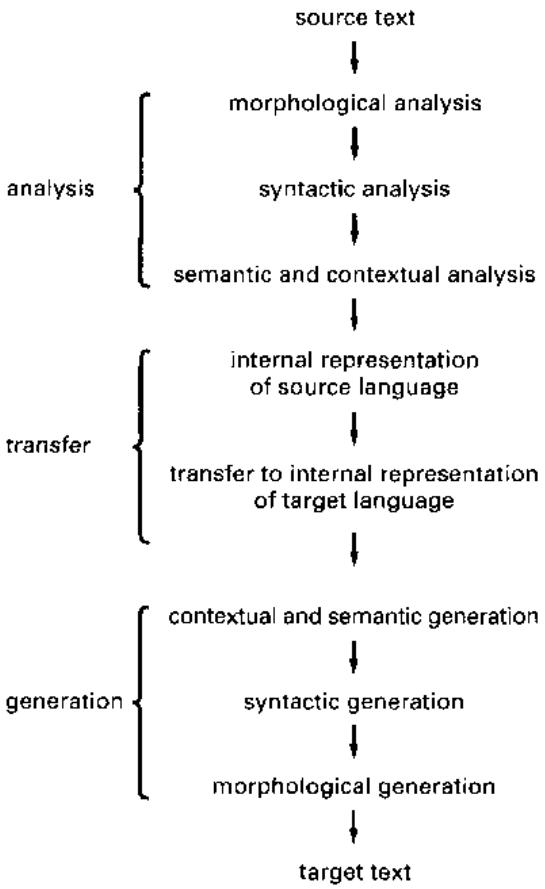
- KIELITEKNOLOGIA
  - puhutun ja kirjoitetun kielen analysointi ja tuottaminen
  - foneettisen, leksikaalisen, morfologisen, syntaktisen, semanttisen ja pragmaattisen tason prosessointi
  - sovellusalueita: tiedonhaku, käyttöliittymät, oikoluku, konekääntäminen
- MERKKIJONOMENETELMÄT
  - hahmonsovitus: merkkijonohahmon etsiminen tekstistä
  - merkkijonomuotoisen tiedon tiivistys
  - merkkijonojen manipulointi
  - sovellusalueita: tiedonhaku, tiedon tallennus

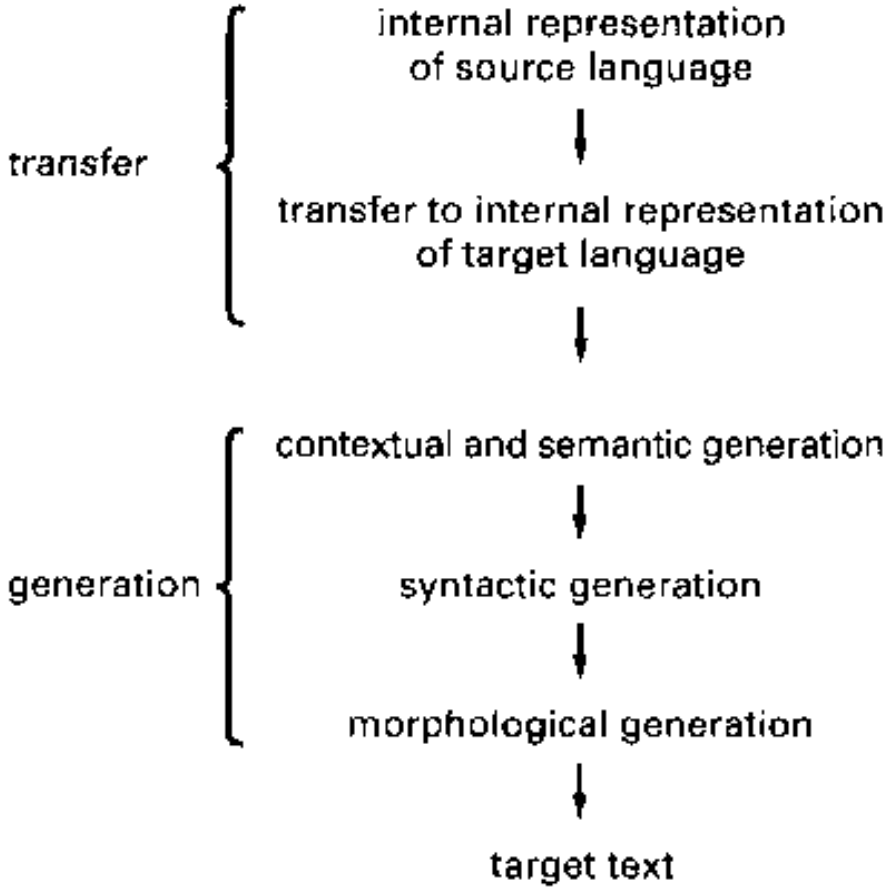
Esimerkki: Perinteinen konekäännösprosessi





Esimerkki: Perinteinen konekäännösprosessi





Oppiva ja hahmopohjainen kieliteknologia

- Laadullinen edistys:
  - Käsitellään kieltä sosiaalisena ja hahmopohjaisena ilmiönä
  - Kehitetään välineitä, joilla pyritään erityisesti ymmärtämisen tukeen
- Määrällinen tehokkuus:
  - Käsin tehtävien resurssien sijasta voidaan laittaa kone oppimaan: tuotetaan kieliteknologisia resursseja täysin tai osin automaattisesti

Esimerkkitutkimuksia

- Kielikoneprojekti
- Morfessor: sanojen segmentoinnin ei-ohjattu oppiminen
- WordICA: sanapiirteiden emergenssi
- Suomen Akatemialle osoitettujen hakemuksien analysointi; mm. termien tilastollinen irrotus
- Tilastollinen konekäännös
- Intersubjektiivisuus kommunikaatiossa


inkorekäpjäivom
eoikäiäpkvrjnom
iräooieäpkmkjvn
ekoirjimvopnäkä
oniäojkmkvmrepä
eopmkinviärojäk
mkvkeijcpoiänär
oeäkoiänipkrmjv
rvoäeiokjmänikp
mjäikpkiänveoro

Kielikone

- Käyttöliittymä, joka ymmärtää sille kirjoitettuja kysymyksiä ja komentoja
- Sitran rahoittama projekti 1980-luvulla

[perinteisiä kalvoja 20 vuoden takaa]


inkorekäpjäivom
eoikäiäpkvrjnom
iräooieäpkmkjvn
ekoirjimvopnäkä
oniäojkikmrepä
eopmkjrviärojäk
mkvkeijopoiänär
oeäkoiänipkrmjv
rvoäeiokjmänikp
mjäikpkiänveoro

Morfeemien automaattinen oppiminen tekstidatasta

- Mathias Creutzin ja Krista Laguksen kehittämä Morfessor-menetelmä
- Menetelmä rakentaa tilastollisin perustein mallin sanojen jakaantumisesta osiinsa perustuen annettuun opetusdataan

Morfessor (Tietokone-lehti)

perjantaina 3.6.2005 klo 9:19

Sanat paloiksi tilastollisesti

Morfessor oppii kieltä maistelemalla

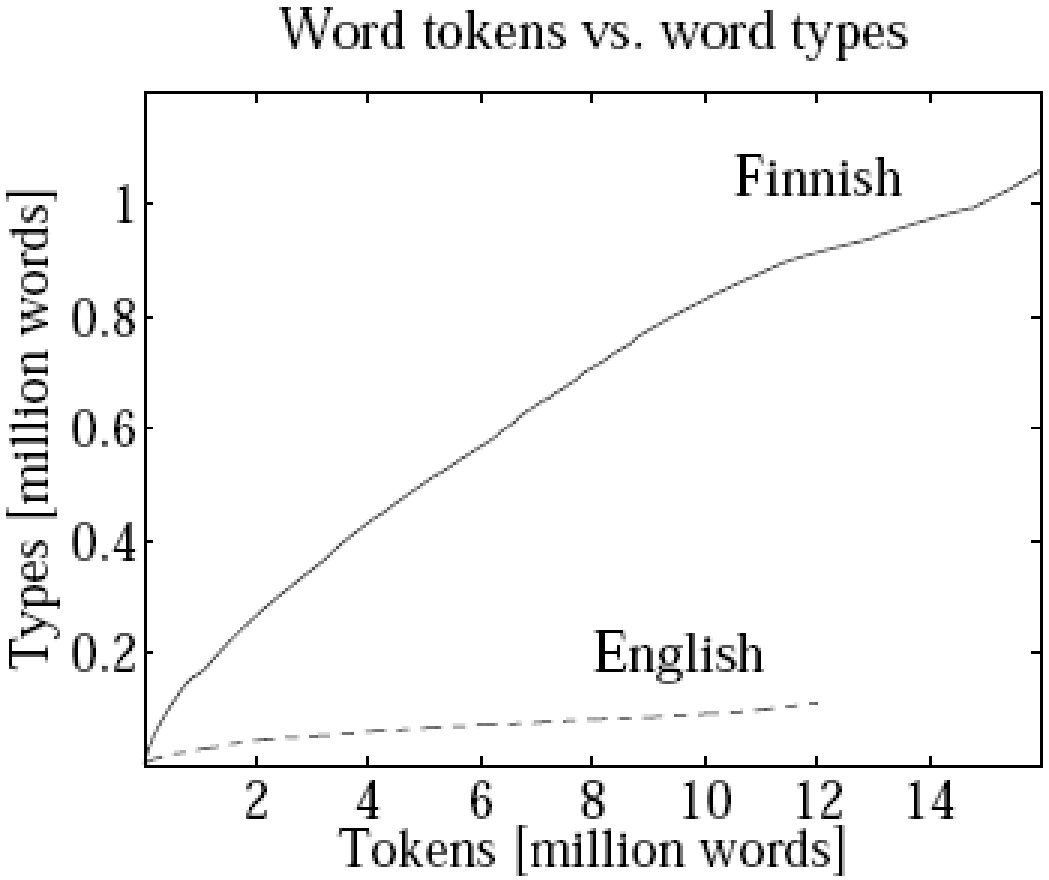
Teknillisen korkeakoulun Informaatiotekniikan laboratoriossa on kehitetty menetelmä sekä Morfessor-tietokoneohjelma, joka oppii sille annetusta tekstiaineistosta automaattisesti analysoimaan sanojen todennäköisiä rakennuspalikoita. Tästä voi olla hyötyä esimerkiksi parempien tulosten saamisessa netin hakukoneita käytettäessä.

Morfessor (Tietokone-lehti)

Internet-hakukoneet eivät yleensä hallitse vaikkapa suomen kielen taivutusmuotoja, saati yhdyssanoja. Morfessoriin ei ole koodattu minkään kielen kielioppia, vaan käytetty menetelmä on oppiva ja täysin tilastollinen. Ohjelmaa on toistaiseksi sovellettu menestyksekkäästi suomen, englannin ja turkin kieliin.

Esimerkiksi suomenkielisestä aineistoista Morfessor-ohjelma on oppinut, että suomen kielessä "ssa" on todennäköinen pääte, sillä se esiintyy yleisesti monissa sananmuodoissa ja pääasiallisesti sanan loppupäässä (esim. "Sisilia + ssa", "auto + ssa + han"). Kohdatessaan uuden sanan "Kaledoniassa" Morfessor-ohjelma osaa siksi päätellä, että kyse luultavasti on paloista "Kaledonia + ssa".

Different kinds of languages regarding morphology



(Creutz, 2006)

Morfessor

- In the following, the basic principles behind the Morfessor method (Creutz & Lagus, 2002, etc.) for unsupervised morph segmentation are outlined (mostly based on Creutz, 2006)

Morphemes as smallest meaningful units

Probabilistic modeling and maximum likelihood (ML) optimization

- Suppose that there is a family of simplistic models, each of which consists of a lexicon of morphs.
- Lexicons emerge from a stochastic process, where letters are chosen by random
- The alphabet consists of the 26 lower-case letters in the English alphabet and of a morph separator (space)
- For simplicity, all letters (including space) have an equal probability

(Creutz, 2006)

Probabilistic modeling and maximum likelihood (ML) optimization

The lexicon is a morph collection in the sense that each space-delimited string is a morph. The probability of the lexicon depends on its size (the number of free parameters in it). Some possible lexicons and their probabilities are (Creutz, 2006):

Lexicon 1 "a␣c␣e␣g␣i␣j␣l␣m␣n␣o␣p␣r␣t␣u␣␣", $P(Lexicon\ 1) = (\frac{1}{27})^{29}$

Lexicon 2 "apple␣juice␣lemon␣orange␣tree␣␣", $P(Lexicon\ 2) = (\frac{1}{27})^{31}$

Lexicon 3 "apple␣applejuice␣appletree␣juice␣lemon␣lemontree␣ orange␣orangejuice␣␣", $P(Lexicon\ 3) = (\frac{1}{27})^{69}$

Probabilistic modeling and maximum likelihood (ML) optimization

Lexicon 1 "a␣c␣e␣g␣i␣j␣l␣m␣n␣o␣p␣r␣t␣u␣␣", $P(Lexicon\ 1) = (\frac{1}{27})^{29}$

Lexicon 2 "apple␣juice␣lemon␣orange␣tree␣␣", $P(Lexicon\ 2) = (\frac{1}{27})^{31}$

Lexicon 3 "apple␣applejuice␣appletree␣juice␣lemon␣lemontree␣ orange␣orangejuice␣␣", $P(Lexicon\ 3) = (\frac{1}{27})^{69}$

The larger the lexicon is, the higher the number of possible configurations is, and the smaller the probability is, that the lexicon actually looks exactly as it happens to do. Therefore, Lexicon 1, which is smallest, is the most probable model, and Lexicon 3 which is largest, is the least probable model, a priori. (Creutz, 2006)

Probabilistic modeling and maximum likelihood (ML) optimization

Given these three models (lexicons), it is possible to compute **probabilities for** a small **data set**, namely the word list: “apple, orange, lemon, juice, applejuice, orangejuice, appletree, lemontree”.

