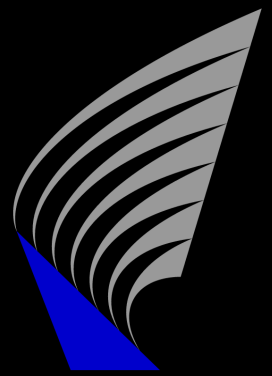
Language modeling using unsupervised learning methods with applications in IR and MT



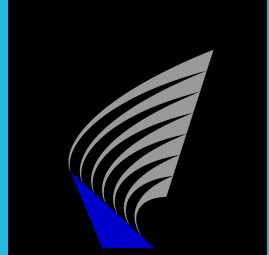
Timo Honkela

Cognitive Systems Research Group
Adaptive Informatics Research Centre
Helsinki University of Technology

www.cis.hut.fi/cis/

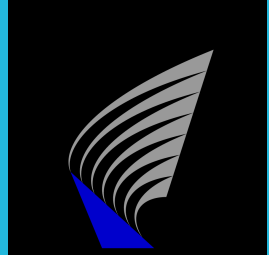
AERFAI summer school

Focus



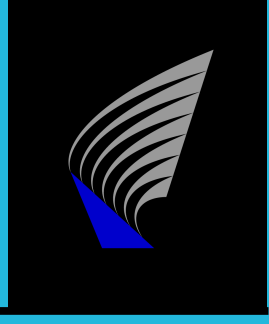
- The main focus in this presentation is in **semantics** and **pragmatics**, and a secondary one in **morphology**
- Applications considered include **information retrieval**, **speech recognition** and **machine translation**

Agenda



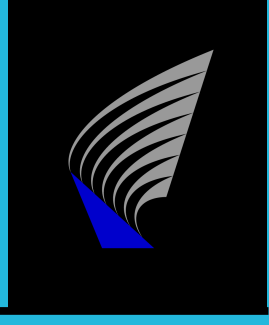
- Learning paradigms
- Philosophical and practical motivation for unsupervised learning
- Overview of unsupervised learning methods
- Case studies
- Conclusions

Case studies



- Word and document maps based on the self-organizing map (SOM)
- Qualitative analysis using text mining based on the self-organizing map
- Emergent word feature models using independent component analysis (ICA)
- Unsupervised morphological modeling using Morfessor and application in speech processing
- MT supported by the SOM and Morfessor

Background

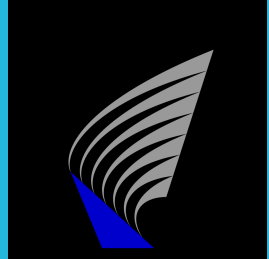


- Considering natural language as a **signal** and **dynamic system** at cognitive and social levels (also in its written form) rather than a symbolic and logical system
- Importance of **embodiment** (cf. e.g. Harnad) and **embeddedness** (cf. e.g. Edelman)
- **Learning** and **pattern recognition** processes are essential (as opposed to the theories presented e.g. by Chomsky, Fodor, Pinker); much of the learning is bound to be **unsupervised**

Agenda

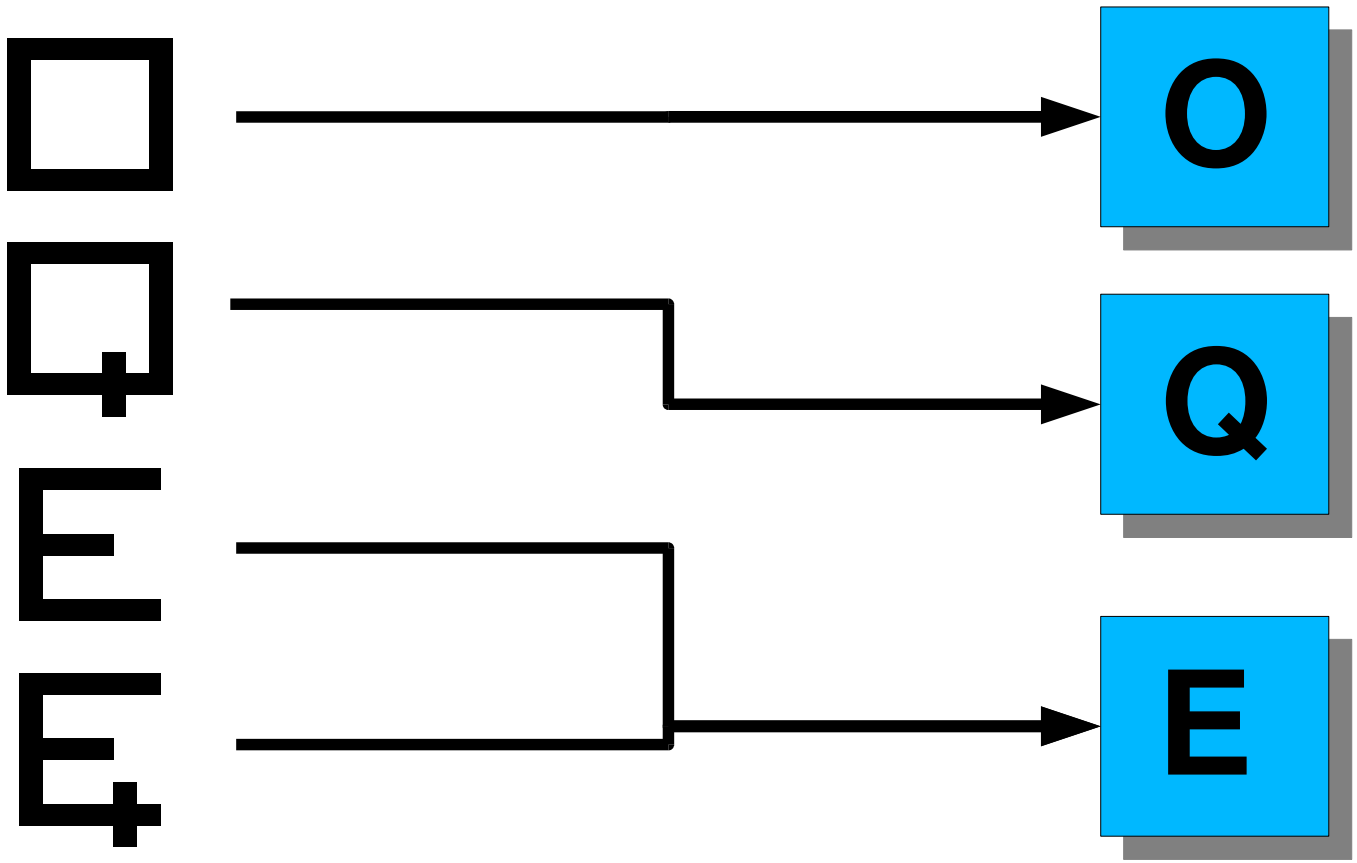
- **Learning paradigms**
- Philosophical and practical motivation for unsupervised learning
- Overview of unsupervised learning methods
- Case studies
- Conclusions

Learning paradigms

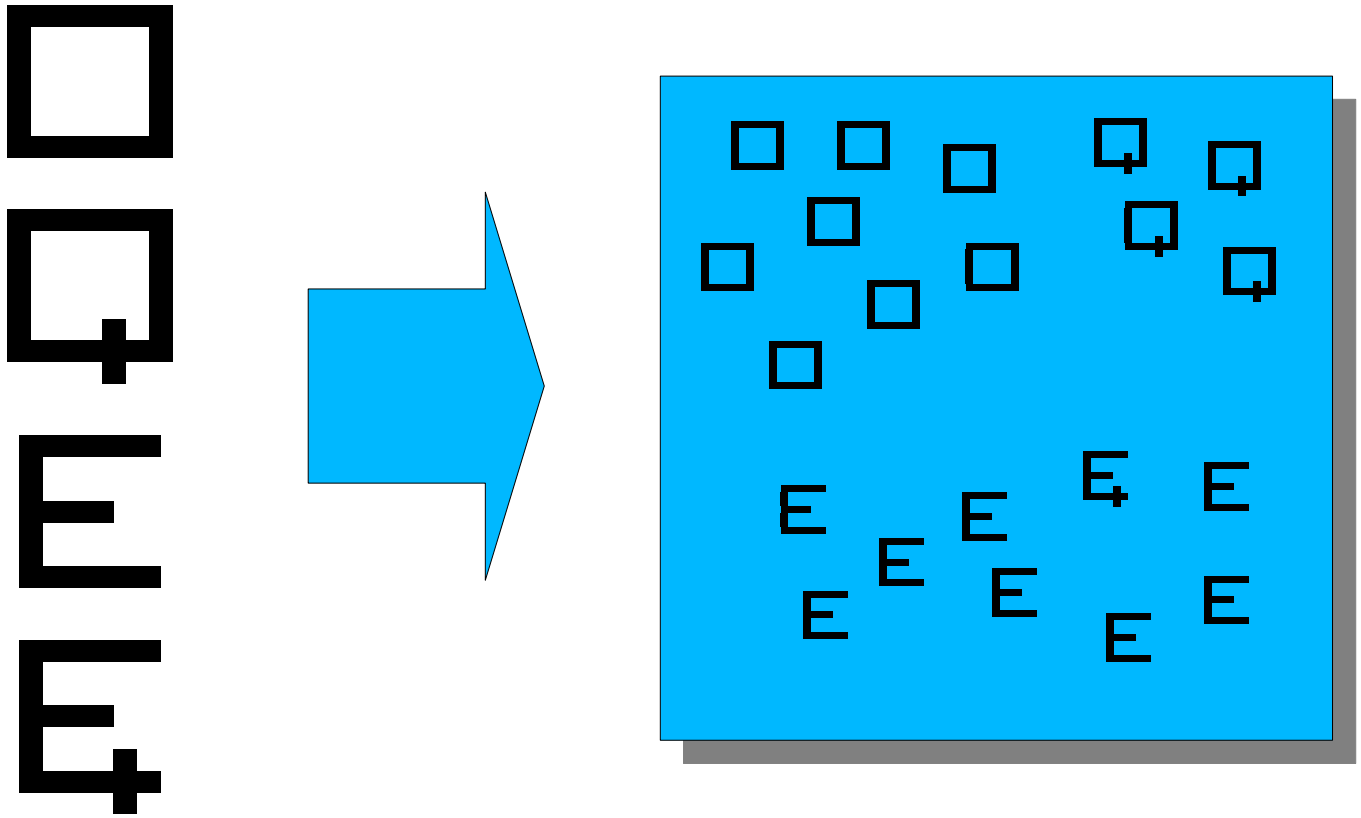


- Supervised learning
- Reinforcement learning
- Unsupervised learning

Supervised learning



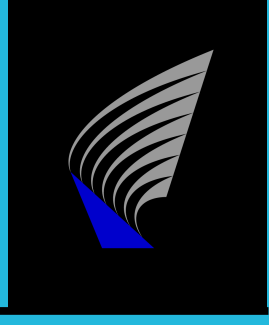
Unsupervised learning



Agenda

- Learning paradigms
- **Philosophical and practical motivation for unsupervised learning**
- Overview of unsupervised learning methods
- Case studies
- Conclusions

Traditional formalization of meaning

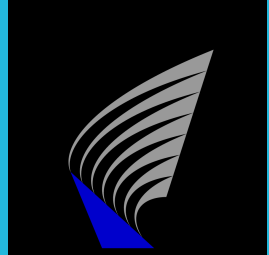


- Formalisms like first-order predicate logic have widely been used as a basis for theories of meaning (epistemology); consider also contemporary efforts such as Semantic Web
- These formalisms provide only limited means for creating in-depth theories of how language is understood

Nature of traditional knowledge representations

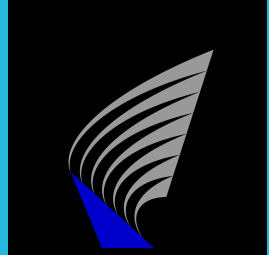
```
event(e781,
 [human(e872),
  name(e872,
   "Peter Gärdenfors"),
  human(e912),
  name(e912,
    "Linda B. Smith"),  ...
  discuss(e872,e661),  ...])
```

Traditional formalization of meaning, cont'd



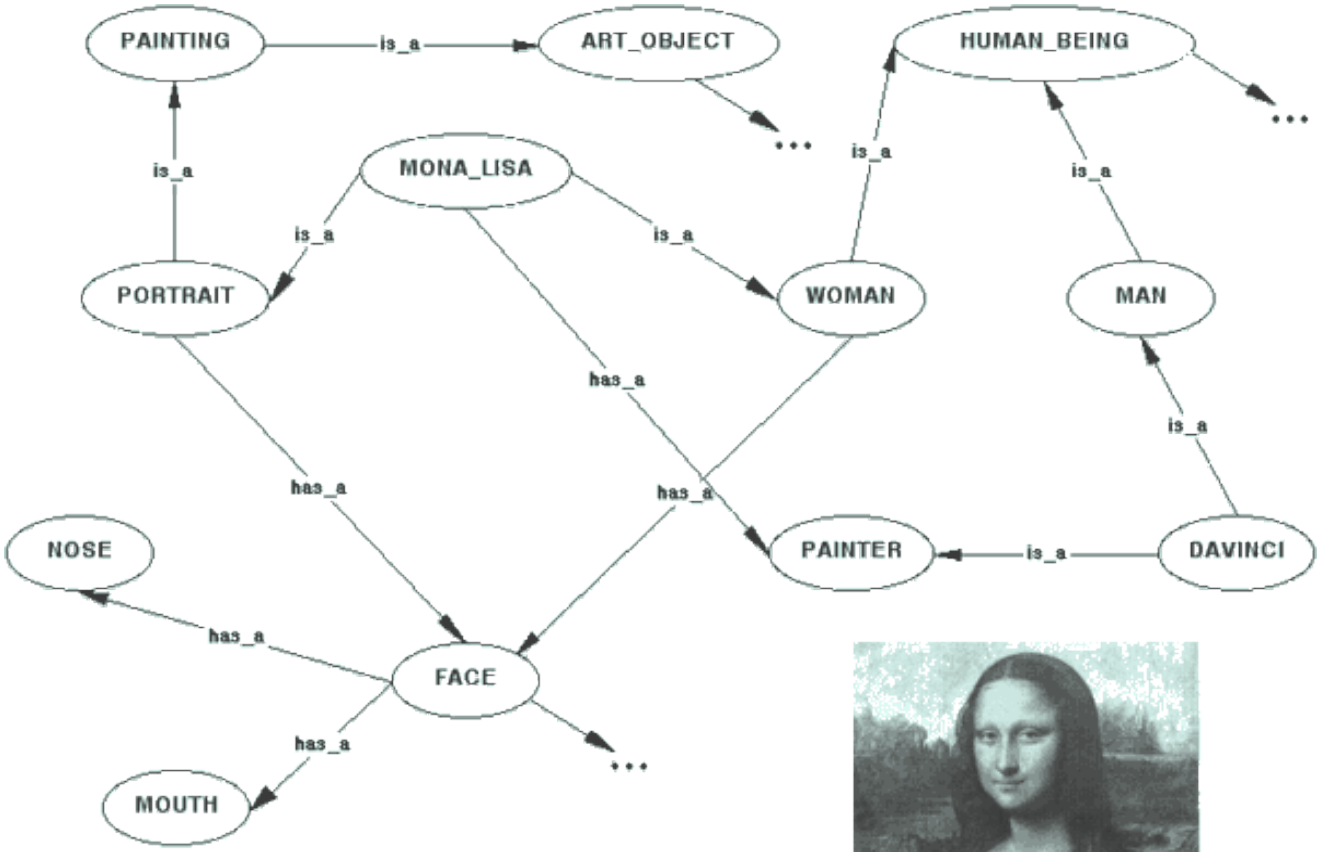
- Traditional logic provides means e.g. for modeling quantification, connectives, analytical truths and conceptual hierarchies
- However, many semantic phenomena are matters of degree. Various proposals that apply Bayesian probability theory or fuzzy sets deal with this