We continue to keep the task simple, and assume that all morphs in a lexicon are equally likely to occur. (Creutz, 2006)

Likelihood with respect to the data

Data | Lexicon 1 "`a p p l e # o r a n g e # l e m o n # j u i c e # a p p l e j u i c e # o r a n g e j u i c e # a p p l e t r e e # l e m o n t r e e # #`" The sequence consists of 69 morphs and its probability conditioned on *Lexicon 1* is: $P(Data \mid Lexicon\ 1) = (\frac{1}{14+1})^{69} \approx 7.1 \cdot 10^{-82}$.

Data | Lexicon 2 "`apple # orange # lemon # juice # apple juice # orange juice # apple tree # lemon tree # #`". The sequence consist of 21 morphs, and $P(Data \mid Lexicon\ 2) = (\frac{1}{5+1})^{21} \approx 4.6 \cdot 10^{-17}$.

Data | Lexicon 3 "`apple # orange # lemon # juice # applejuice # orangejuice # appletree # lemontree # #`". The sequence consists of 17 morphs, and $P(Data \mid Lexicon\ 3) = (\frac{1}{8+1})^{17} \approx 6.0 \cdot 10^{-17}$.

Probabilistic modeling and maximum likelihood (ML) optimization

When the lexicons were compared according to their prior probabilities the following ranking was obtained: P(Lexicon 1) > P(Lexicon 2) > P(Lexicon 3)

If the lexicons are compared according to their likelihood with respect to the data, the opposite ranking ensues: P(Data | Lexicon 3) > P(Data | Lexicon 2) > P(Data | Lexicon 1)

Selecting the model that assigns the highest probability to the data is called **maximum likelihood** (ML) optimization.

Model selection

The proposed model selection procedure is based on maximizing the posterior probability of the model, P(Lexicon X |Data).

The posterior can be rewritten using Bayes' rule:

$$P(Lexicon\ X \mid Data) = \frac{P(Lexicon\ X) \cdot P(Data \mid Lexicon\ X)}{P(Data)}.$$

(Creutz, 2006)

Model selection

If the intention is to compare different models on the very same data set, the probability of the data is a constant that does not affect the result of the comparison.

Therefore, the probability of the data can be ignored, and we obtain:

$$P(Lexicon\ X \mid Data) \propto P(Lexicon\ X) \cdot P(Data \mid Lexicon\ X).$$

(Creutz, 2006)

Minimum description length (MDL)
(Rissanen, 1978, etc.)

Regardless of version, the fundamental idea of MDL is to view data compression as the basis. Any regularity in data can be used for compressing the data. Therefore, the more compact description one can obtain for a data set, the more regularity one has discovered and the more one has learned about the data.

Another consistent theme in the MDL methodology is the rejection of Bayesian prior probabilities. The Bayesian approach leaves room for subjectivity.

Recursive segmentation

Recursive segmentation. The search for the optimal morph segmentation proceeds recursively. First, the word as a whole is considered to be a morph and added to the codebook. Next, every possible split of the word into two parts is evaluated.

The algorithm selects the split (or no split) that yields the minimum total cost. In case of no split, the processing of the word is finished and the next word is read from input. Otherwise, the search for a split is performed recursively on the two segments. The order of splits can be represented as a binary tree for each word, where the leafs represent the morphs making up the word, and the tree structure describes the ordering of the splits.

(Creutz, 2006)

Recursive segmentation and MDL cost

$$
\begin{aligned}
C &= \text{Cost(Source text)} + \text{Cost(Codebook)} \\
&= \sum_{\text{tokens}} -\log p(m_i) + \sum_{\text{types}} k * l(m_j)
\end{aligned}
$$

(Creutz, 2006)

Morfessor Categories-ML and HMM

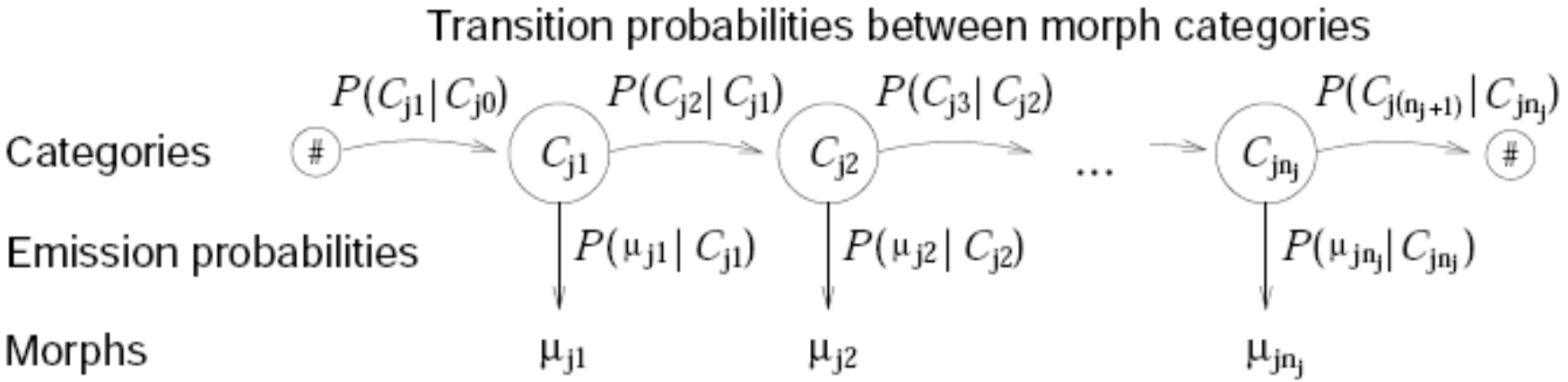
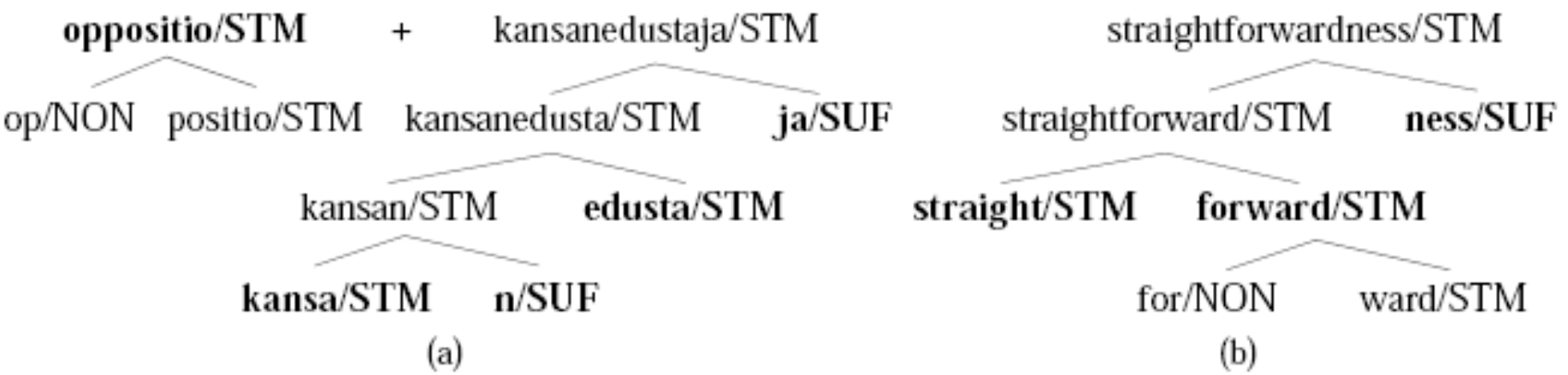


Figure 3.3: The HMM model of a word according to Equation 3.8. The word consists of a sequence of morphs which are emitted from latent categories. For instance, a possible category sequence for the English word "unavailable" would be "prefix + stem + suffix" and the corresponding morphs would be "un + avail + able".

(Creutz, 2006)

Example



(Creutz, 2006)

Examples in Finnish

Baseline-Length	Categories-MAP	Hutmegs Gold Standard
aarre kammioissa	[**aarre kammio**] *issa*	**aarre kammio** *i ssa*
aarre kammioon	[**aarre kammio**] *on*	**aarre kammio** *on*
bahama laiset	**bahama** *laiset*	**bahama** *laise t*
bahama saari en	**bahama** [**saari** *en*]	**bahama saar** *i en*
epä esteettis iksi	epä [[**esteet** *ti*] *s*] *iksi*	**epäesteett** *is i ksi*
epätasapaino inen	[epä [[tasa **paino**] *inen*]]	epätasa **painoinen**
haapa koskeen	[**haapa** [**koskee** *n*]]	**haapa koske** *en*
haapa koskella	[**haapa** [**koske** *lla*]]	**haapa koske** *lla*
ja n ille	**jani** *lle*	**jani** *lle*
jäädyttä ä kseen	[**jäädy ttää**] *kseen*	**jäädy** *ttä ä kse en*
ma clare n	**maclare** *n*	–
nais autoilija a	[nais [**autoili** *ja*]] *a*	**nais autoili** *ja a*
pää aiheesta	**pää** [**aihe** *esta*]	**pää aihee** *sta*
pää aiheista	[pää [**aihe** *ista*]]	**pää aihe** *i sta*
päähän	[pää **hän**]	**pää** *hän*
sano ttiin ko	[**sano** *ttiin*] *ko*	**sano** *tt i in ko*
työ tapaaminen	**työ** [**tapaa** *minen*]	**työ tapaa** *minen*
töhri misistä	**töhri** (*mis istä*)	**töhri** *mis i stä*
voi mmeko	[[**voi** *mme*] *ko*]	**voi** *mme ko*

(Creutz, 2006)

Examples in English

Baseline-Length	Categories-MAP	Hutmegs Gold Standard
accomplish es	[**accomplish** *es*]	**accomplish** *es*
accomplish ment	[**accomplish ment**]	**accomplish** *ment*
beautiful ly	[**beautiful** *ly*]	**beauti** *ful ly*
configu ration	[**configur ation**]	con **figur** *ation*
dis appoint	**disappoint**	dis **appoint**
express ive ness	[**expressive** *ness*]	**express** *ive ness*
flu s ter ed	[**fluster** *ed*]	**fluster** *ed*
insur e	**insure**	in **sure**
insur ed	[**insur** *ed*]	in **sur** *ed*
insur es	[**insure** *s*]	in **sure** *s*
insur ing	[**insur** *ing*]	in **sur** *ing*
long fellow 's	[[**long fellow**] *'s*]	–
master piece s	[[**master piece**] *s*]	**master piece** *s*
micro organism s	[**micro** [**organism** *s*]]	micro **organ** *ism s*
photograph ers	[[[**photo graph**] *er*] *s*]	photo **graph** *er s*
re side d	**resided**	**resid** *ed*
re side s	[**reside** *s*]	**reside** *s*
re s id ing	[re **siding**]	**resid** *ing*
un expect ed ly	[[un [**expect** *ed*]] *ly*]	–

(Creutz, 2006)


inkorekäpjäivom
eoikäiäpkvrjnom
iräooieäpkmkjvn
ekoirjimvopnäkä
oniäojkikvmrepä
eopmkinvärojäk
mkvkeiöpoiänär
oeäkoiänipkrmjv
rvoäeiokjmänikp
mjäikpkiänveoro
3

WordICA

- Menetelmä sanapiirteiden automaattiseen tuottamiseen riippumattomien komponenttien menetelmää (independent component analysis, ICA) käyttämällä (Honkela & Hyvärinen 2004, Väyrynen, Honkela, Hyvärinen 2005, jne.)

ICA-menetelmästä

The classic version of the ICA model can be expressed as

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{1}$$

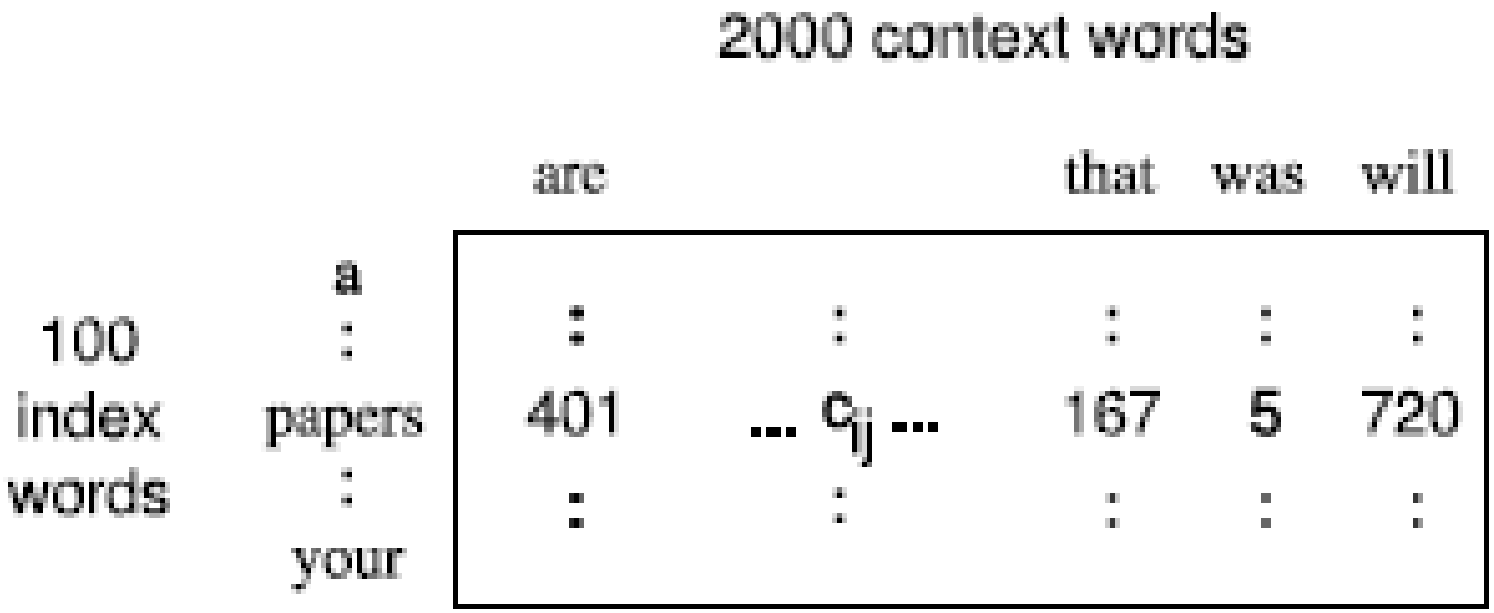
where $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ is the vector of observed random variables, the vector of the independent latent variables is denoted by $\mathbf{s} = (s_1, s_2, \ldots, s_n)^T$ (the "independent components"), and $\mathbf{A}$ is an unknown constant matrix, called the mixing matrix. If we denote the columns of matrix $\mathbf{A}$ by a_j the model can be written as

$$\mathbf{x} = \sum_{i=1}^{n} \mathbf{a}_i s_i \tag{2}$$

The goal in ICA is to learn the decomposition in Eq. (1) in an unsupervised manner. That is, we only observe $\mathbf{x}$ and want to estimate both $\mathbf{A}$ and s.

(Honkela & Hyvärinen, 2004) (Hyvärinen, Karhunen & Oja, 2001)

Word-context input matrix



Word ICA

6	7	8	9	10
a	the	neural	their	will
the	an	computational	our	can
and	and	cognitive	your	may
or	or	network	my	should
their	their	adaptive	learning	would
its	its	control	research	must
your	are	learning	processing	did
...	...	...	...	...

Fig. 12. The most representative words for the last five features (components), in the order of representativeness, top is highest.

(Honkela & Hyvärinen, 2004)

Word ICA: example

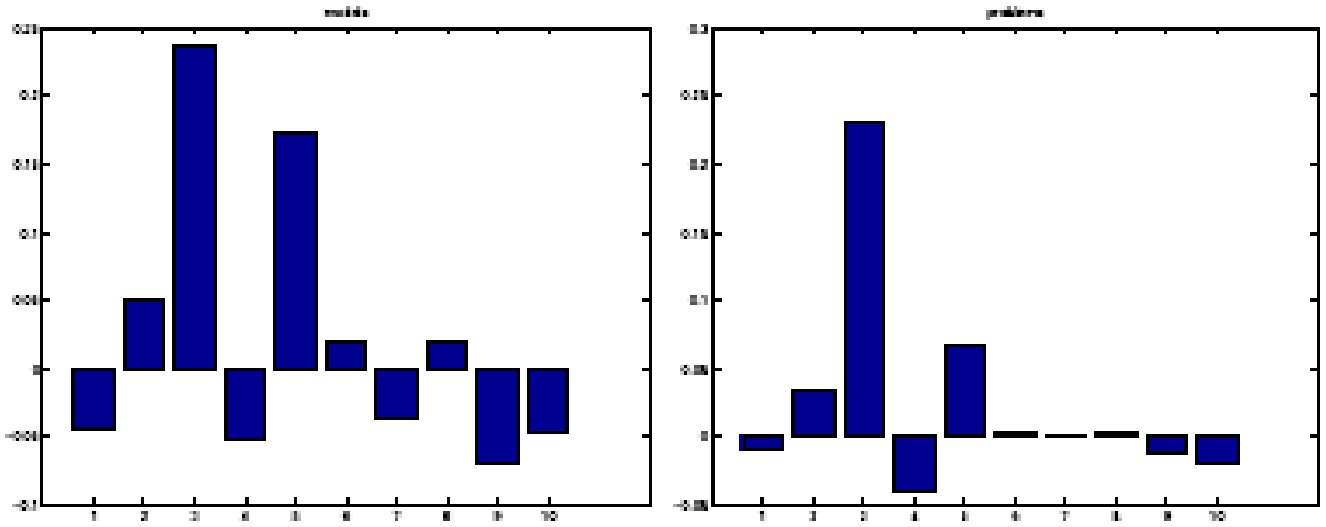


Fig. 4. ICA features for "models" and "problems".

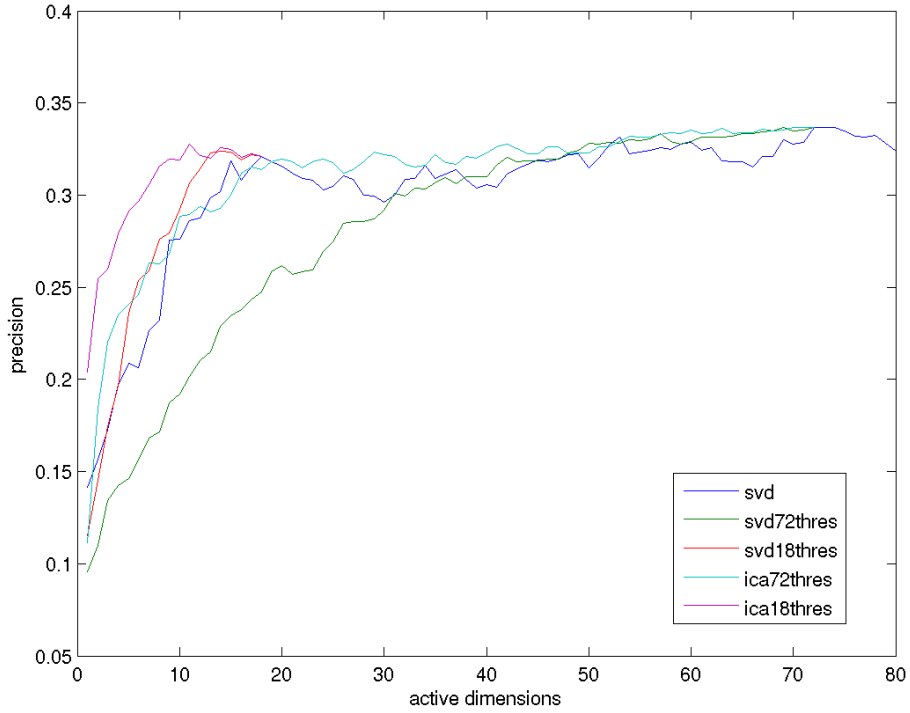
(Honkela & Hyvärinen, 2004)