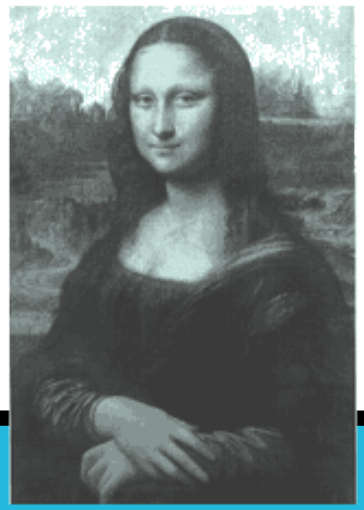
Neglected pattern recognition processes



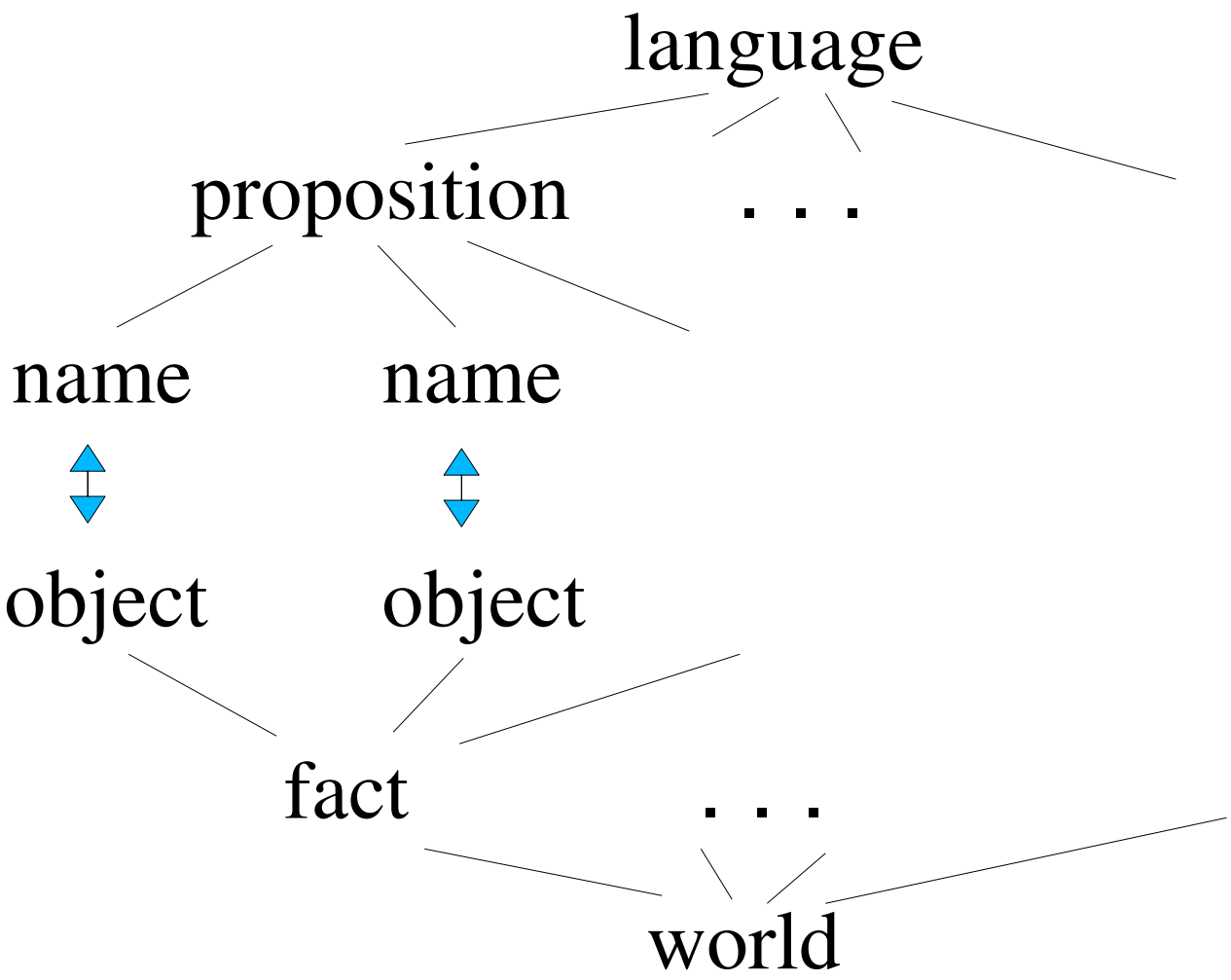
- Even these methodological extensions do not suffice if the pattern recognition processes are not taken into account
- The world is not straightforwardly experienced as discrete objects and events but there are complex underlying cognitive processes involved

“symbols” versus “reality”

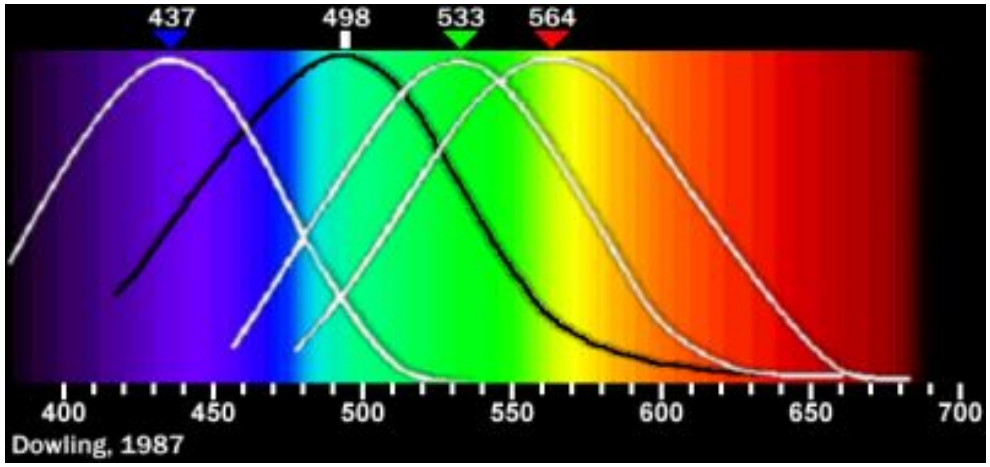


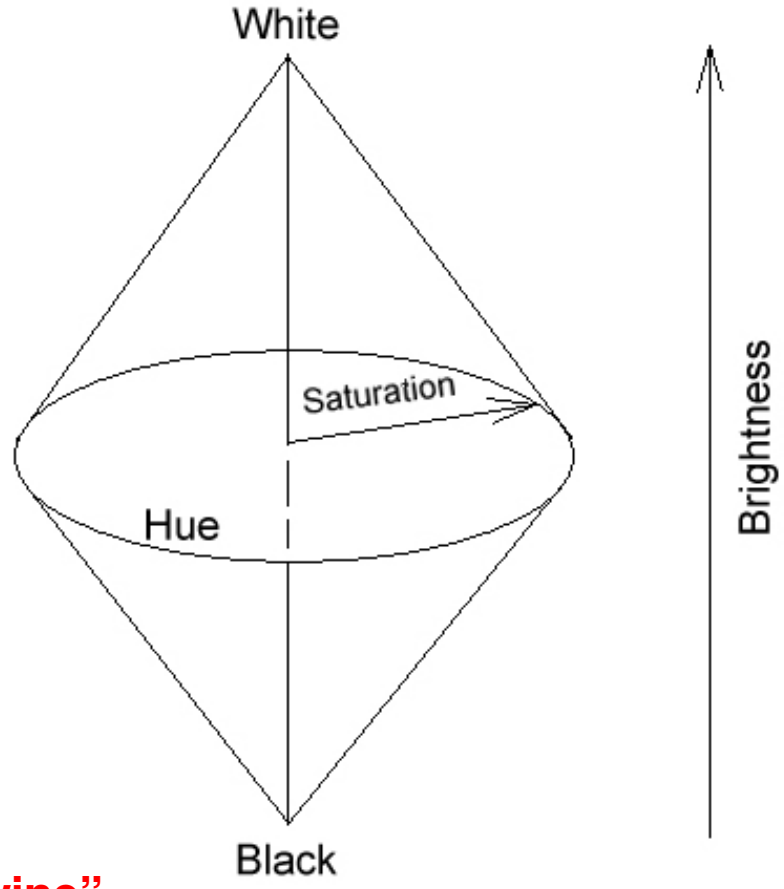


Formal semantics:
Early Wittgenstein



Example: Color naming





Human vision: rods, cones,...

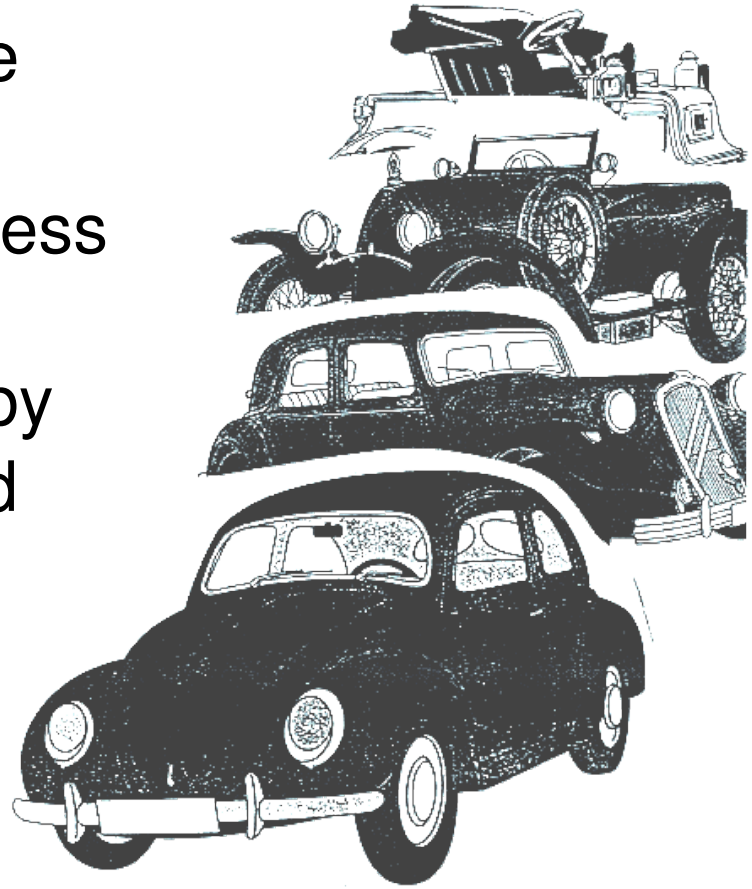
Physical reasons for color

Contextuality of naming

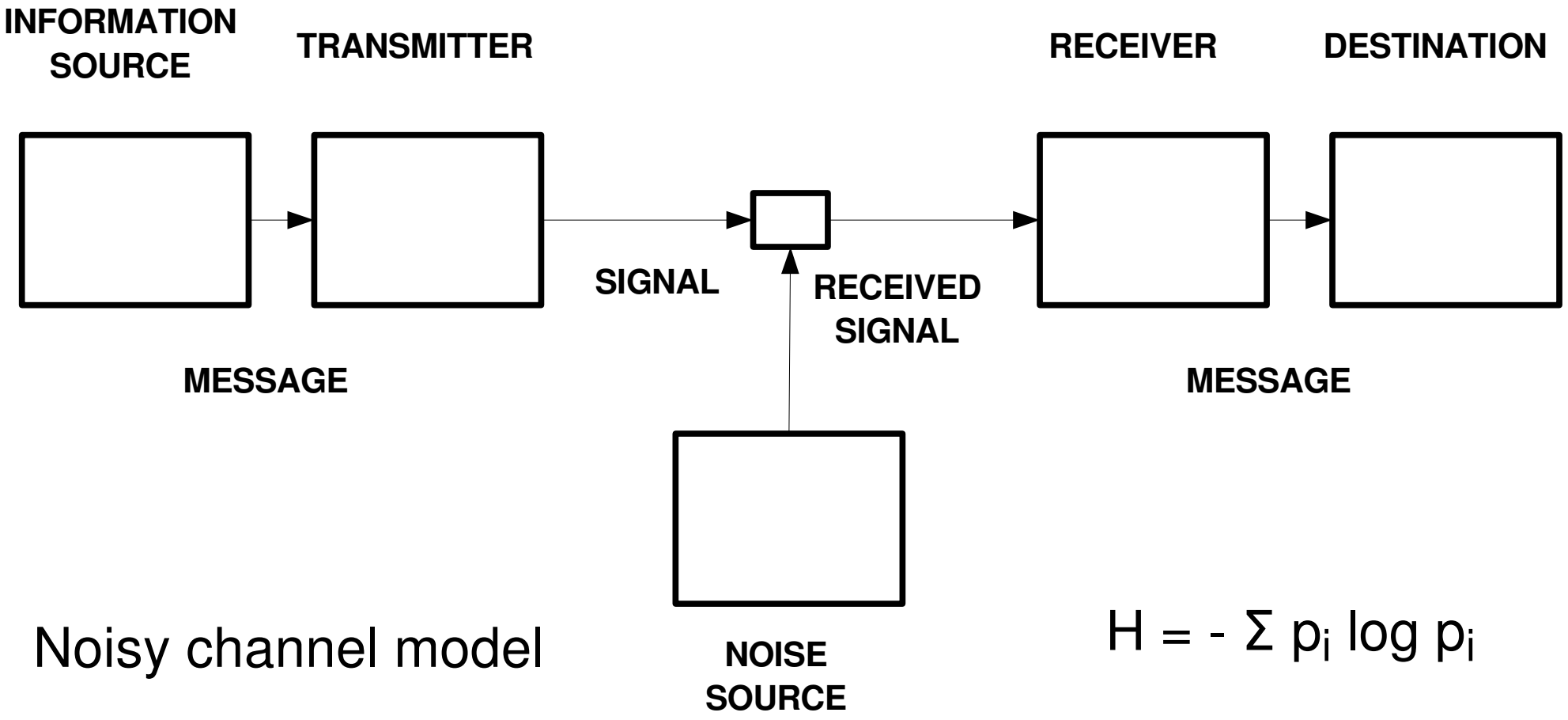
“red wine”

Experiential grounding of human knowledge

Human understanding of the world and of the relationship between language use and perception and action within the world is based on a long learning process for which the genotype gives a certain basis but which is mainly determined by the individual interaction with the world including other human beings and the social and cultural context; partial rejection of “Poverty of Stimulus” argument



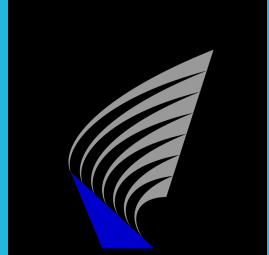
General communication system and measuring information (Shannon & Weaver)



Noisy channel model

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

Weaver on Shannon

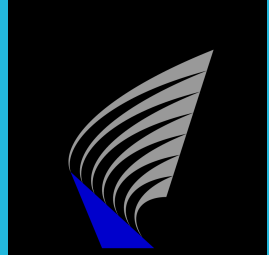


- “Relative to the broad subject of communication, there seem to be problems at three levels. [...]
  - LEVEL A. How accurately can the symbols of communication be transmitted? (The technical problem)
  - LEVEL B. How precisely do the transmitted symbols convey the desired meaning? (The semantic problem)
  - 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)

Agenda

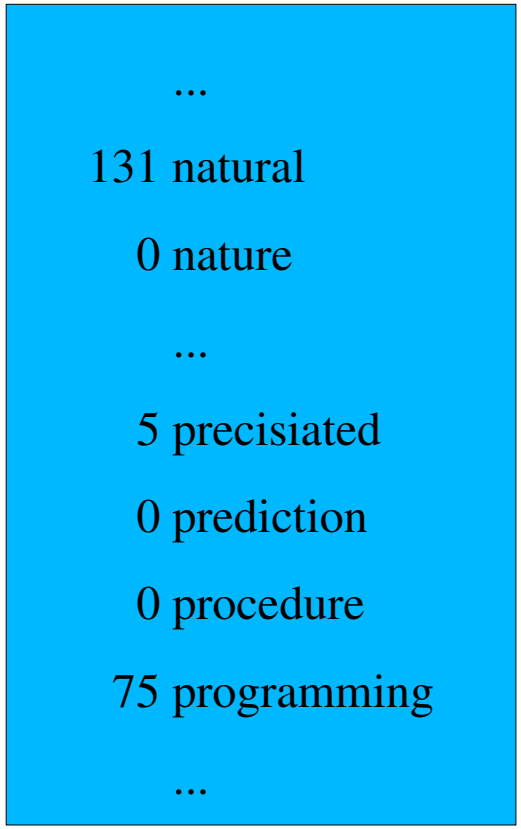
- Learning paradigms
- Philosophical and practical motivation for unsupervised learning
- **Overview of unsupervised learning methods for language modeling**
- Case studies
- Conclusions

Modeling language learning



- There have been a large number of efforts to device systems that learn syntactic level of language (cf. e.g. the classical PDP work on learning past tense of English verbs, McClelland & Rumelhart, 1986)
- Learning the semantics of expressions is typically based on applying contextual information

Language in context



+

language

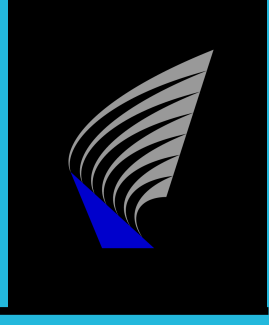
sky

rainbow

cloud

sea

Modeling distributional similarity: word space models

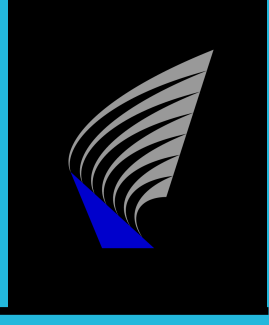


- Word space models represent meaning as points or areas in a high dimensional vector space
  - Self-Organizing Semantic Maps (Ritter and Kohonen 1989)
  - LSA/LSI (Landauer & Dumais 1997)
  - HAL (Lund & Burgess 1996)
  - Conceptual spaces (Gärdenfors 2000) (based on SOMs)
  - Word ICA (Honkela, Hyvärinen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

Agenda

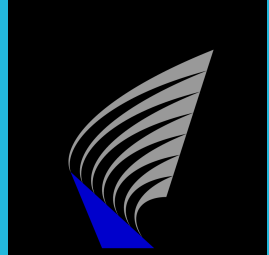
- Learning paradigms
- Philosophical and practical motivation for unsupervised learning
- Overview of unsupervised learning methods
- **Case studies**
- Conclusions

Case studies



- **Word and document maps based on the self-organizing map (SOM)**
- Qualitative analysis using text mining based on the self-organizing map
- Emergent word feature models using independent component analysis (ICA)
- Unsupervised morphological modeling using Morfessor and application in speech processing
- MT supported by the SOM and Morfessor

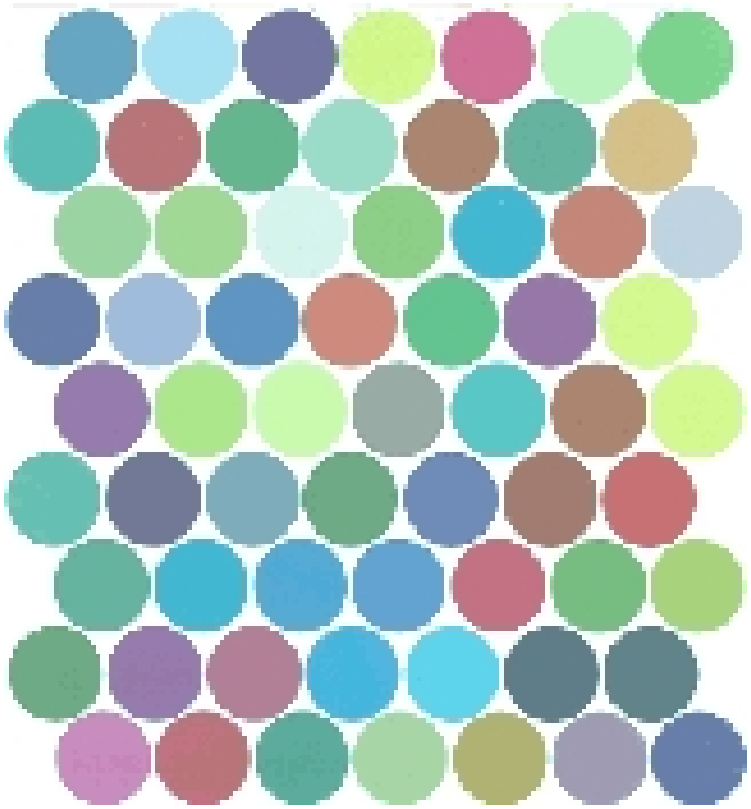
Self-Organizing Map
(Kohonen 1982, 2001)



Unsupervised learning method that:

- *provides a mapping from a high-dimensional space into a low-dimensional space* thus providing a suitable means for visualization of complex data
- *reveals clustering structure in the data by providing a representation of the topological structure of the original data* (the topological distance between two points in the map is proportional to the distance between the points in the original input space)

Self-organizing process

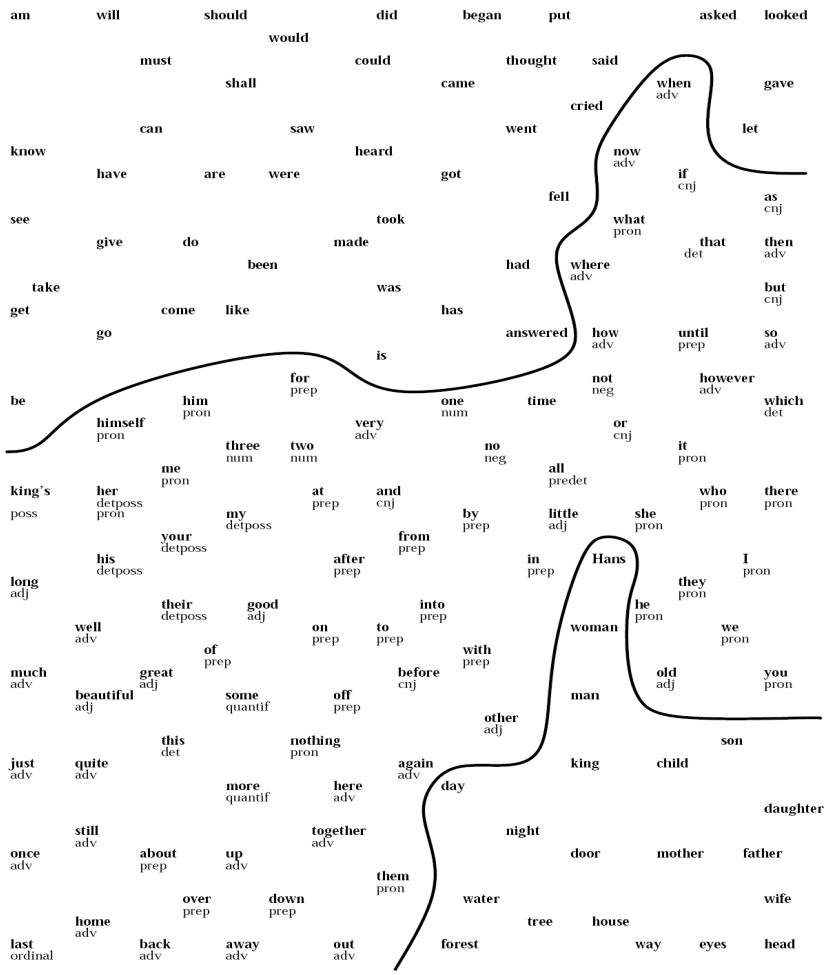


Initial random order

Organized map

SOM analysis of Grimm fairy tales

- Analysis was made by feeding word trigrams to the SOM
- No prior syntactic or semantic categorization of the words
- Distinct areas of verbs and nouns, clusters of numerals, and possessives emerged
- Distinction between animate and inanimate nouns emerged

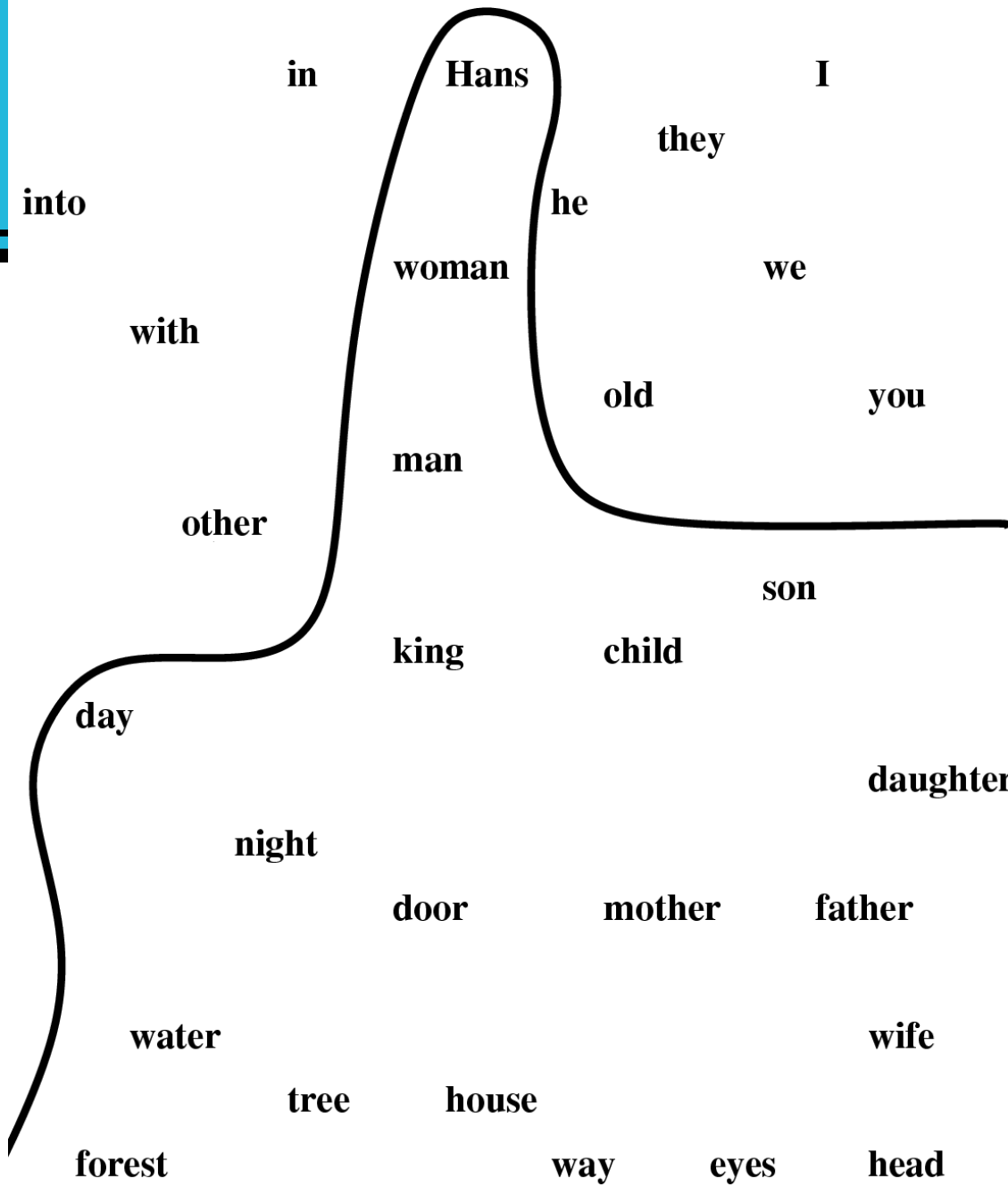


(Honkela, Pulkki & Kohonen, 1995)