Word ICA

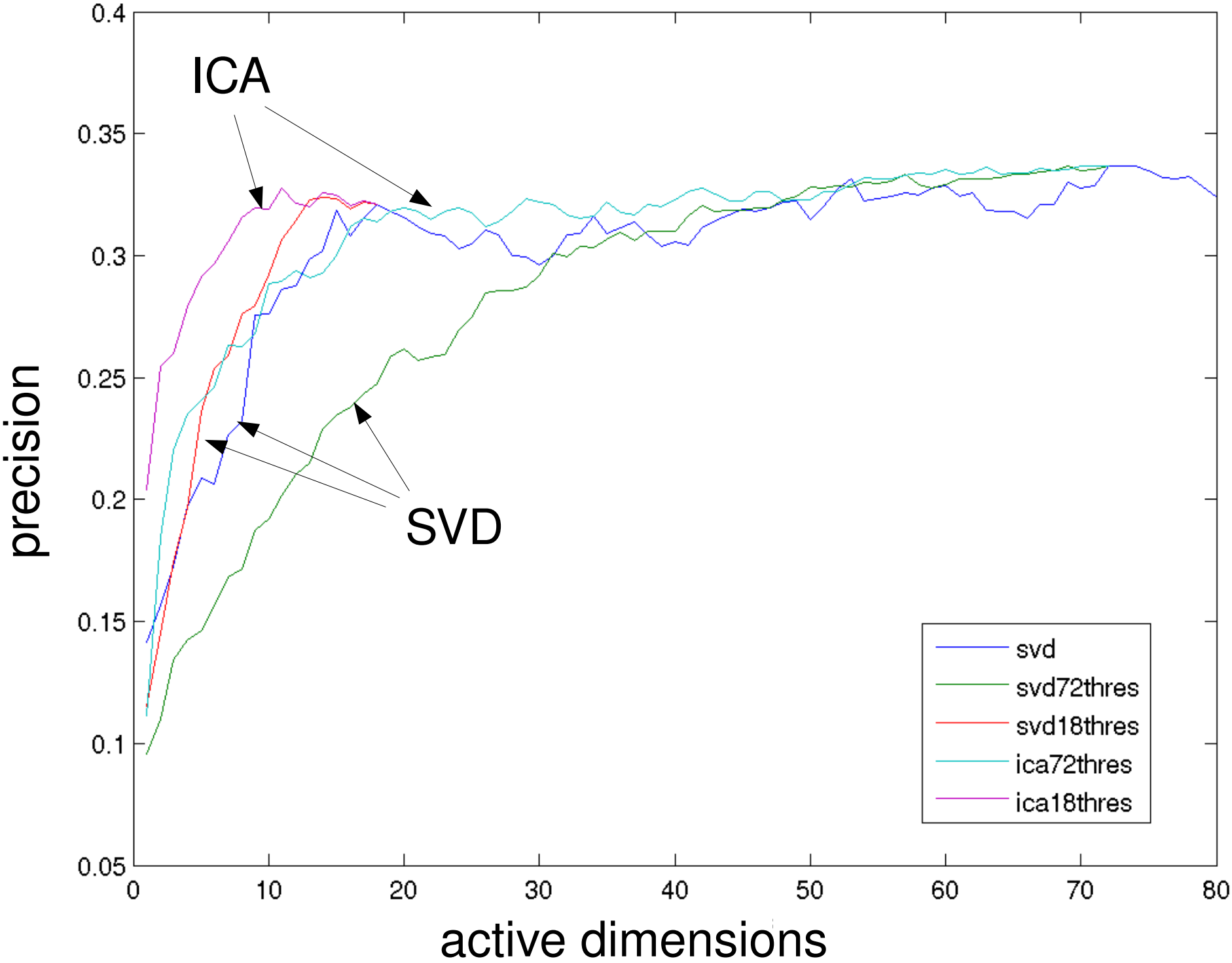
Comparison with SVD/LSA

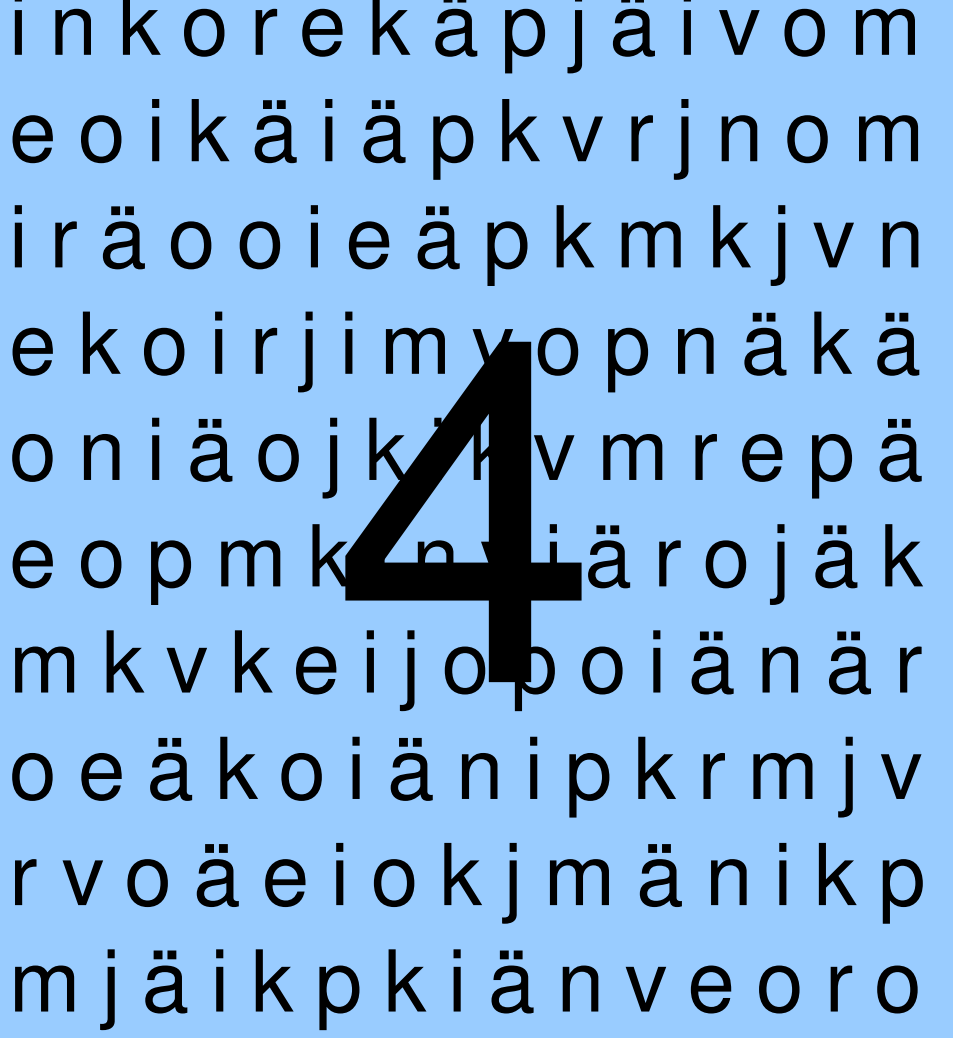
Effect of sparseness and meaningful emergent components

Data: TOEFL tests



(Väyrynen, Lindqvist, Honkela 2007)


ICA
SVD
precision
active dimensions
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
10
20
30
40
50
60
70
80
svd
svd72thres
svd18thres
ica72thres
ica18thres

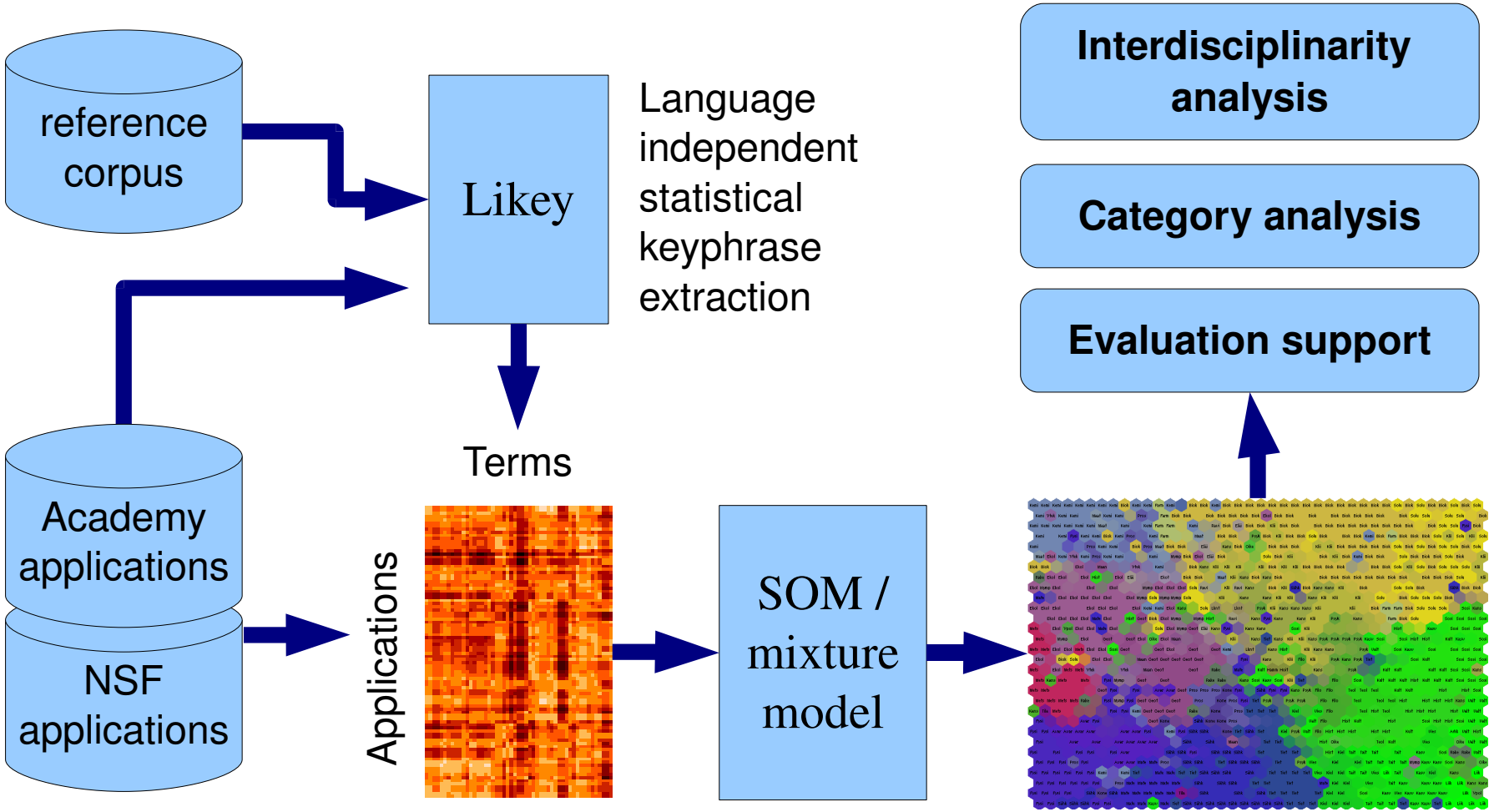

i n k o r e k ä p j ä i v o m
e o i k ä i ä p k v r j n o m
i r ä o o i e ä p k m k j v n
e k o i r j i m v o p n ä k ä
o n i ä o j k i k v m r e p ä
e o p m k n i ä r o j ä k
m k v k e i j o p o i ä n ä r
o e ä k o i ä n i p k r m j v
r v o ä e i o k j m ä n i k p
m j ä i k p k i ä n v e o r o
4

Suomen Akatemian hakemuksien analyysi

- Tarkoituksena oli selvittää, miten automaattisilla tekstinlouhinnan menetelmillä voitaisiin toteuttaa hakemusten sisältöanalyysi ja tutkia mm. hakemusten monitieteisyyttä

(Honkela ja Klami, 2007, julkaisumaton)

Menetelmäkaavio



Termien valintaprosessi

- Hakemustekstien esikäsittelyssä:
  - poimittiin automaattisesti englanninkieliset osuudet
  - poistettiin välimerkit ja muunnettiin tekstit pienaakkosille
- Merkitykselliset termit poimittiin teksteistä:
  - hyödyntämällä referenssikorpusta (EU-parlamenttitekstit)
  - soveltamalla entropiamittaa, joka kertoo termien erottelukyvystä
- Lopuksi termiehdokkaat käytiin käsin läpi ja valittiin 1200 termiä

Raakatermit eli yleisimmät sanat ja fraasit

the 1276847
of 1067918
and 817852
in 625330
to 357453
for 225307
is 205723
on 162509
research 157251
be 151475
with 136854
will 135992
as 122707
are 116508
by 113878
university 98003
...

of the 237053
in the 139990
university of 73841
will be 72631
to the 48573
on the 48522
and the 47784
for the 44311
the research 34639
at the 34019
the project 32963
et al 29326
from the 26554
department of 25568
with the 25087
to be 24870
...

university of helsinki 17784
of the research 13902
as well as 13638
the university of 12286
of the project 12140
academy of finland 11365
university of technology 9213
member of the 7932
in order to 7270
university of turku 7092
the academy of 6717
university of oulu 6673
...

the academy of finland 6490
at the university of 5994
helsinki university of technology 5118
...

Termien sijalukuvertailua: hakemukset versus Europarl

1.	the	1276847	
2.	of	1067918	
3.	and	817852	
4.	in	625330	
5.	to	357453	
6.	for	225307	
7.	is	205723	
8.	on	162509	
9.	**research**	157251	
10.	be	151475	
11.	with	136854	
12.	will	135992	
13.	as	122707	
14.	are	116508	
15.	by	113878	
16.	**university**	98003	
...			

1.	the	2023617
2.	of	945622
3.	to	883206
4.	and	717718
5.	in	611421
6.	that	473739
7.	a	445775
8.	is	445119
9.	**we**	305590
10.	for	296092
11.	**i**	290412
12.	this	286924
13.	on	274614
14.	it	251343
15.	be	246917
16.	are	197082
...		

Sijalukuvertailun avulla tehty valinta

university
finland
finnish
studies
helsinki
et
al
et al
department
study
project
research
academy
cell
thesis
cells
science
laboratory

institute
prof
e.g
professor
researcher
analysis
doctoral
methods
phd
turku
data
dr
students
journal
school
ja
design
technology

molecular
oulu
models
publications
gene
studied
theory
tampere
researchers
protein
chemistry
properties
student
using
forest
pp
processes
sciences

species
scientific
graduate
team
results
genes
proteins
physics
biology
model
collaboration
applications
engineering
related
interaction
novel
theoretical
experiments

experimental
techniques
research project
ph.d
factors
expression
materials
knowledge
jyväskylä
measurements
material
surface
program
structure
helsinki
university
activity
articles
physical

Entropiamitan mukainen järjestys

0.0299	3578 13	xxxxxxx
0.0661	3728 5	xxxx
0.1138	3898 11	xxxxxxxx xx
0.1491	2053 20	xxxxxxxx
0.1743	3215 18	xxxxxxxxx
0.2004	3702 24	algebras
0.2017	3699 27	xxxxx xxxxxxx
0.2218	3539 26	cellular neural
0.2450	2132 83	cmos
0.2789	3449 61	circuit design
0.2860	2701 71	vlsi
0.2917	3917 64	symposium on circuits
0.3493	3278 21	xxxxxxxx
0.3628	2334 86	splicing
0.3751	2518 91	on circuits
0.3846	3097 26	xxxxxxxx
0.3934	997 52	xxxxxxxxxx
0.4143	3631 28	photogrammetry
0.4212	3109 30	cnn
0.4362	3812 5	xxxxxxx
0.4377	2426 29	knip
0.4395	2625 31	p38
0.4445	3248 36	ion exchange
0.4449	2849 26	xxxx xxxxx
0.4717	1115 60	xxxxxxxx
0.4784	2868 58	on-chip
0.4797	1744 48	xxxxxx
0.4821	1848 38	xxxxxx
0.4933	3602 8	xxxxxxxxx
0.4942	2785 24	xxxxxx
0.5449	3736 62	cdma
0.5544	1638 86	social work
. . .		