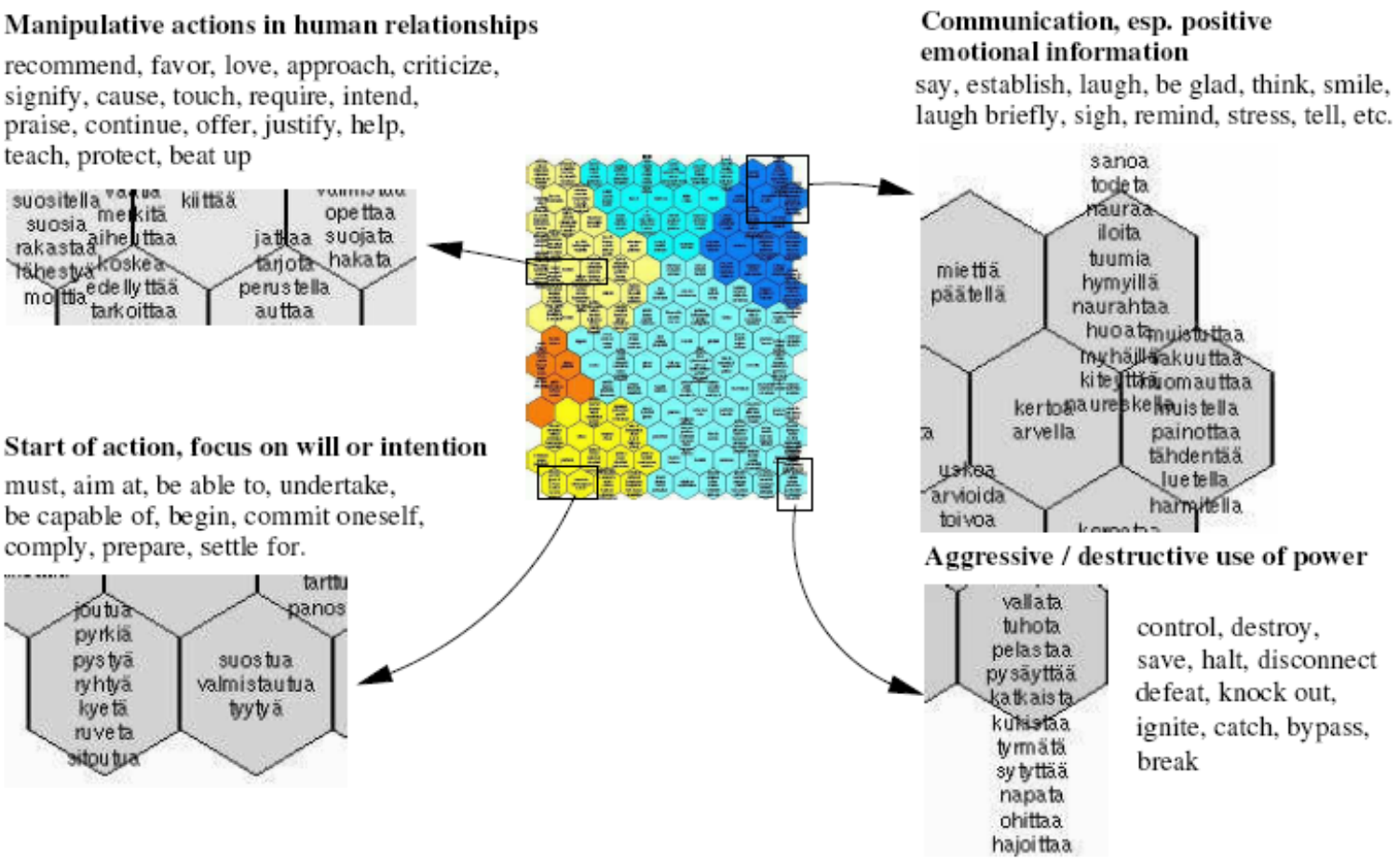
Emergent implicit categories

→

implicit knowledge (versus meta-linguistic knowledge)



Map of verbs



(Lagus, Airola & Creutz, 2002)

Emergence of ontological relations

Figure 6: Ordering of object classes on MPEG-7 EdgeHistogram SOM

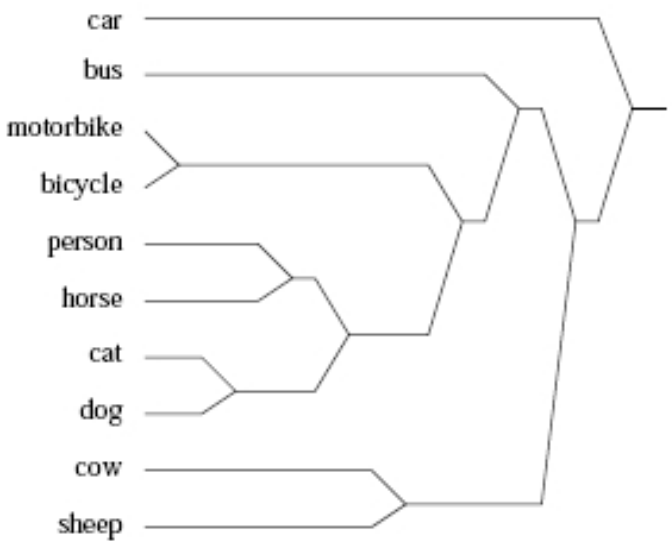
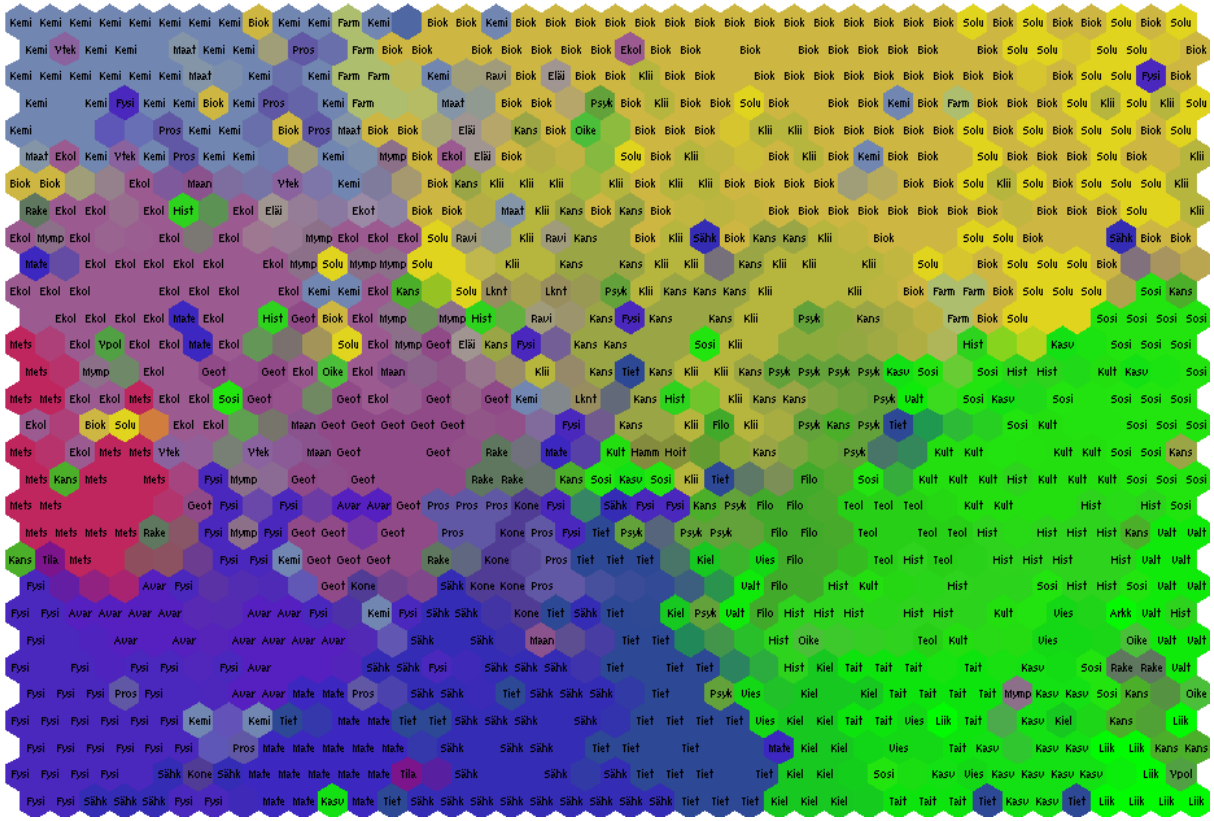


Figure 4: Taxonomy tree of visual similarity of the classes

(Laaksonen &Viitaniemi, 2006)

Map of documents: WEBSOM

(Honkela, Kaski, Kohonen & Lagus, 1996, etc.)



A self-organized semantic space of documents

Application example: Document map of research

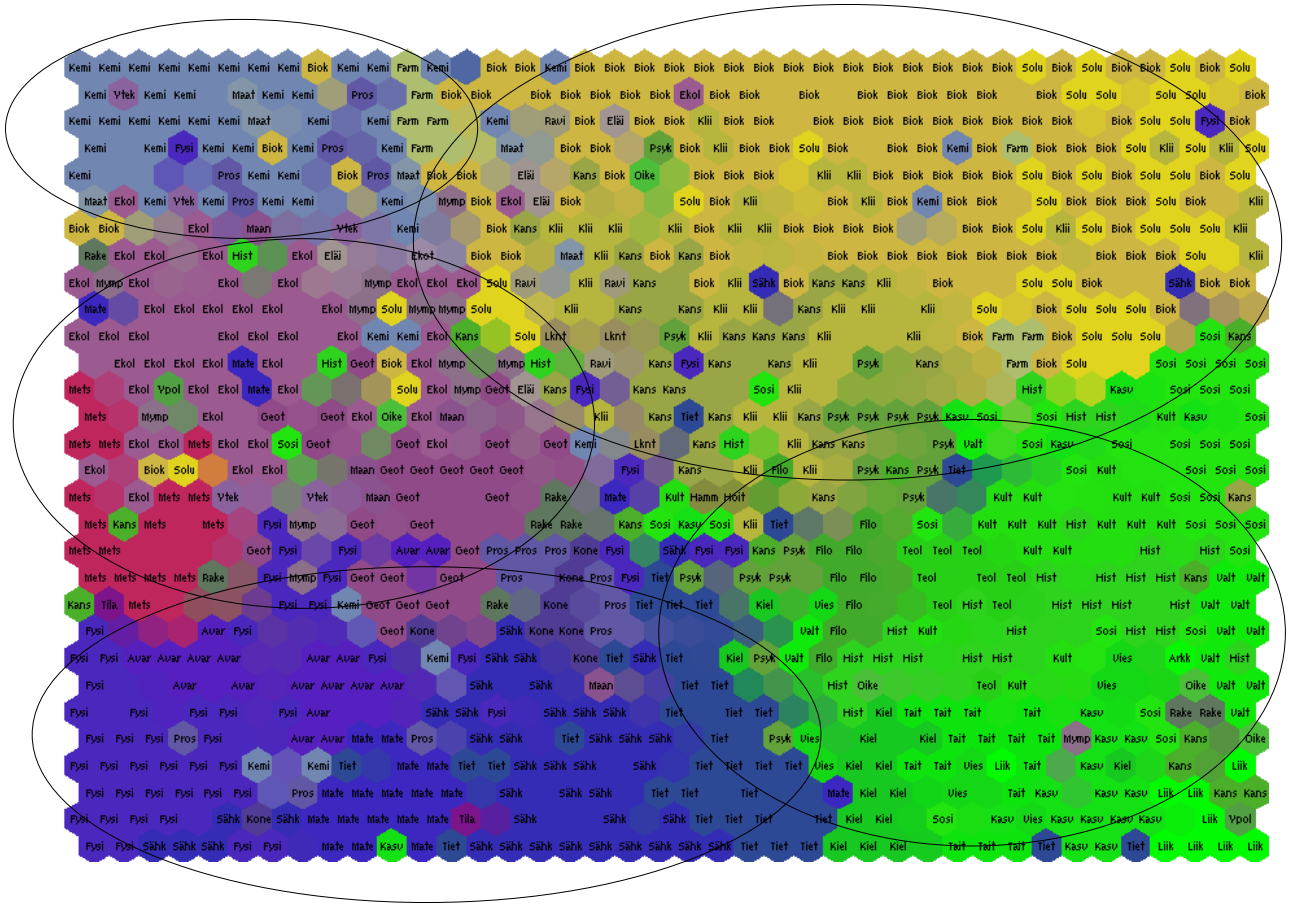
chemistry

biosciences

physics and engineering

medicine

culture and society

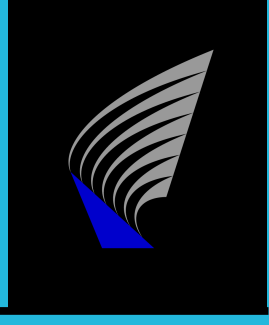


A map of 3224 funding applications to Academy of Finland
(Honkela & Klami, 2007)

WEBSOM: Modes of use

- Search
  - One can find documents in which the search term does not appear but which have similar semantic overall content
  - User is not aware of the near misses when using the traditional search engines
- Exploration
  - The map gives a holistic view on the document collection
- Filtering

Case studies



Word and document maps based on the self-organizing map (SOM)

- **Qualitative analysis using text mining based on the self-organizing map**

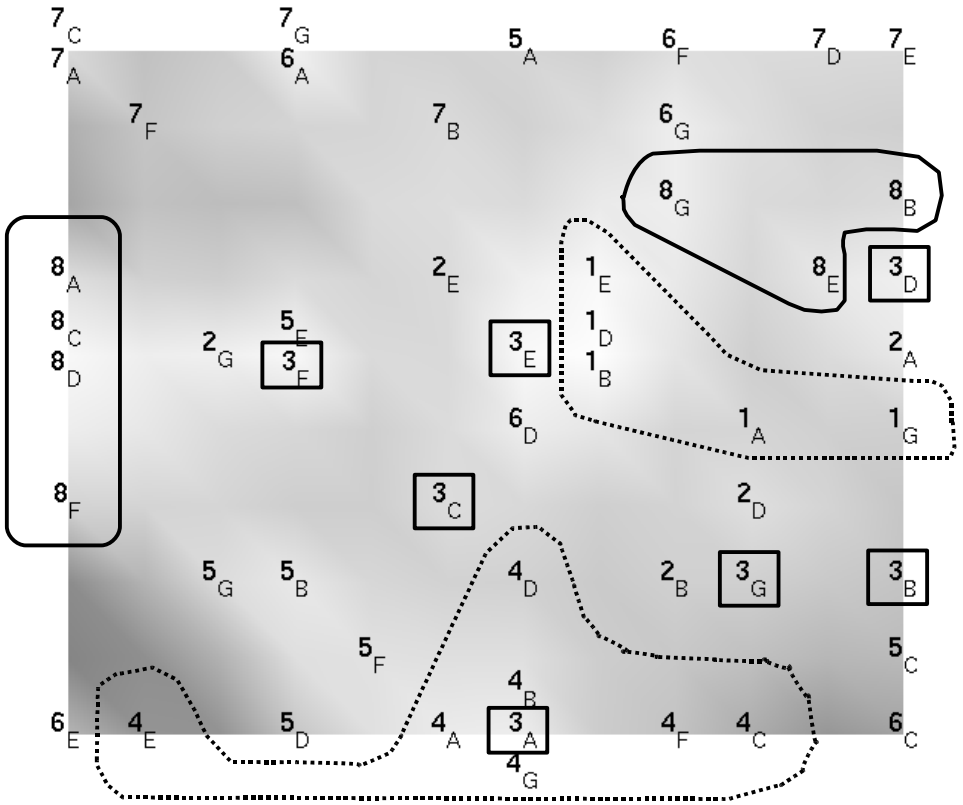
  Emergent word feature models using independent component analysis (ICA)

  Unsupervised morphological modeling using Morfessor and application in speech processing

  MT supported by the SOM and Morfessor

Text mining based on the SOM for qualitative research

(Janasik, Honkela & Bruun, 2008)

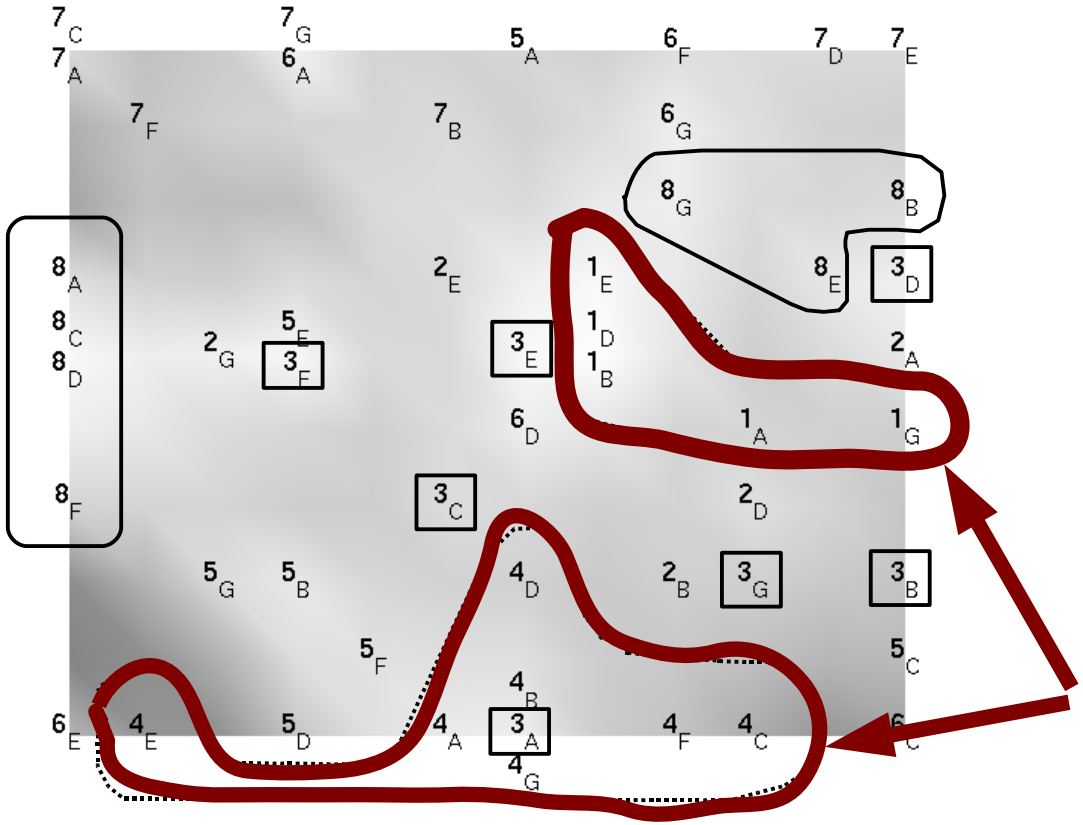


Numbers: projects

Letters: persons

Coherent project views

(Janasik, Honkela & Bruun, 2008)



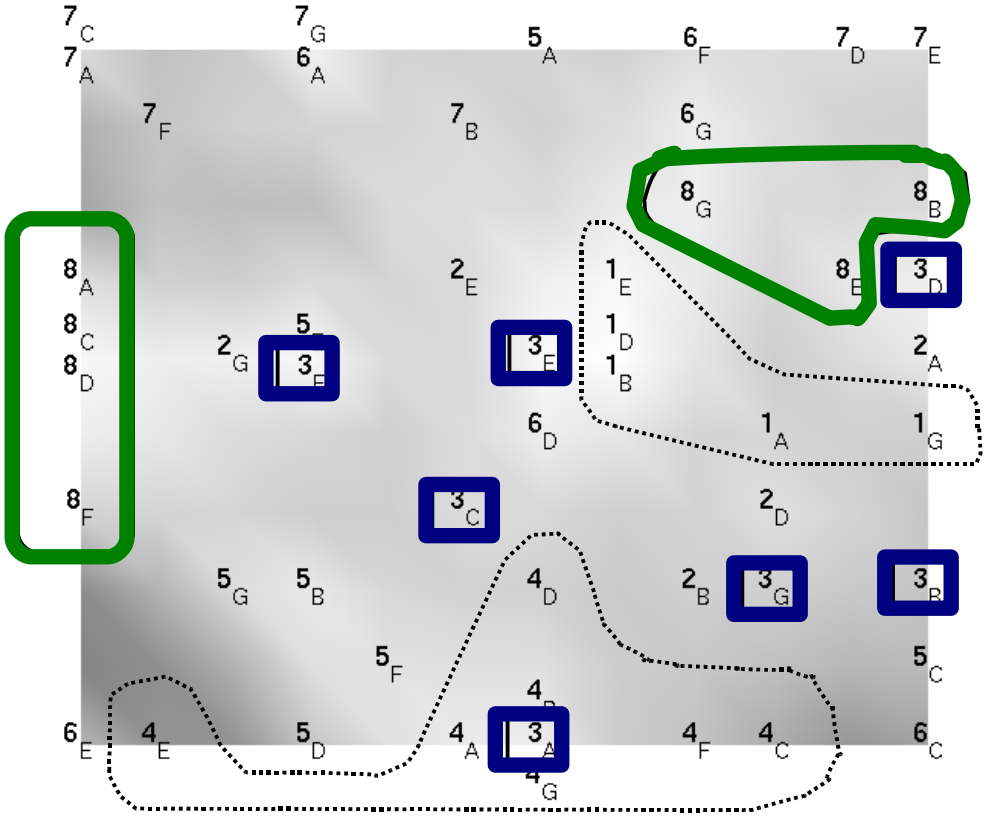
Numbers: projects

Letters: persons

Coherent projects

Scattered project views

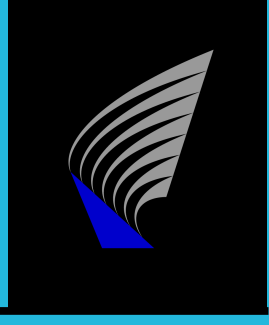
(Janasik, Honkela & Bruun, 2008)



Numbers: projects

Letters: persons

Case studies



Word and document maps based on the self-organizing map (SOM)

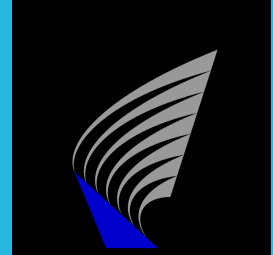
Qualitative analysis using text mining based on the self-organizing map

- **Emergent word feature models using independent component analysis (ICA)**

  Unsupervised morphological modeling using Morfessor and application in speech processing

  MT supported by the SOM and Morfessor

Independent component analysis



The classic version of the ICA model can be expressed as

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

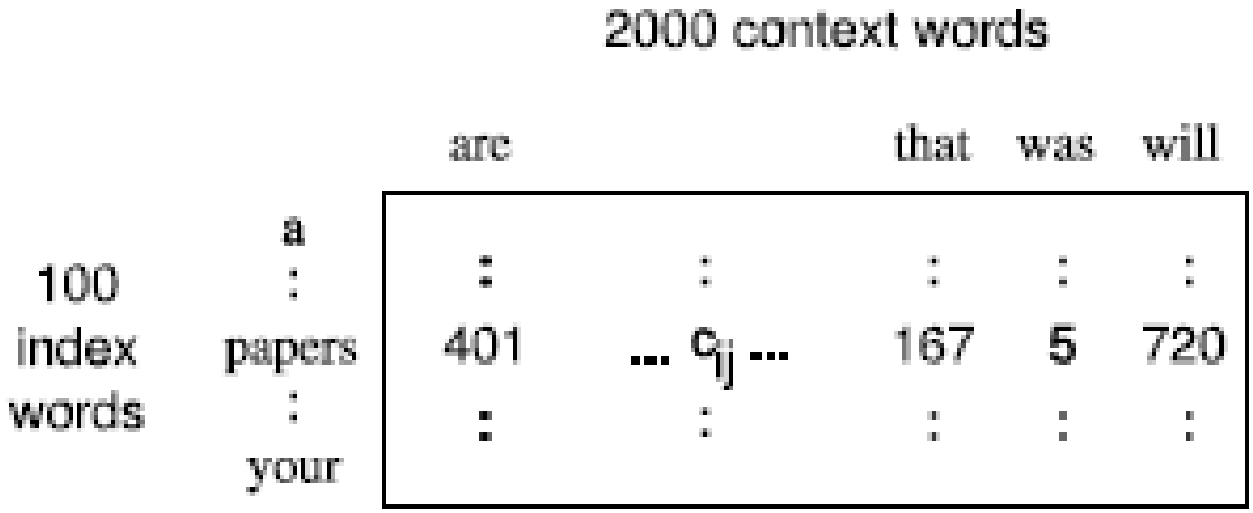
where $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ is the vector of observed random variables, the vector of the independent latent variables is denoted by $\mathbf{s} = (s_1, s_2, \ldots, s_n)^T$ (the "independent components"), and $\mathbf{A}$ is an unknown constant matrix, called the mixing matrix. If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

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

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

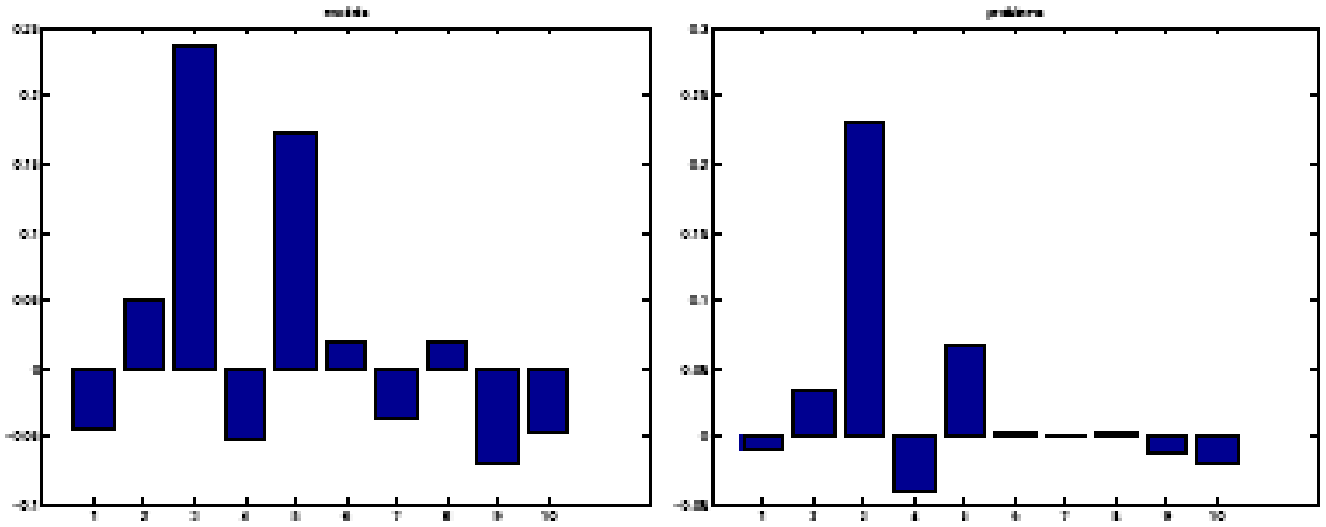


Fig. 4. ICA features for "models" and "problems".