Käsin tehty karsinta

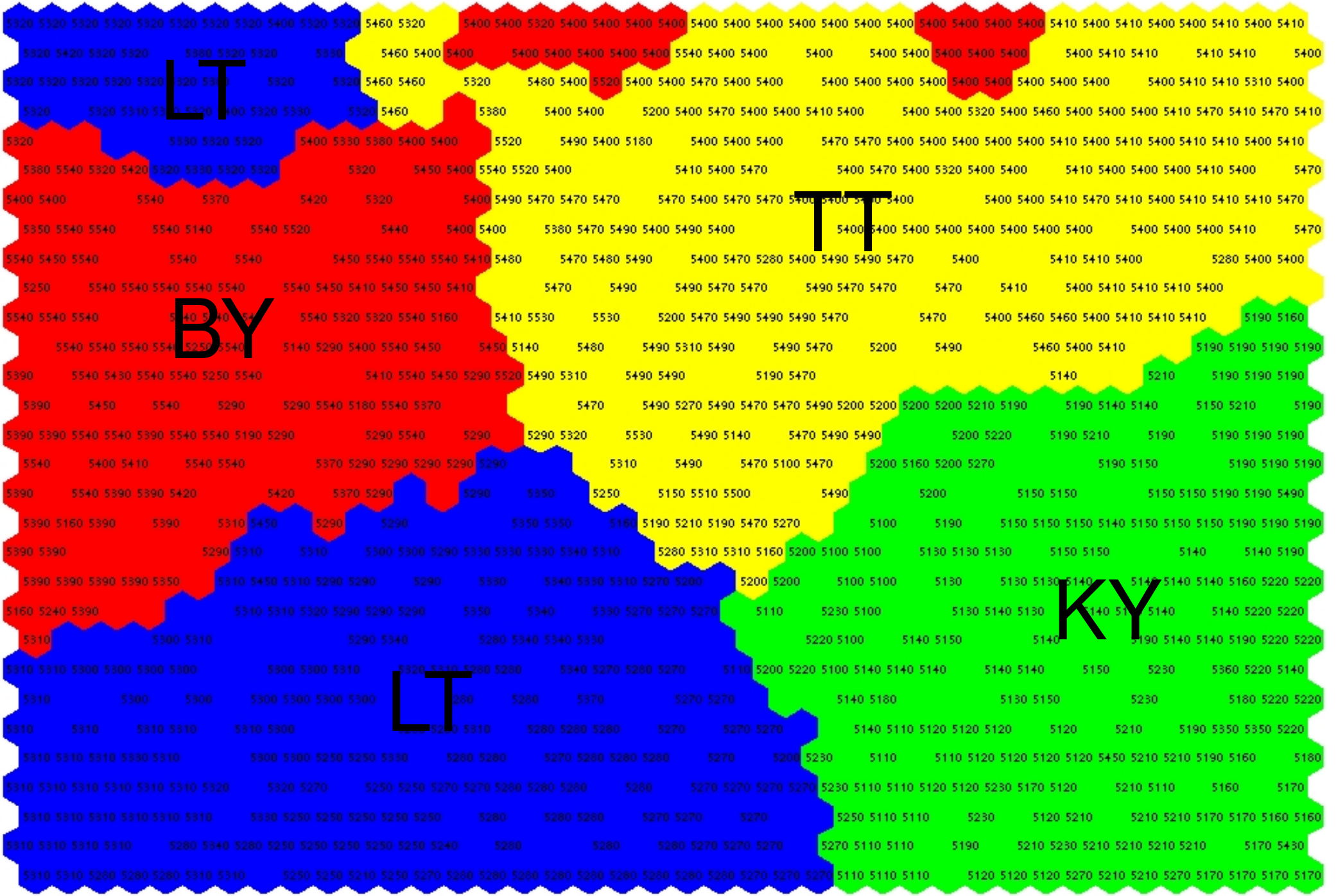
algebras
cellular neural
cmos
circuit design
vlsi
symposium on circuits
splicing

photogrammetry
cnn
knip
p38
ion exchange
on-chip
cdma
..

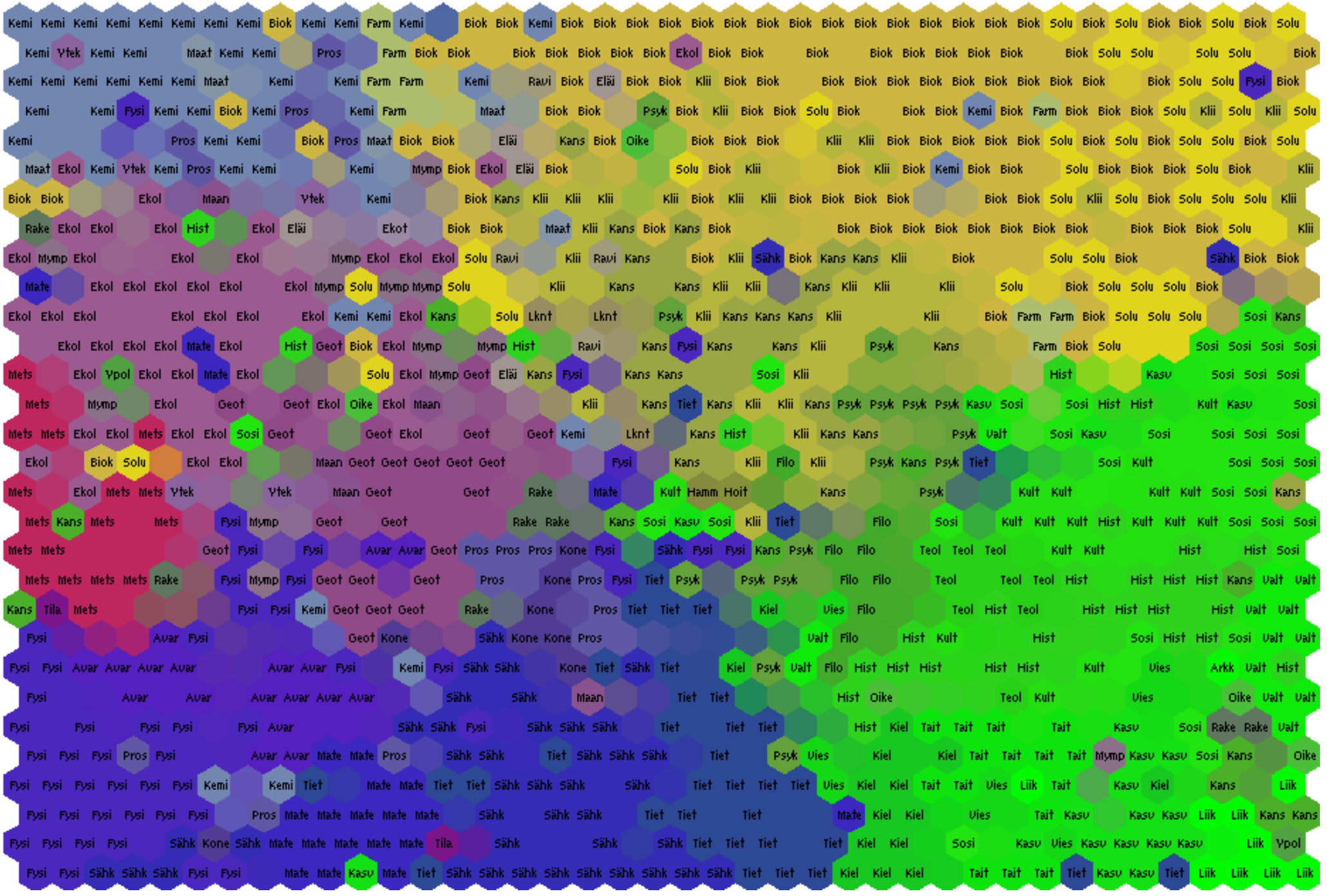
1331 → 1200 termiä (9,8% poistettu)

Karsintakriteerit:

- henkilönimet
- kaupunkien nimet
- organisaatiot
- fraasifragmentit (“researchers will”)
- numerot (“ii”, tms.)

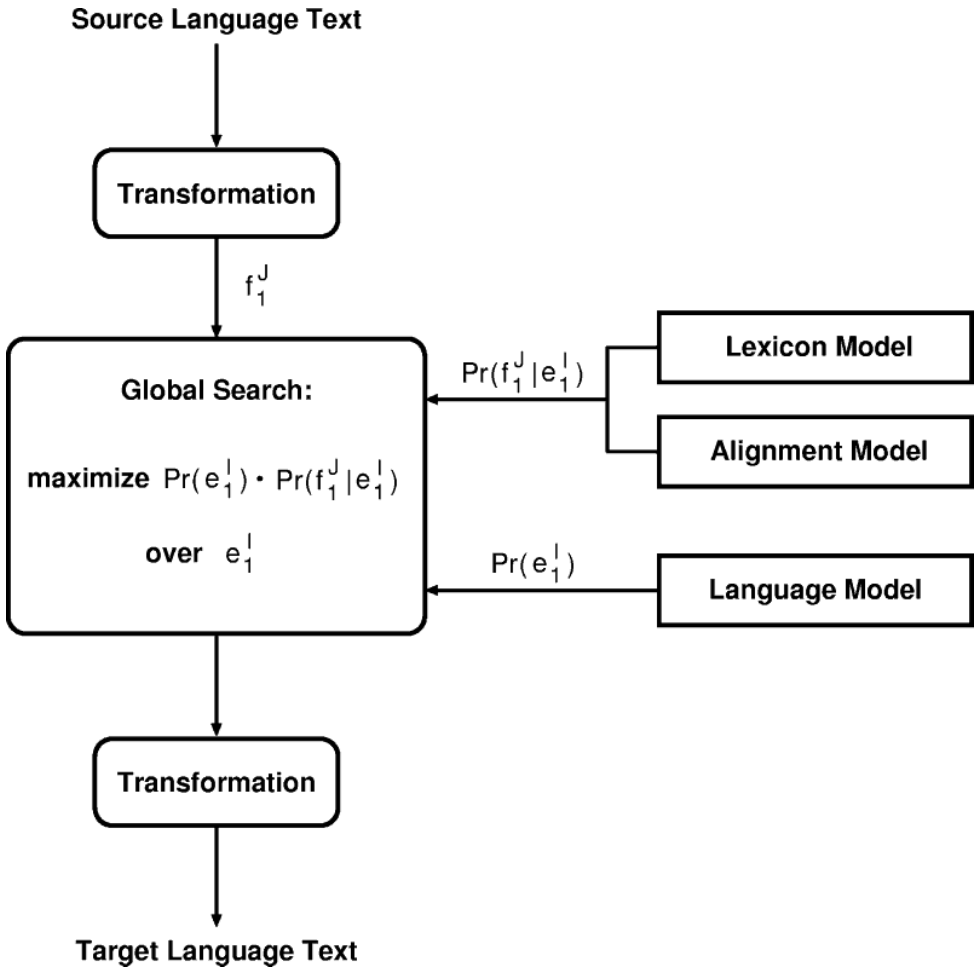


KY = kulttuurin ja yhteiskunnan tutkimus, LT = luonnontieteiden ja tekniikan tutkimus,