(Honkela & Hyvärinen, 2004)

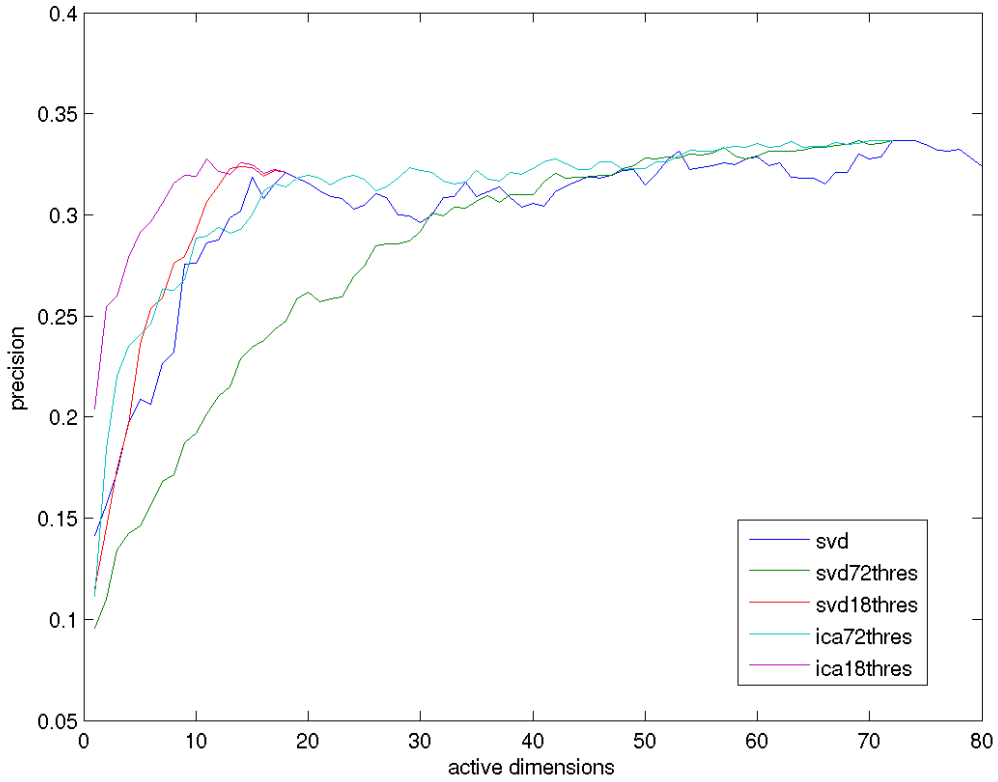
Word ICA

ICA of word contexts; nonlinearity through thresholding

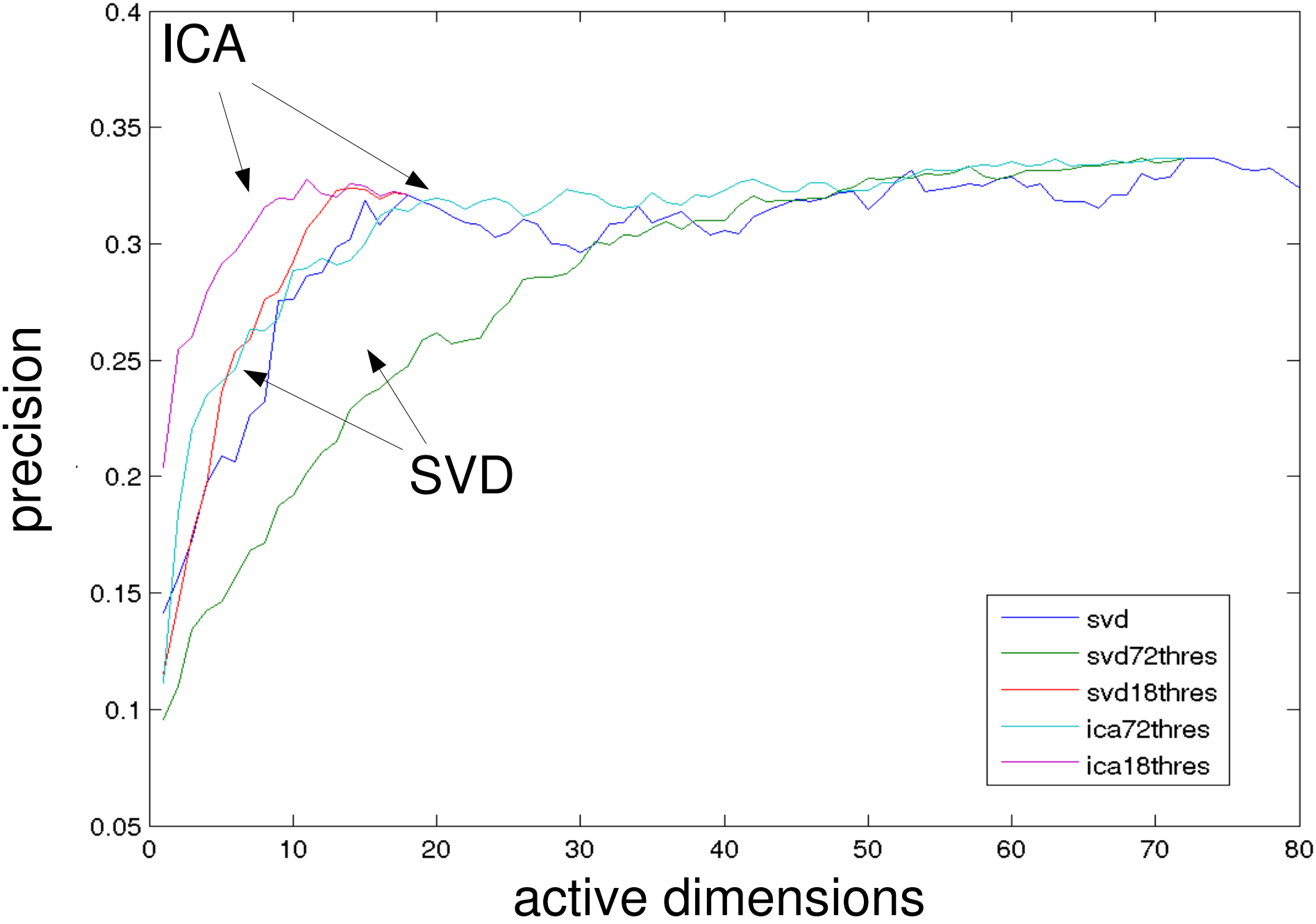
Comparison with SVD/LSA

Effect of sparseness and meaningful emergent components

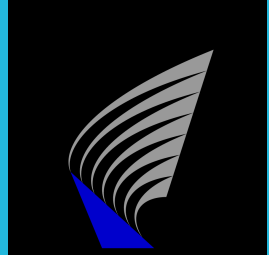
Data: TOEFL tests



(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

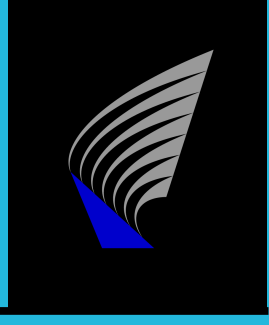
Relation to Koehn's presentation



- **Sparse encoding** of words created **by thresholded word ICA** can be used **as factored representation** of words
- Väyrynen's experiments show that components created by Word ICA coincide reasonably well with traditional linguistic categories. This kind of emergence of explicit meaningful features is not gained with LSA/LSI.

Time for a break ...

Case studies



Word and document maps based on the self-organizing map (SOM)

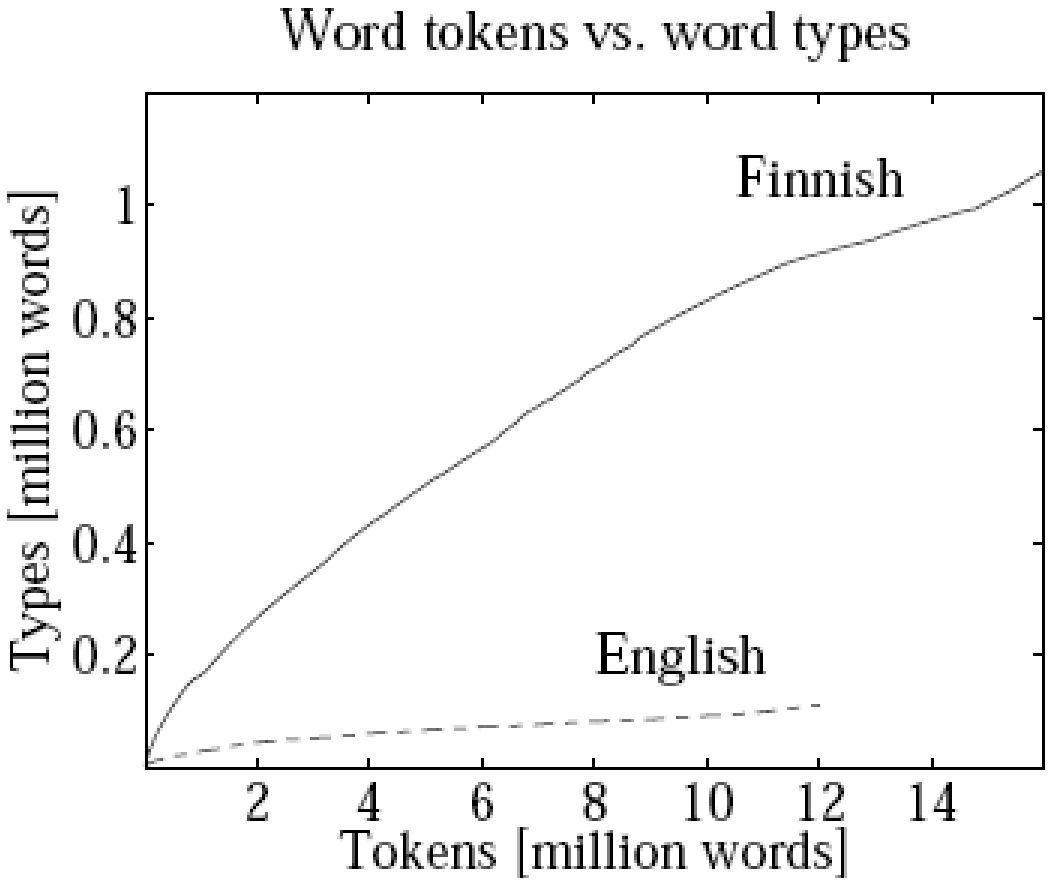
Qualitative analysis using text mining based on the self-organizing map

Emergent word feature models using independent component analysis (ICA)

- **Unsupervised morphological modeling using Morfessor and application in speech processing**

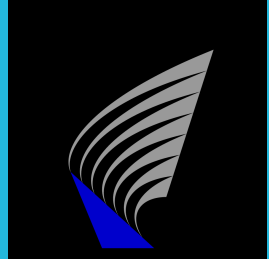
  MT supported by the SOM and Morfessor

Different kinds of languages regarding morphology



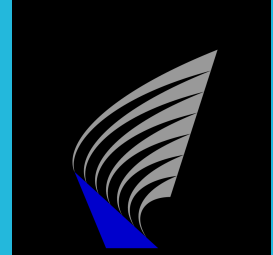
(Creutz, 2006)

Different kinds of languages, cont'd



- In classical morphological typology, the world's languages are characterized by their position on two continua:
  - **isolating** vs. **synthetic**, and
  - **agglutinative** vs. **fusional**

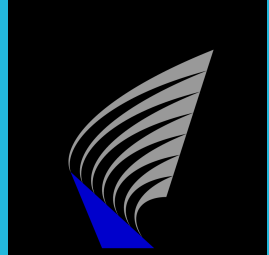
Morfessor



- 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

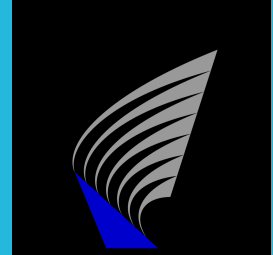
Probabilistic modeling and maximum likelihood (ML) optimization



- 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



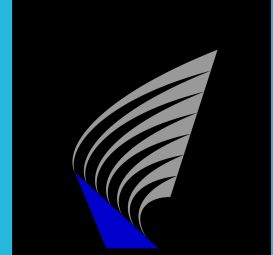
The lexicon is a morph collection in the sense that each space-delimited string is a morph. The probability of the lexicon depends on its size (the number of free parameters in it). Some possible lexicons and their probabilities are (Creutz, 2006):

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



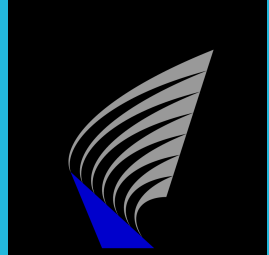
Lexicon 1 "a␣c␣e␣g␣i␣j␣l␣m␣n␣o␣p␣r␣t␣u␣␣", $P(Lexicon\ 1) = (\frac{1}{27})^{29}$

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)

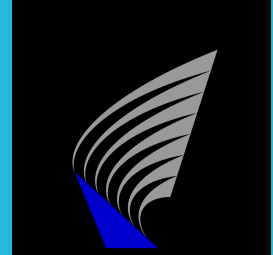
Probabilistic modeling and maximum likelihood (ML) optimization



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

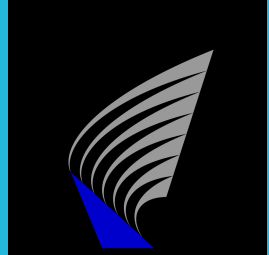


Data | Lexicon 1 "a p p l e # o r a n g e # l e m o n # j u i c e # a p p l e j u i c e # o r a n g e j u i c e # a p p l e t r e e # l e m o n t r e e # #" The sequence consists of 69 morphs and its probability conditioned on *Lexicon 1* is: $P(Data \mid Lexicon\ 1) = (\frac{1}{14+1})^{69} \approx 7.1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

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

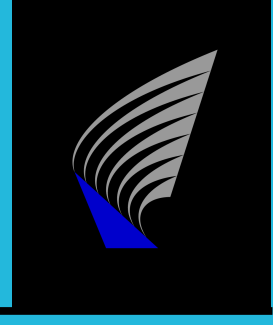


When the lexicons were compared according to their prior probabilities the following ranking was obtained: P(Lexicon 1) > P(Lexicon 2) > P(Lexicon 3)

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



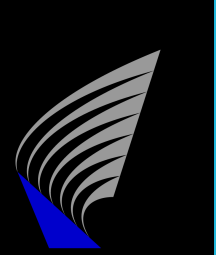
The proposed model selection procedure is based on maximizing the posterior probability of the model, P(Lexicon X |Data).

The posterior can be rewritten using Bayes' rule:

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

(Creutz, 2006)

Model selection



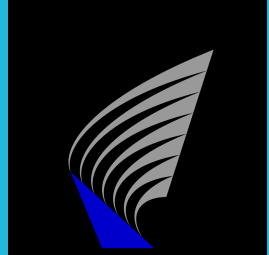
If the intention is to compare different models on the very same data set, the probability of the data is a constant that does not affect the result of the comparison.

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

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

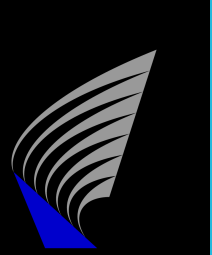
(Creutz, 2006)

Maximum a posteriori (MAP) estimation and model selection



$$\arg\max_{\mathcal{M}} P(\mathcal{M} \mid \mathit{corpus}) = \arg\max_{\mathcal{M}} P(\mathit{corpus} \mid \mathcal{M}) \cdot P(\mathcal{M}), \text{ where}$$
$$P(\mathcal{M}) = P(\mathit{lexicon}, \mathit{grammar}).$$

MAP



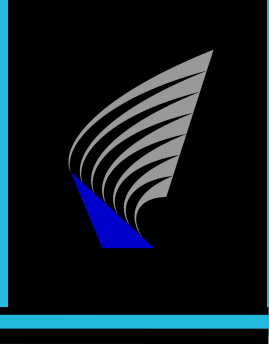
The model selection procedure is based on maximizing the posterior probability of the model $P(\textit{Lexicon } X \,|\, \textit{Data})$. The posterior can be rewritten using the Bayes' rule:

$$\begin{aligned} P(\textit{Lexicon } X \,|\, \textit{Data}) &= \frac{P(\textit{Lexicon } X) \cdot P(\textit{Data} \,|\, \textit{Lexicon } X)}{P(\textit{Data})} \\ &\propto P(\textit{Lexicon } X) \cdot P(\textit{Data} \,|\, \textit{Lexicon } X). \end{aligned}$$

$P(\textit{Lexicon 2}) \cdot P(\textit{Data} \,|\, \textit{Lexicon 2}) \approx 1.9 \cdot 10^{-61} > P(\textit{Lexicon 3}) \cdot P(\textit{Data} \,|\, \textit{Lexicon 3}) \approx 1.0 \cdot 10^{-115} > P(\textit{Lexicon 1}) \cdot P(\textit{Data} \,|\, \textit{Lexicon 1}) \approx 2.2 \cdot 10^{-123}$.