Kemi Kemi Kemi Kemi Kemi Kemi Kemi Kemi Biok Kemi Kemi Farm Kemi Biok Biok Kemi Biok Biok Biok Biok Biok Biok Biok Biok Biok Biok Biok Biok Biok Biok Biok Solu Biok Solu Biok Biok Solu Biok Solu
Kemi Ytek Kemi Kemi Maat Kemi Kemi Pros Farm Biok Biok Biok Biok Biok Biok Biok Biok Ekol Biok Biok Biok Biok Biok Biok Biok Biok Biok Biok Biok Solu Solu Solu Solu Biok
Kemi Kemi Kemi Kemi Kemi Kemi Maat Kemi Kemi Farm Farm Kemi Ravi Biok Eläi Biok Biok Klii Biok Biok Biok Biok Biok Biok Biok Biok Biok Biok Biok Biok Biok Solu Solu Fysi Biok
Kemi Kemi Fysi Kemi Kemi Biok Kemi Pros Kemi Farm Maat Biok Biok Psyk Biok Klii Biok Biok Solu Biok Biok Biok Kemi Biok Farm Biok Biok Biok Solu Klii Solu Klii Solu
Kemi Pros Kemi Kemi Biok Pros Maat Biok Biok Eläi Kans Biok Oike Biok Biok Biok Klii Klii Biok Biok Biok Biok Biok Solu Biok Solu Biok Solu Solu Biok Solu
Maat Ekol Kemi Ytek Kemi Pros Kemi Kemi Kemi Mymp Biok Ekol Eläi Biok Solu Biok Klii Biok Klii Biok Kemi Biok Biok Solu Biok Biok Biok Solu Biok Klii
Biok Biok Ekol Maan Ytek Kemi Biok Kans Klii Klii Klii Klii Biok Klii Klii Biok Biok Biok Biok Biok Biok Solu Klii Solu Biok Solu Solu Solu Klii
Rake Ekol Ekol Ekol Hist Ekol Eläi Ekot Biok Biok Maat Klii Kans Biok Kans Biok Biok Biok Biok Biok Biok Biok Biok Biok Biok Biok Biok Biok Biok Biok Solu Klii
Ekol Mymp Ekol Ekol Ekol Mymp Ekol Ekol Ekol Solu Ravi Klii Ravi Kans Biok Klii Sähk Biok Kans Kans Klii Biok Solu Solu Biok Sähk Biok Biok
Mate Ekol Ekol Ekol Ekol Ekol Ekol Mymp Solu Mymp Mymp Solu Klii Kans Kans Klii Klii Kans Klii Klii Klii Solu Biok Solu Solu Solu Biok
Ekol Ekol Ekol Ekol Ekol Ekol Ekol Kemi Kemi Ekol Kans Solu Lknt Lknt Psyk Klii Kans Kans Kans Klii Klii Biok Farm Farm Biok Solu Solu Solu Sosi Kans
Ekol Ekol Ekol Ekol Mate Ekol Hist Geot Biok Ekol Mymp Mymp Hist Ravi Kans Fysi Kans Kans Klii Psyk Kans Farm Biok Solu Sosi Sosi Sosi Sosi
Mets Ekol Ypol Ekol Ekol Mate Ekol Solu Ekol Mymp Geot Eläi Kans Fysi Kans Kans Sosi Klii Hist Kasv Sosi Sosi Sosi
Mets Mymp Ekol Geot Geot Ekol Oike Ekol Maan Klii Kans Tiet Kans Klii Klii Kans Psyk Psyk Psyk Psyk Kasv Sosi Sosi Hist Hist Kult Kasv Sosi
Mets Mets Ekol Ekol Mets Ekol Ekol Sosi Geot Geot Ekol Geot Geot Kemi Lknt Kans Hist Klii Kans Kans Psyk Valt Sosi Kasv Sosi Sosi Sosi Sosi
Ekol Biok Solu Ekol Ekol Maan Geot Geot Geot Geot Geot Fysi Kans Klii Filo Klii Psyk Kans Psyk Tiet Sosi Kult Sosi Sosi Sosi
Mets Ekol Mets Mets Ytek Ytek Maan Geot Geot Rake Mate Kult Hamm Hoit Kans Psyk Kult Kult Kult Kult Sosi Sosi Kans
Mets Kans Mets Mets Fysi Mymp Geot Geot Rake Rake Kans Sosi Kasv Sosi Klii Tiet Filo Sosi Kult Kult Kult Hist Kult Kult Kult Sosi Sosi Sosi
Mets Mets Geot Fysi Fysi Avar Avar Geot Pros Pros Pros Kone Fysi Sähk Fysi Fysi Kans Psyk Filo Filo Teol Teol Teol Kult Kult Hist Hist Sosi
Mets Mets Mets Mets Rake Fysi Mymp Fysi Geot Geot Geot Pros Kone Pros Fysi Tiet Psyk Psyk Psyk Filo Filo Teol Teol Teol Hist Hist Hist Hist Kans Valt Valt
Kans Tila Mets Fysi Fysi Kemi Geot Geot Geot Rake Kone Pros Tiet Tiet Tiet Kiel Vies Filo Teol Hist Teol Hist Hist Hist Hist Valt Valt
Fysi Avar Fysi Geot Kone Sähk Kone Kone Pros Valt Filo Hist Kult Hist Sosi Hist Hist Sosi Valt Valt
Fysi Fysi Avar Avar Avar Avar Avar Avar Fysi Kemi Fysi Sähk Sähk Kone Tiet Sähk Tiet Kiel Psyk Valt Filo Hist Hist Hist Hist Hist Kult Vies Arkk Valt Hist
Fysi Avar Avar Avar Avar Avar Avar Sähk Sähk Maan Tiet Tiet Hist Oike Teol Kult Vies Oike Valt Valt
Fysi Fysi Fysi Fysi Fysi Avar Sähk Sähk Fysi Sähk Sähk Sähk Tiet Tiet Tiet Hist Kiel Tait Tait Tait Tait Kasv Sosi Rake Rake Valt
Fysi Fysi Fysi Pros Fysi Avar Avar Mate Mate Pros Sähk Sähk Tiet Sähk Sähk Sähk Tiet Psyk Vies Kiel Kiel Tait Tait Tait Tait Mymp Kasv Kasv Sosi Kans Oike
Fysi Fysi Fysi Fysi Fysi Fysi Kemi Kemi Tiet Mate Mate Tiet Tiet Sähk Sähk Sähk Sähk Tiet Tiet Tiet Tiet Vies Kiel Kiel Tait Tait Vies Liik Tait Kasv Kiel Kans Liik
Fysi Fysi Fysi Fysi Fysi Fysi Pros Mate Mate Mate Mate Mate Sähk Sähk Sähk Tiet Tiet Tiet Mate Kiel Kiel Vies Tait Kasv Kasv Kasv Liik Liik Kans Kans
Fysi Fysi Fysi Fysi Sähk Kone Sähk Mate Mate Mate Mate Mate Tila Sähk Sähk Sähk Tiet Tiet Tiet Tiet Kiel Kiel Sosi Kasv Vies Kasv Kasv Kasv Kasv Liik Ypol
Fysi Fysi Sähk Sähk Sähk Fysi Fysi Mate Mate Kasv Mate Tiet Sähk Sähk Sähk Sähk Sähk Sähk Sähk Sähk Tiet Tiet Tiet Kiel Kiel Kiel Tait Tait Tait Tiet Kasv Kasv Tiet Liik Liik Liik Liik


inkorekäpjäivom
eoikäiäpkvrjnom
iräooieäpkmkjvn
ekoirjimvopnäkä
oniäojjikvmrepä
eopmkinvärojäk
mkvkeljpoiänär
oeäkoiänipkrmjv
rvoäeiokjmänikp
mjäikpkiänveoro
5

Tilastollinen konekäännös



Käännösmalli

Kielimalli

Suomen kielen haastavuus konekäännöksen kohdekielenä

Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. MT Summit 2005.

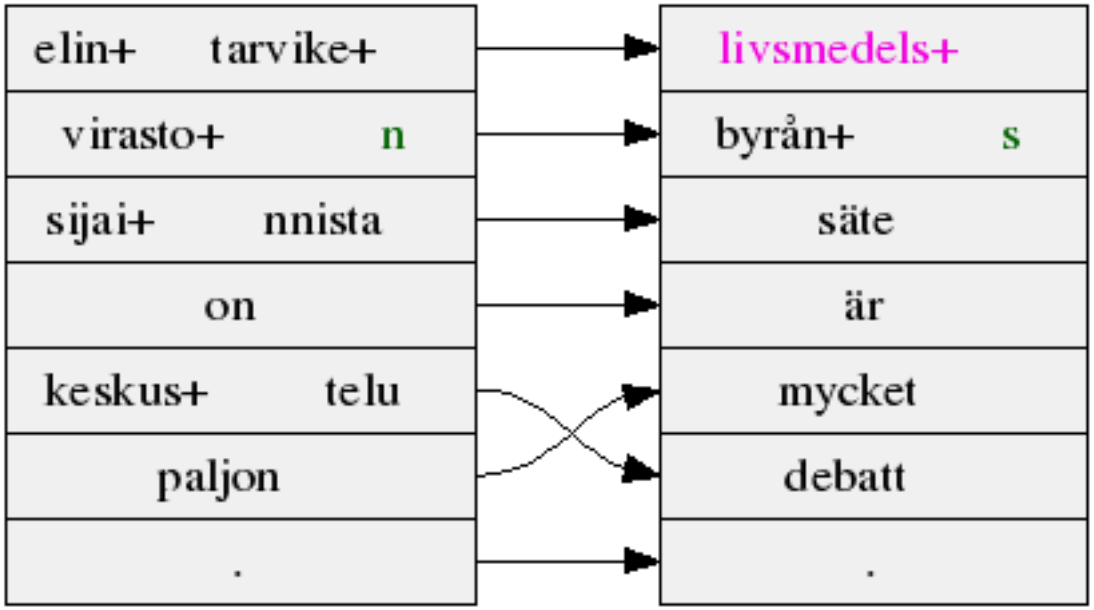
Source Language	Target Language										
	da	de	el	en	es	fr	fi	it	nl	pt	sv
da	-	18.4	21.1	28.5	26.4	28.7	14.2	22.2	21.4	24.3	28.3
de	22.3	-	20.7	25.3	25.4	27.7	11.8	21.3	23.4	23.2	20.5
el	22.7	17.4	-	27.2	**31.2**	**32.1**	11.4	26.8	20.0	27.6	21.2
en	25.2	17.6	23.2	-	**30.1**	**31.1**	13.0	25.3	21.0	27.1	24.8
es	24.1	18.2	28.3	**30.5**	-	**40.2**	12.5	**32.3**	21.4	**35.9**	23.9
fr	23.7	18.5	26.1	**30.0**	**38.4**	-	12.6	**32.4**	21.1	**35.3**	22.6
fi	20.0	14.5	18.2	21.8	21.1	22.4	-	18.3	17.0	19.1	18.8
it	21.4	16.9	24.8	27.8	**34.0**	**36.0**	11.0	-	20.0	**31.2**	20.2
nl	20.5	18.3	17.4	23.0	22.9	24.6	10.3	20.0	-	20.7	19.0
pt	23.2	18.2	26.4	**30.1**	**37.9**	**39.0**	11.9	**32.0**	20.2	-	21.9
sv	**30.3**	18.9	22.8	**30.2**	28.6	29.7	15.3	23.9	21.9	25.9	-

Table 2: BLEU scores for the 110 translation systems trained on the Europarl corpus

Morphologically rich languages

- Language may contain many word forms that are not present in the training corpus (“eläinlääkintäviranomaisetkin”)
- Automatically learned segmentation with Morfessor
  - morph = statistical morpheme (only segmentation, no lemmatisation)
  - eläin + **lääkintä** + **viranomais** + *et* + *kin*
- Application in SMT: all models are created using morphs rather than words

Elintarvikeviraston sijainnista on keskustelu paljon.



livsmedelsbyråns säte är mycket debatt .

(Virpioja, Väyrynen, Creutz, Sadeniemi, 2007)

Effect of Morfessor on BLEU

	→ da	→ fi	→ sv
da →		-0.60	-0.52
fi →	-1.23		**-2.14**
sv →	-0.46	**-1.14**	

Table 7: Absolute changes in BLEU scores from word-based translations to morph-based translations. The maximum phrase length was 7 for words and 10 for morphs. 4-gram language models were used for both.

(Virpioja, Väyrynen, Creutz, Sadeniemi, 2007)

Effect of Morfessor on OOV words

word / morph	→ da	→ fi	→ sv
da →		128 / 31	74 / 12
fi →	189 / 41		195 / 44
sv →	76 / 21	132 / 42	

Table 9: Number of sentences not fully translated out of 1 000 with word-based and morph-based phrases. The numbers were the same with all of the tested language models and maximum phrase length combinations.

(Virpioja, Väyrynen, Creutz, Sadeniemi, 2007)

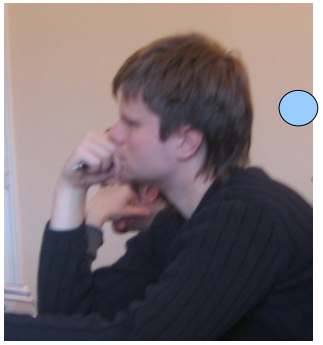
Experiment in perception-based translation

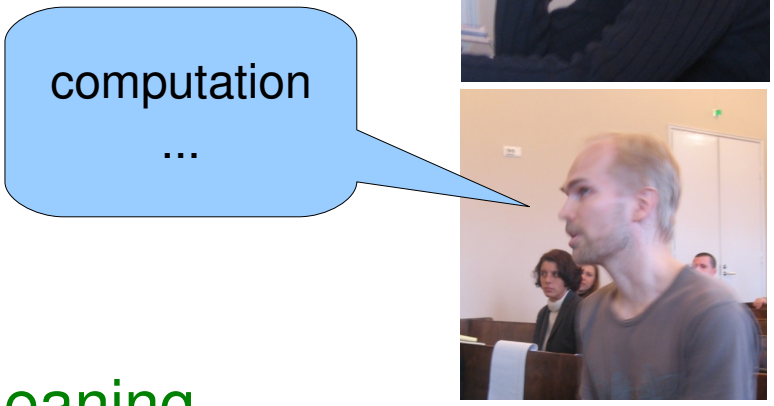
- We conducted an analysis of some tens of thousands of images and associated words
- The description was either in English or in French
- Through detection of the similarity of images we were able to find automatically correspondences between English and French words visible in the images, such as 'crayon' – 'pen' and 'pencil'

(Sjöberg, Viitaniemi, Laaksonen & Honkela, 2006)


i n k o r e k ä p j ä i v o m
e o i k ä i ä p k v r j n o m
i r ä o o i e ä p k m k j v n
e k o i r j i m v o p n ä k ä
o n i ä o j k i k v m r e p ä
e o p m k n v ä r o j ä k
m k v k e i j o p o i ä n ä r
o e ä k o i ä n i p k r m j v
r v o ä e i o k j m ä n i k p
m j ä i k p k i ä n v e o r o
6

Intersubjektiivisuus kommunikaatiossa

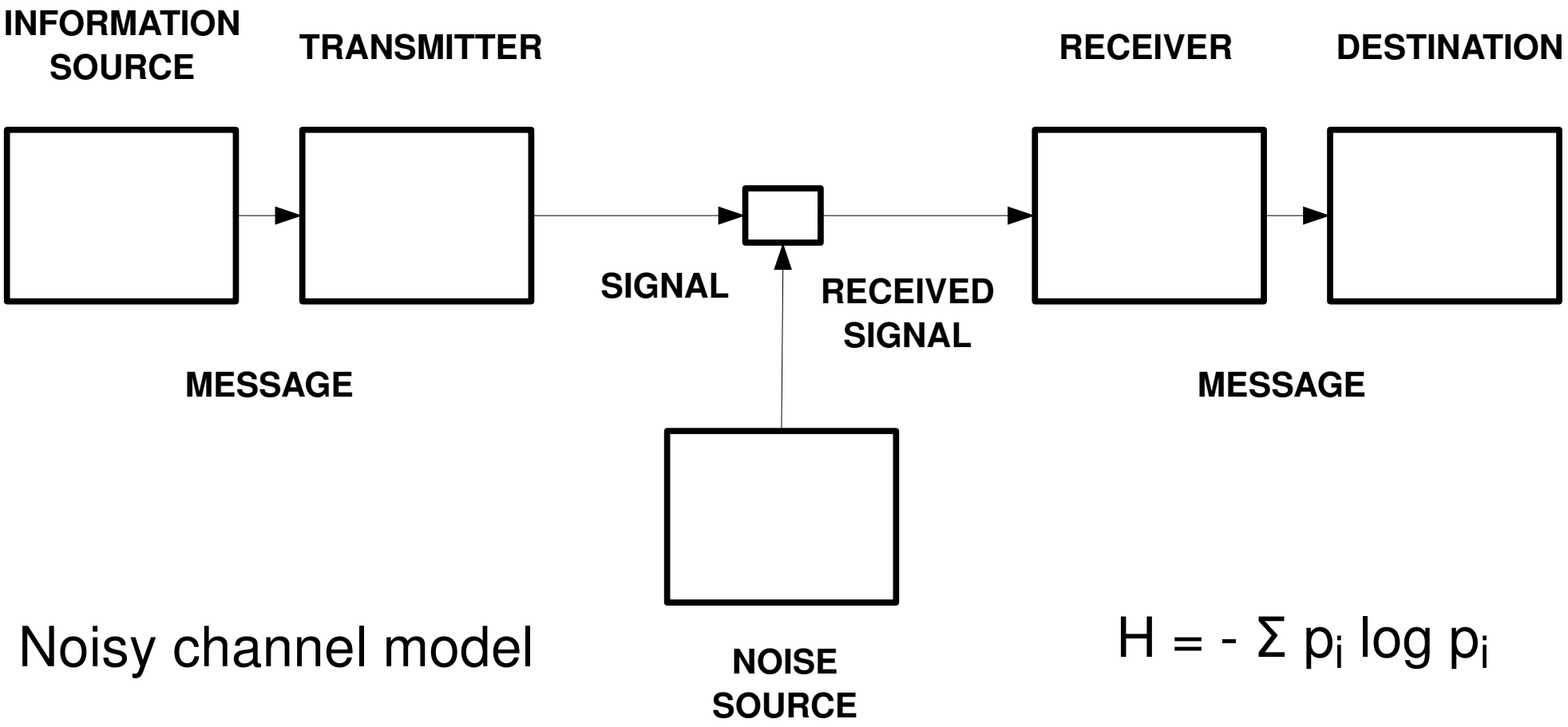




Subjectivity of interpretation

Meaning negotiation

General communication system and measuring information (Shannon & Weaver)



Noisy channel model

$$H = - \Sigma p_i \log p_i$$

Weaver on Shannon

- “Relative to the broad subject of communication, there seem to be problems at three levels. [...]
  - LEVEL A. How accurately can the symbols of communication be transmitted? (The technical problem)
  - LEVEL B. How precisely do the transmitted symbols convey the desired meaning? (The semantic problem)
  - **LEVEL C. How effectively does the received meaning affect conduct in the desired way? (The effectiveness problem)**”
- “The semantic problems are concerned with the identity, or satisfactorily close approximation, in the interpretation of meaning by the receiver, as compared with the intended meaning of the sender.” (1949, p. 4)

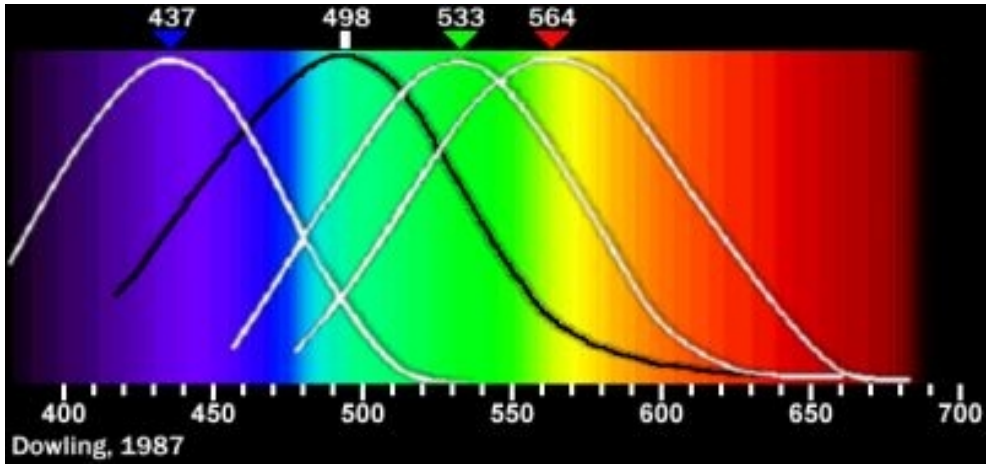
Modeling distributional similarity: word space models

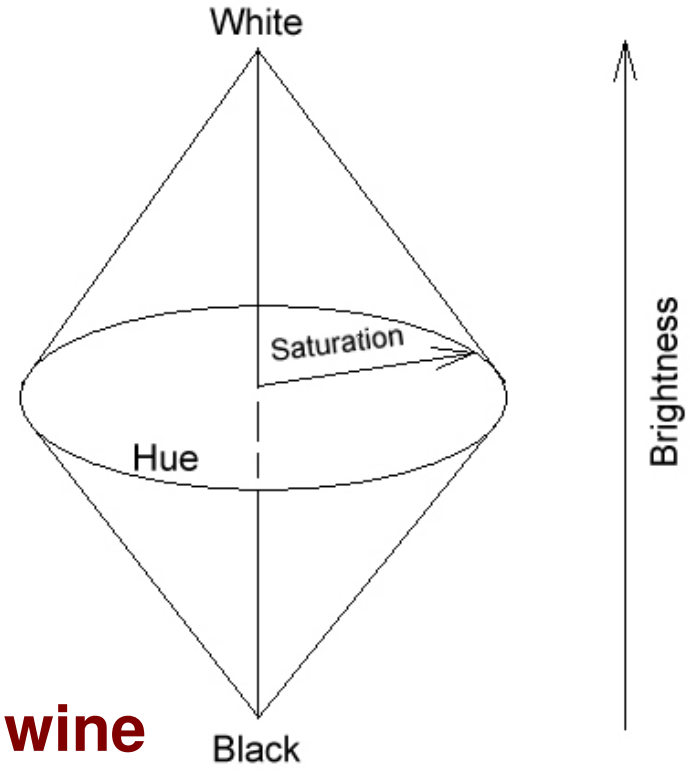
- Word space models represent meaning as points or areas in a high dimensional vector space
  - Self-Organizing Semantic Maps (Ritter and Kohonen 1989)
  - LSA (Landauer & Dumais 1997)
  - HAL (Lund & Burgess 1996)
  - Conceptual spaces (Gärdenfors 2000)
  - Word ICA (Honkela, Hyvärinen & Väyrynen 2004)
  - etc. etc.

Distributional hypothesis

- Two words are semantically similar to the extent that their contextual representations are similar (Miller & Charles 1991)
- The meaning of words is in their use (Wittgenstein)

Simple challenge: Color naming



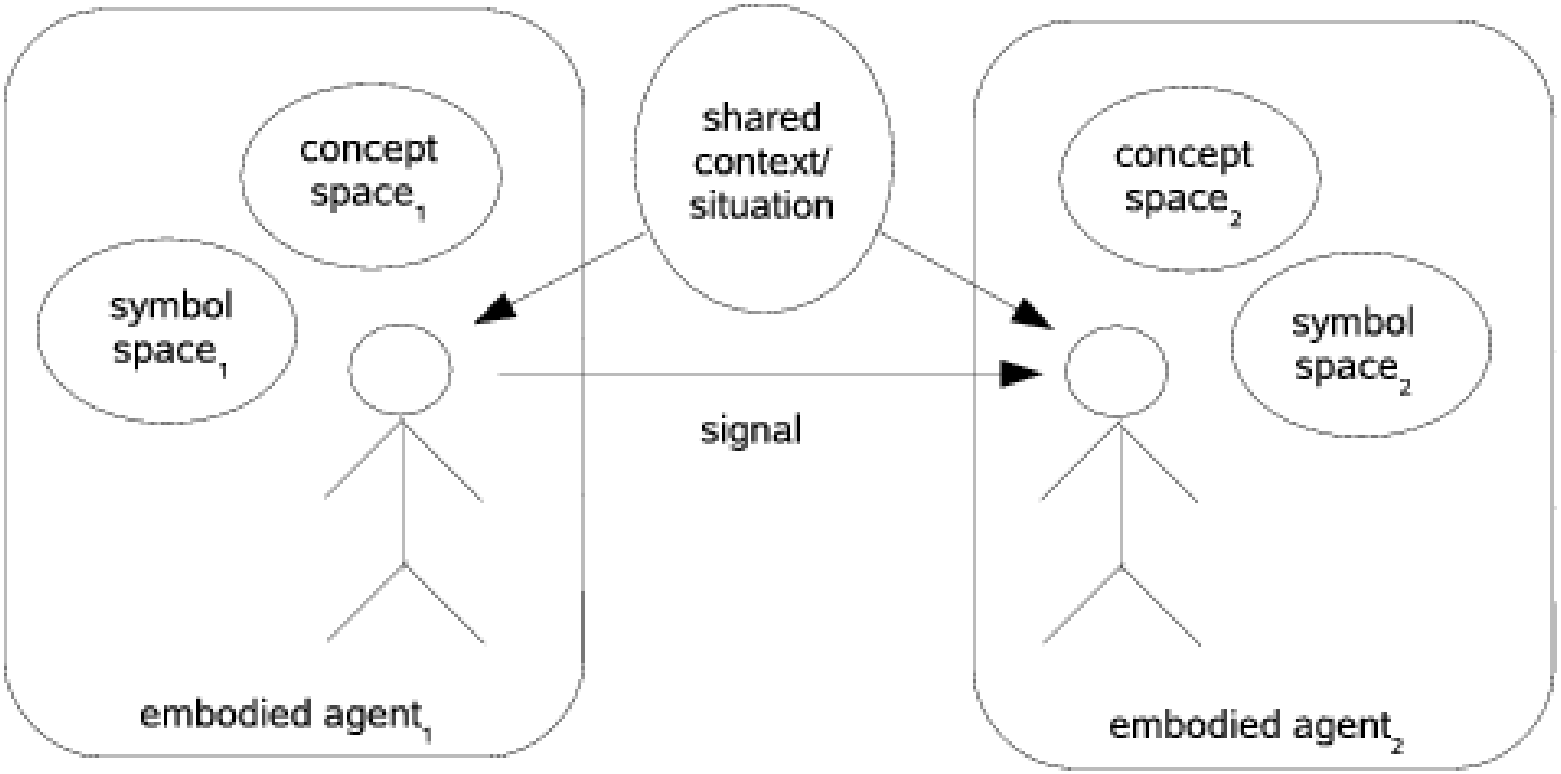


Human vision: rods, cones,...
Physical reasons for color
Contextuality of naming

red wine
red skin
red shirt

→ Hardin
→ Gärdenfors

Theory of Intersubjective Meaning Spaces (IMS)



(Honkela, Könönen, Lindh-Knuutila & Paukkeri, 2008)

IMS: Theoretical framework

- Agent's internal view of its context, *the concept space*
- The concept space is spanned by a number of features.
- Dimensionalities of the concept spaces can be different for each agent
- Therefore, we denote the features used by agent 1 f_i^1, $i = 1 \ldots N$
- Similarly, agent 2 uses features f_i^2, $i = 1 \ldots M$
- Thus, the concept space of agent 1 is an N-dimensional metric space C^1, and for agent 2, C^2

(Honkela, Könönen, Lindh-Knuutila & Paukkeri, 2008)

IMS: Theoretical framework

- In the framework, there exist two distance measures, namely ω and λ. ω gives a distance between two points inside the concept space of the agent, i.e. $\omega : C^i \times C^i \rightarrow \mathbb{R}, i = 1, 2,$
- λ gives a distance between two points in the concept spaces of the different agents, i.e. $\lambda : C^i \times C^j \rightarrow \mathbb{R}, i \neq j$
- The symbol space S^1 of agent 1 is its vocabulary that consists of discrete symbols

(Honkela, Könönen, Lindh-Knuutila & Paukkeri, 2008)