In this comparison the complexity of the model has been balanced against the fit of the training data, which favors a good compromise, that is, a model that does not overlearn and that adequately generalizes to unseen data.

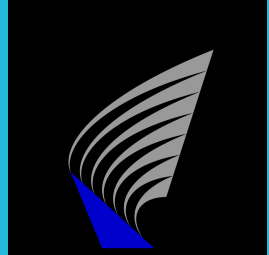
Minimum description length (MDL)
(Rissanen, 1978, etc.)



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

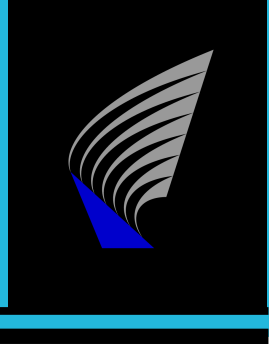


Recursive segmentation. The search for the optimal morph segmentation proceeds recursively. First, the word as a whole is considered to be a morph and added to the codebook. Next, every possible split of the word into two parts is evaluated.

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)

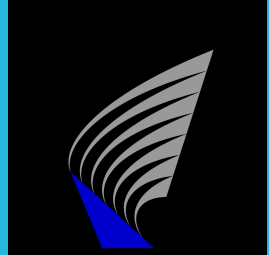
Recursive segmentation and MDL cost



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

(Creutz, 2006)

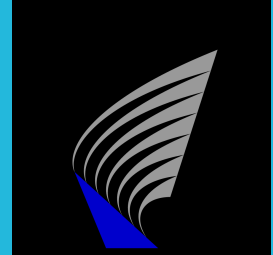
Sequential segmentation



1. Initialize segmentation by splitting words into morphs at random intervals, starting from the beginning of the word. The lengths of intervals are sampled from the Poisson distribution with $\lambda = 5.5$. If the interval is larger than the number of letters in the remaining word segment, the splitting ends.

(Creutz, 2006)

Sequential segmentation



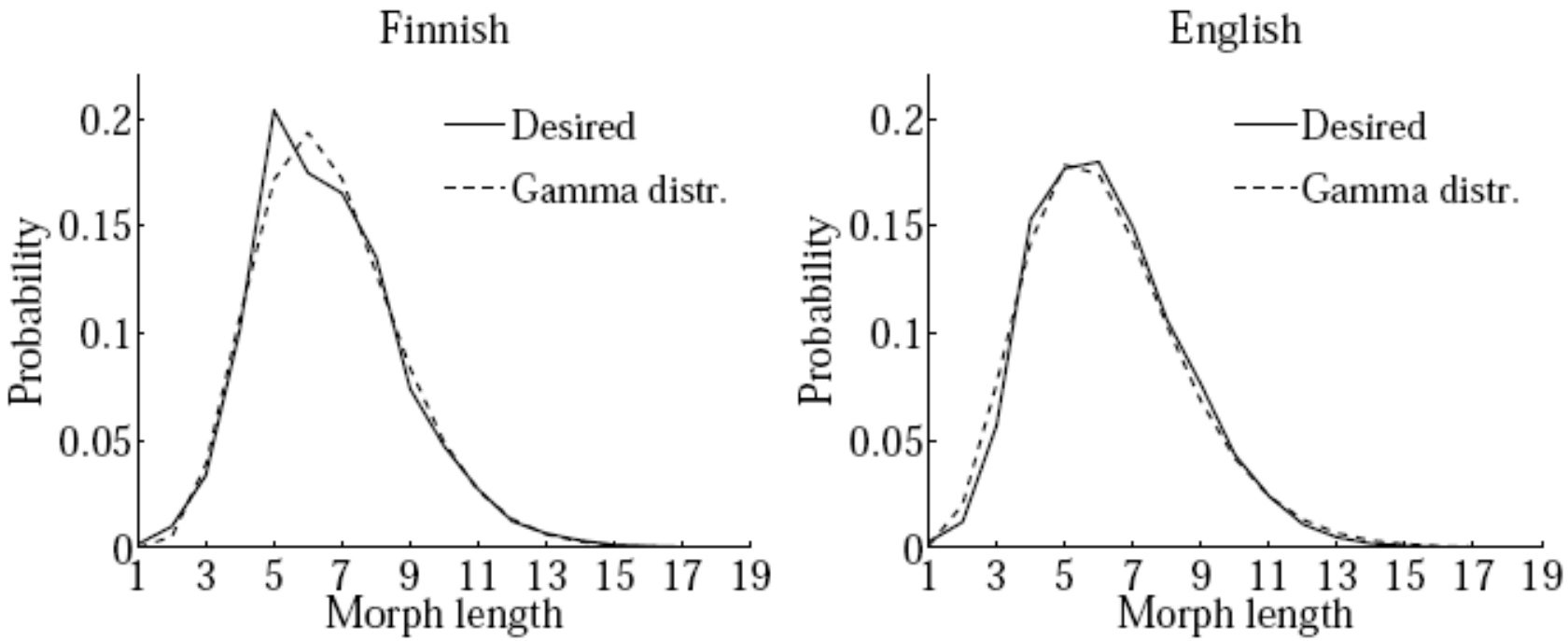
2. Repeat for a number of iterations:
   (a) Estimate morph probabilities for the given splitting.
   (b) Given the current set of morphs and their probabilities, re-segment the text using the Viterbi algorithm for finding the segmentation with lowest cost for each word.
   (c) If not the last iteration: Evaluate the segmentation of a word against rejection criteria. If the proposed segmentation is not accepted, segment this word randomly (as in the Initialization step).

(Creutz, 2006)

Sequential segmentation and ML cost

$$\text{Cost}(\text{Source text}) = \sum_{\text{morph tokens}} -\log p(m_i), \quad (1)$$

Morph length prior



(Creutz, 2006)

Morfessor Categories-ML and HMM

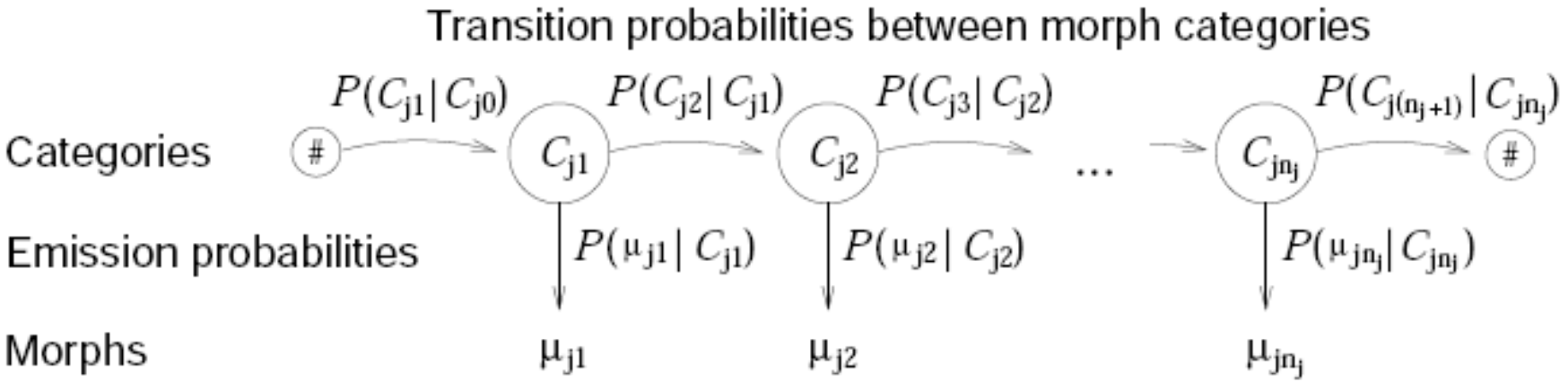
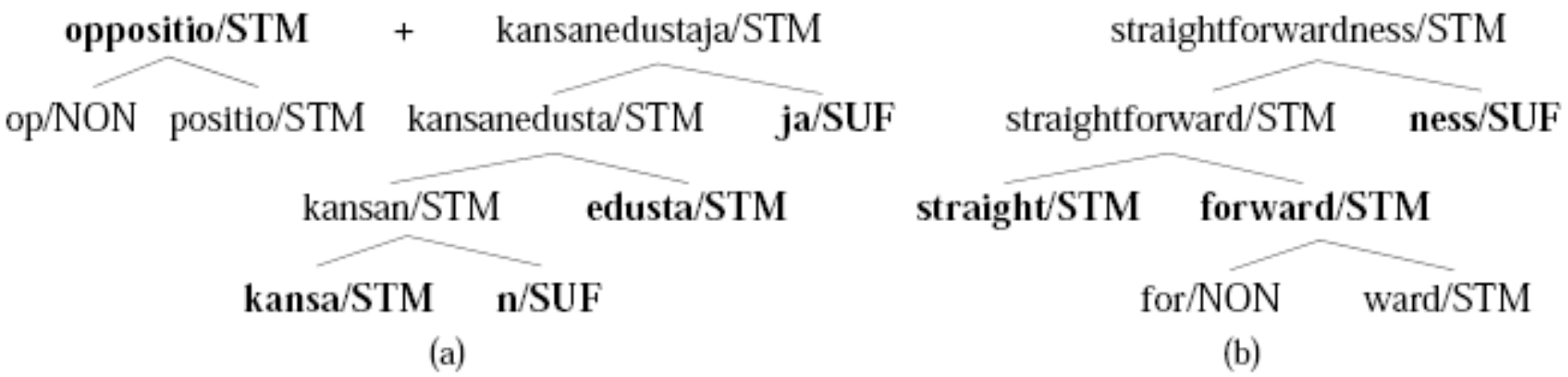


Figure 3.3: The HMM model of a word according to Equation 3.8. The word consists of a sequence of morphs which are emitted from latent categories. For instance, a possible category sequence for the English word "unavailable" would be "prefix + stem + suffix" and the corresponding morphs would be "un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

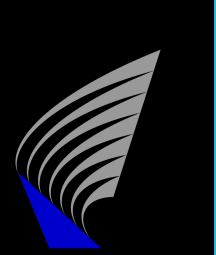
(Creutz, 2006)

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*

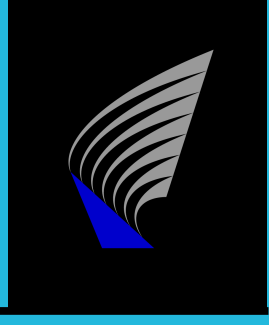
(Creutz, 2006)

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)

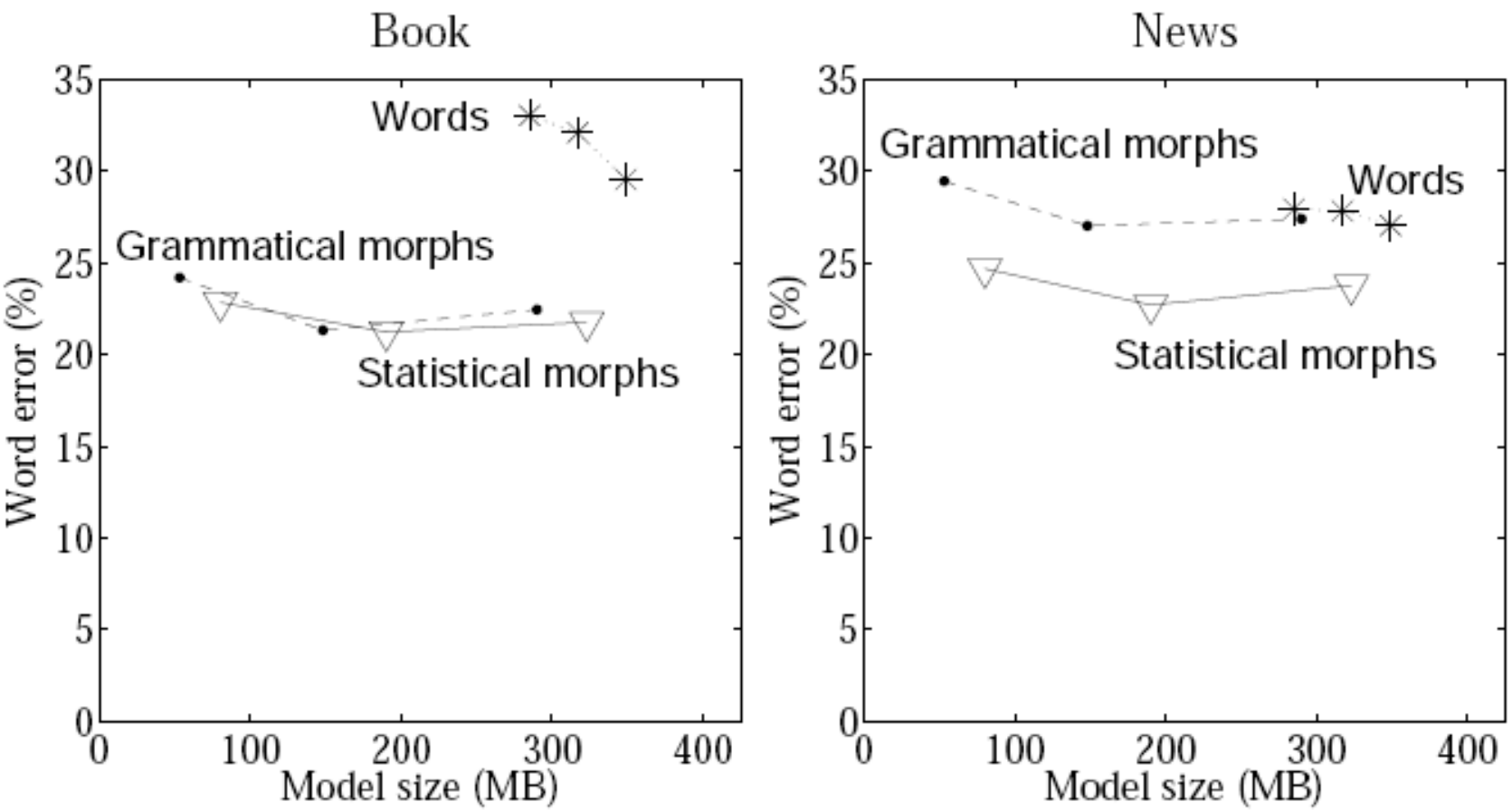
F-measures for Morfessor



Corpus size / Method	Finnish	English	Turkish	Egyptian Arabic
Word tokens	32 000 000	24 000 000	17 000 000	150 000
Word types	1 600 000	170 000	580 000	17 000
Baseline	54.2	66.0	51.3	41.7
Categories-ML	67.0	69.0	69.2	67.9
Categories-MAP	66.4	66.2	70.7	68.1