IMS: Theoretical framework

- An agent i has an individual mapping function ξ^i that maps the symbol $s^i \in S^i$ to C^i
- An agent i expresses each symbol $s^i \in S^i$ as a signal d in the signal space D
- We assume that the signal space D is multidimensional, continuous and shared between the agents
- However, each agent i has an individual mapping function ϕ^i from its vocabulary to the signal space, i.e. $\phi^i : S^i \rightarrow D$ and an inverse mapping ϕ^{-i} from the signal space to the symbol space
- In the single-agent communication model, the sender is communicating with the receiver but it has only its own conceptual space available

IMS: Theoretical framework

- The sender, agent 1, selects a symbol that corresponds best to the current context by the means of its own conceptual model
- Therefore this model can be seen as an optimal decision problem.
- Let us denote the symbol agent 1 selects and communicates to agent 2 as s^* and the features agent 1 observes corresponding to the current context as f^1. Then agent 1 selects the symbol that corresponds the current observations best by the means of some distance measure ω

(Honkela, Könönen, Lindh-Knuutila & Paukkeri, 2008)

IMS: Theoretical framework

- One suitable choice for ω in that case is for example Euclidean distance
- Formally this can be represented as in Eq. (1).

$$s^* = \arg\min_{s \in S^1} \omega(f^1, \xi^1(s)) \qquad (1)$$

- After symbol selection process, agent 1 communicates the symbol s^* to agent 2, i.e.:

$$d = \phi^1(s^*) \qquad (2)$$

(Honkela, Könönen, Lindh-Knuutila & Paukkeri, 2008)

IMS: Theoretical framework

- When agent 2 observes d, it maps it to some $s^2 \in S^2$ by using the function ϕ^{-2}
- Then it maps the symbol to some point in its conceptual space by using ξ^2

(Honkela, Könönen, Lindh-Knuutila & Paukkeri, 2008)

IMS: Theoretical framework

- If this point is very near of its own observation f^2, we can say that the communication process has succeeded

- Mathematically this is as follows:

$$||\xi^2(\phi^{-2}(d)) - f^2|| \leq \epsilon, \qquad (3)$$

where ϵ is a small constant and $||\cdot||$ is some suitable norm in C^2.

(Honkela, Könönen, Lindh-Knuutila & Paukkeri, 2008)

IMS: Theoretical framework

- The two-agent model relates to the single-agent model very closely but now the sender has some estimate of the receiver's conceptual space available
- This model can be learned from communication samples or it can be known a priori
- The symbol selection process is formally given in Eq. (4)

$$s^* = \arg\min_{s \in \tilde{S}^2} \lambda(f^1, \tilde{\xi}^2(s)) \qquad (4)$$

(Honkela, Könönen, Lindh-Knuutila & Paukkeri, 2008)

IMS: Theoretical framework

- In this equation, $\tilde{\xi}^2$ is the model of the receiver
- Note that also the vocabulary of the receiver can be unknown and should be estimated by $\tilde{S}^2$

(Honkela, Könönen, Lindh-Knuutila & Paukkeri, 2008)

IMS: Theoretical framework

- As the dimensionalities of the conceptual spaces C^1 and C^2 can be different, we should use some special method to calculate the distance between points in these two spaces
- Dynamic programming and neural networks may provide solutions to this task
- A neural network solution for the general case when agents have different representations for the same phenomenon has been considered for a supervised learning task by Laakso and Cottrell (2000) and for unsupervised learning by Raitio, Vigário, Särelä and Honkela (2004)

(Honkela, Könönen, Lindh-Knuutila & Paukkeri, 2008)

IMS: Theoretical framework

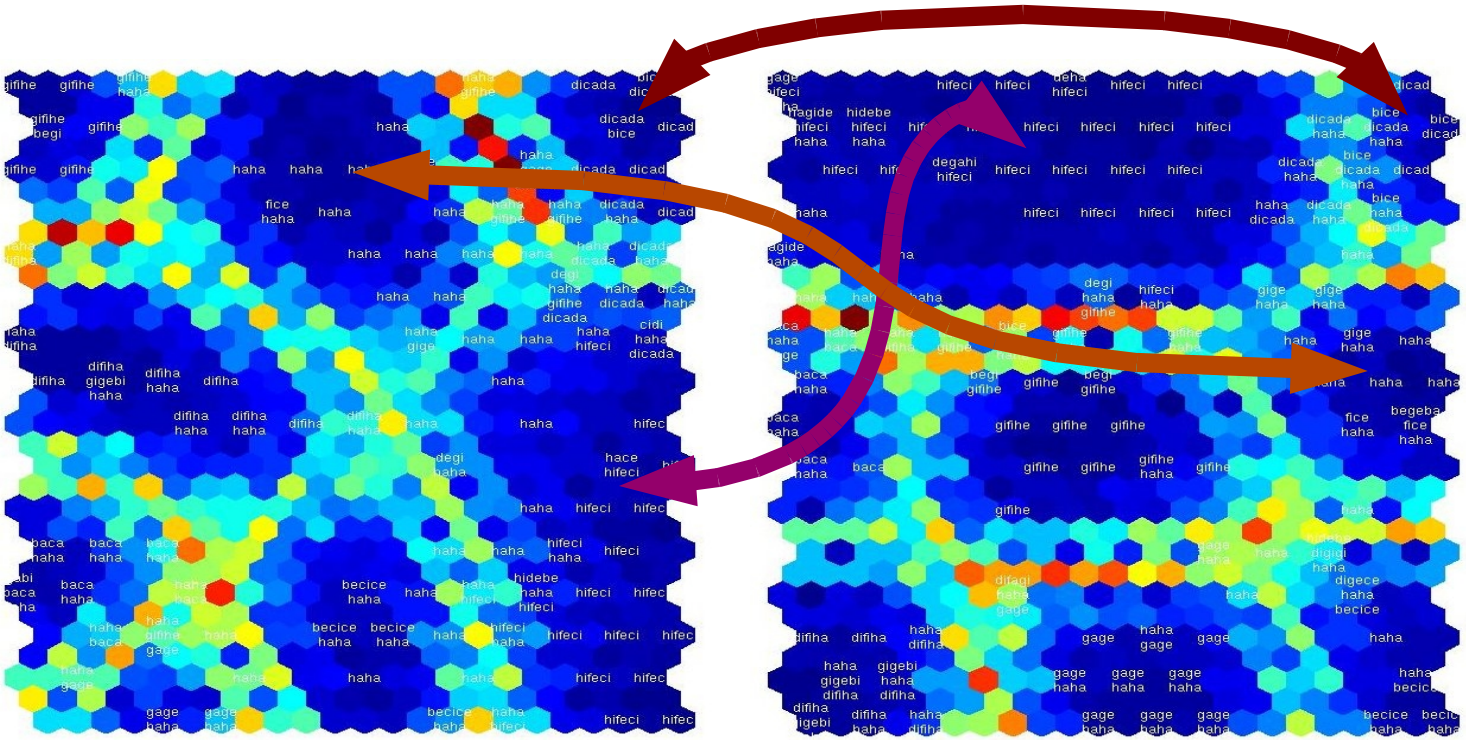
- Other parts of the communication process remain the same as in the single-agent case
- As agent 1 utilizes the explicit model of agent 2, the symbol selection problem can be seen as a game theoretical problem

IMS: Theoretical framework

- As an example related to the vocabulary problem, two persons may have different conceptual or terminological "density" of the topic under consideration
- A layperson, for instance, is likely to describe a phenomenon in general terms whereas an expert uses more specific terms
- In our framework, symbols generate clusters in the conceptual spaces of the agents

(Honkela, Könönen, Lindh-Knuutila & Paukkeri, 2008)

Emergence of a coherent lexicon in a community of interacting SOM-based agents



(Lindh-Knuutila, Lagus & Honkela, SAB'06)

Related to e.g. Steels and Vogt on language games

Practical consequences

- The traditional notion of uncertainty in decision making does not cover the uncertainties caused by differences in conceptual systems of individual agents within a community
- In many transactions, including symbolic/linguistic communication, the differences in the underlying conceptual systems play an important role
- Serious efforts have been made to harmonize or to standardize the classification systems or ontologies used by agents
- Even if standardization is conducted, there can not be any true guarantee that all participating agents would share the meaning of all the expressions used in the transactions in various contexts


i n k o r e k ä p j ä i v o m
e o i k ä i ä p k v r j n o m
i r ä o o i e ä p k m k j v n
e k o i r j i m v o p n ä k ä
o n i ä o j k i k v m r e p ä
e o p m k i n v i ä r o j ä k
m k v k e i j o p o i ä n ä r
o e ä k o i ä n i p k r m j v
r v o ä e i o k j m ä n i k p
m j ä i k p k i ä n v e o r o

Language as a complex phenomenon

- A large proportion of modern human activity in its different forms (science, industry, society, culture, etc.) is based on the use of language
- There are at least 6000 languages in the world and many more dialects
- Each language has the order of 10^5 to 10^{10} different word forms
- Each word is understood differently by each speaker of that language at least to some degree

Viitteitä, 2007-2008

Timo Honkela, Ville Könönen, Tiina Lindh-Knuutila, and Mari-Sanna Paukkeri. Simulating processes of concept formation and communication. Journal of Economic Methodology, 15(3):245–259, 2008.

Mari-Sanna Paukkeri, Ilari T. Nieminen, Matti Pöllä, and Timo Honkela. A language-independent approach to keyphrase extraction and evaluation. In Coling 2008: Companion volume, pages 83–86, 2008.

Mikaela Klami and Timo Honkela. Self-organized ordering of terms and documents in NSF awards data. In Proceedings of WSOM'07. University Library of Bielefeld, 2007.

Matti Pöllä and Timo Honkela. Probabilistic text change detection using an immune model. In Proceedings of the International Joint Conference on Neural Networks, IJCNN 2007, pages 1109–1114, Orlando, Florida, August 2007.

Sami Virpioja, Jaakko J. Väyrynen, Mathias Creutz, and Markus Sadeniemi. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In Proceedings of the Machine Translation Summit XI, pages 491–498, September 2007.

Viitteitä, 2006-2007

Jaakko J. Väyrynen, Lasse Lindqvist, and Timo Honkela. Sparse distributed representations for words with thresholded independent component analysis. In Proceedings of IJCNN'07, pages 1031–1036, 2007.

Mathias Creutz. Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition. Doctoral thesis, Dissertations in Computer and Information Science, Report D13, Helsinki University of Technology, Espoo, Finland, 2006.

Tiina Lindh-Knuutila, Timo Honkela, and Krista Lagus. Simulating meaning negotiation using observational language games. In P. Vogt et al., editor, Symbol Grounding and Beyond: Proceedings of the Third International Workshop on the Emergence and Evolution of Linguistic Communication, pages 168–179, Rome, Italy, 2006. Springer, Berlin/Heidelberg.

Mats Sjöberg, Ville Viitaniemi, Jorma Laaksonen, and Timo Honkela. Analysis of semantic information available in an image collection augmented with auxiliary data. In Ilias Maglogiannis, Kostas Karpouzis, and Max Bramer, editors, Proceedings of AIAI'06, Artificial Intelligence Applications and Innovations, volume 204, pages 600–608. Springer, 2006.

Viitteitä, 1995-2004

Mathias Creutz and Krista Lagus. Induction of a simple morphology for highly-inflecting languages. In Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON), pages 43–51, Barcelona, July 2004.

Timo Honkela and Aapo Hyvärinen. Linguistic feature extraction using independent component analysis. In Proceedings of IJCNN'07, pages 279–284, Budabest, Hungary, July 2004.

Mathias Creutz, and Krista Lagus. Unsupervised discovery of morphemes. In Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02, pages 21-30, Philadelphia, Pennsylvania, USA, July 11, 2002.

Hyvärinen A., Karhunen J., and Oja E. Independent component analysis. Wiley, 2001.

Timo Honkela, Ville Pulkki, and Teuvo Kohonen. Contextual relations of words in Grimm tales, analyzed by self-organizing map. In F. Fogelman-Soulié and P. Gallinari, editors, Proc. of ICANN'95, International Conference on Artificial Neural Networks, volume II, pages 3–7, Nanterre, France, 1995. EC2.

Viitteitä, 1988-1992

Riitta Alkula and Timo Honkela. Development of text storage and information retrieval methods with natural language processing components. Final report of the FULLTEXT project (in Finnish). VTT, Espoo, Finland, 1992.

Harri Jäppinen, Timo Honkela, Heikki Hyötyniemi, and Aarno Lehtola. Hierarchical multilevel processing model for natural language database interface. In Proceedings of CAIA'88, 4th Conference on Artificial Intelligence Applications, pages 332–337. IEEE (The Computer Society of the IEEE), 1988.

Harri Jäppinen, Timo Honkela, Heikki Hyötyniemi, and Aarno Lehtola. A multilevel natural language processing model. Nordic Journal of Linguistics, 11:69–82, 1988.