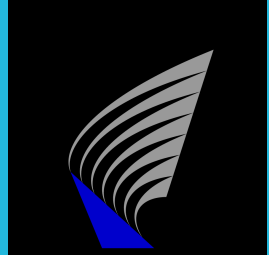
(Creutz, 2006)

Applying Morfessor in speech recognition



(Creutz, 2006)

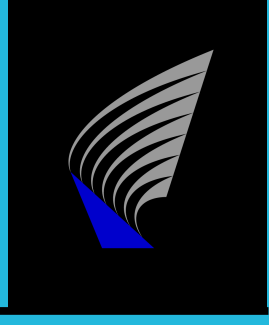
Recognition of an audio book



n	Statistical morphs			Grammatical morphs			Words		
	size [MB]	H [bits]	WER [%]	size [MB]	H [bits]	WER [%]	size [MB]	H [bits]	WER [%]
2	19	16.11	–	14	16.71	–	241	16.15	–
3	80	14.95	22.85	53	15.10	24.20	285	15.59	33.04
4	190	14.41	21.24	148	14.51	21.33	317	15.07	32.11
5	323	14.40	21.76	290	14.36	22.46	349	14.67	29.55
6	441	14.41	–	445	14.37	–	385	14.45	–
7	538	14.43	–	545	14.42	–	422	14.39	–

(Creutz, 2006)

Case studies



Word and document maps based on the self-organizing map (SOM)

Qualitative analysis using text mining based on the self-organizing map

Emergent word feature models using independent component analysis (ICA)

Unsupervised morphological modeling using Morfessor and application in speech processing

- **SMT supported by Morfessor and the SOM**

Motivation

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

Translation model

		Scores			
Source language (f)	Target language (e)	phi(e|f)	lex(e|f)	phi(f|e)	lex(f|e)
...					
göteborgin pöytäkirja	göteborgsprotokollet	0.3333	0.0496	0.7500	0.0215
göteborgin pöytäkirja	protokollet från götebörg	1.0000	0.0532	0.2500	0.0050
...					
göteborgissa	från göteborg	0.3571	0.1478	0.0270	0.0008
göteborgissa	göteborg	0.0952	0.2955	0.1297	0.4941
göteborgissa	i göteborg ,	0.2308	0.1479	0.0162	0.0049

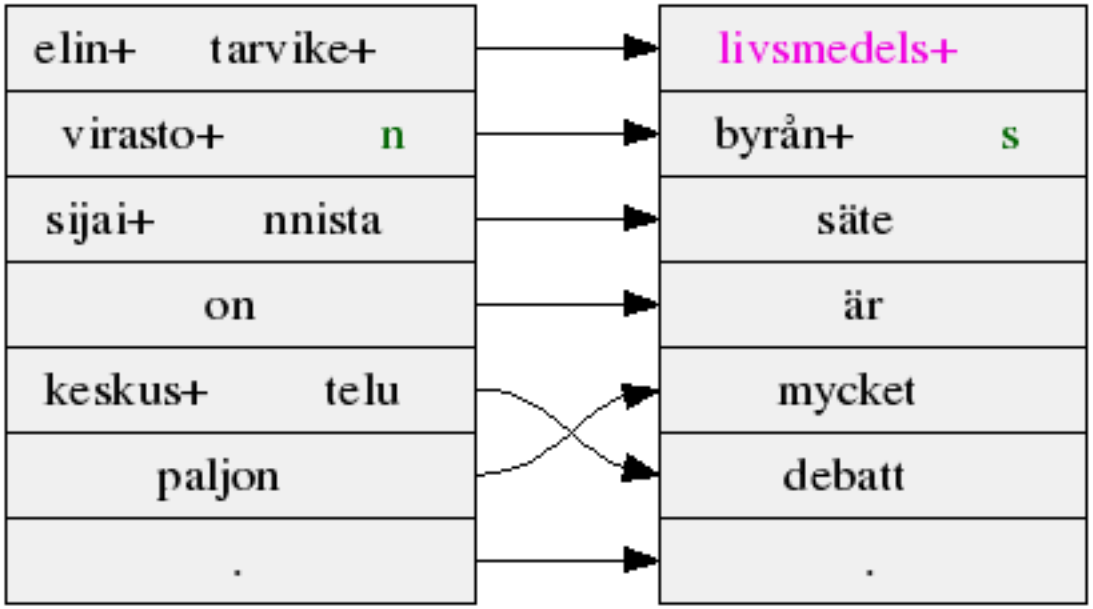
- Scores are estimated from the phrase aligned Europarl corpus

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

Example

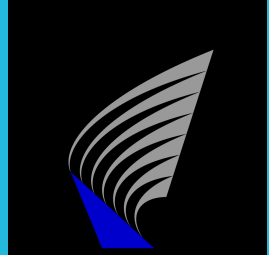
Elintarvikeviraston sijainnista on keskustelu paljon.



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)

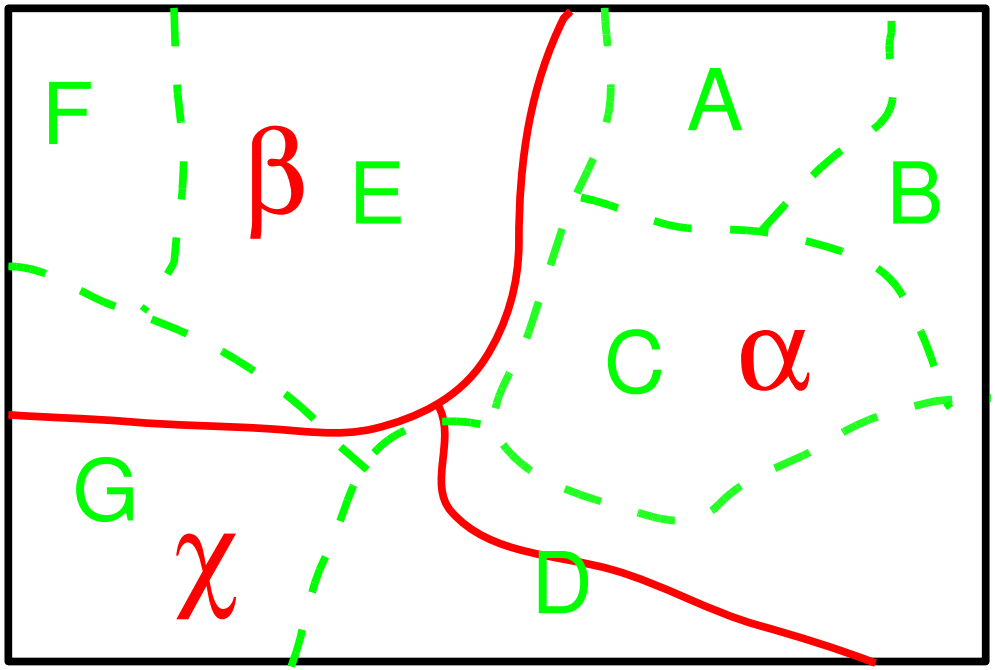
Effect of Morfessor on OOV words

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.

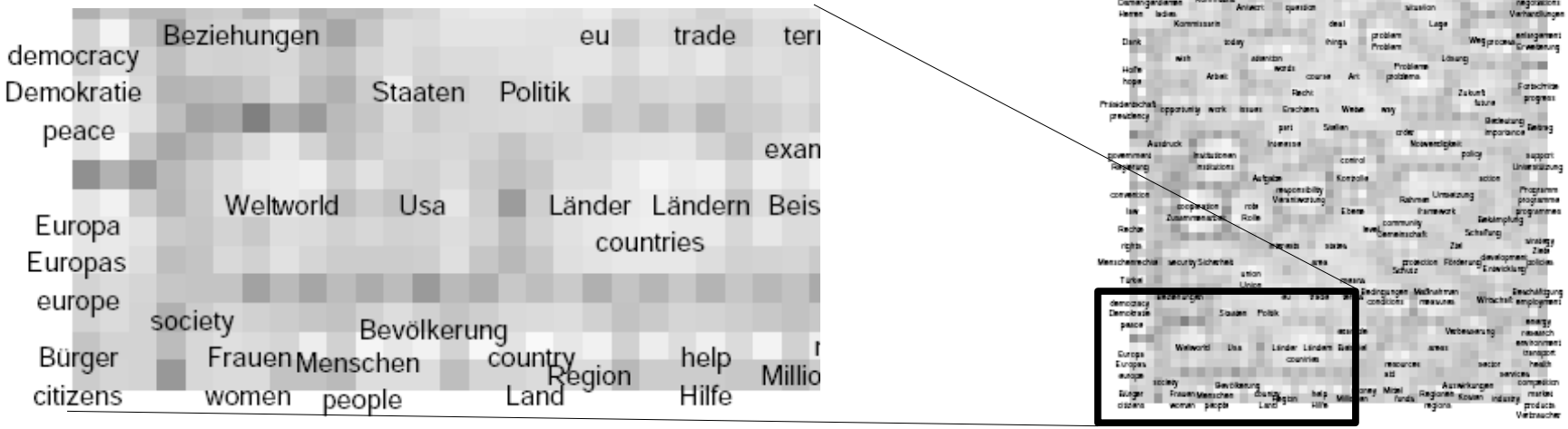
(Virpioja, Väyrynen, Creutz, Sadeniemi, 2007)

Conceptual territories



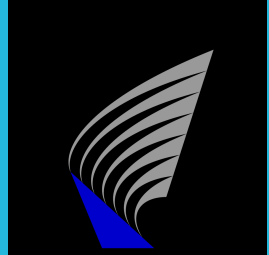
Each language divides the conceptual space differently
(cf. e.g. research by M. Bowerman)

Multilingual map of words
(based on Europarl data)



(Honkela, Virpioja & Väyrynen, 2008)

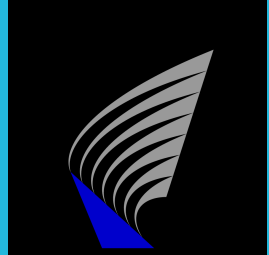
Experiment in perception-based translation



- 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'

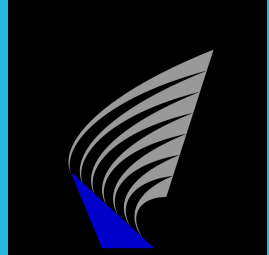
(Sjöberg, Viitaniemi, Laaksonen & Honkela, 2006)

Agenda



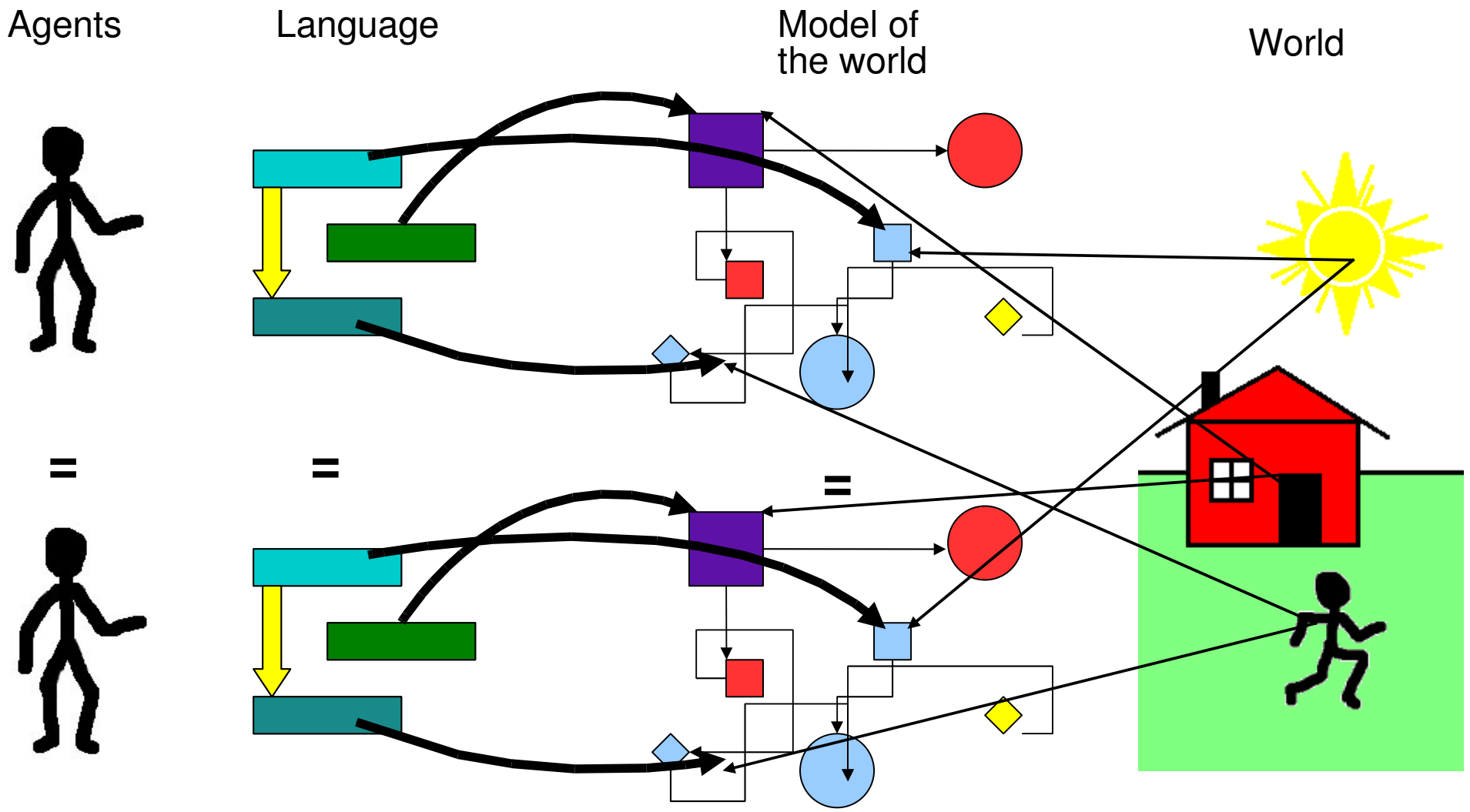
- Learning paradigms
- Philosophical and practical motivation for unsupervised learning
- Overview of unsupervised learning methods for language modeling
- Case studies
- **Additional concerns, conclusions and discussion**

Subjectivity of Meaning



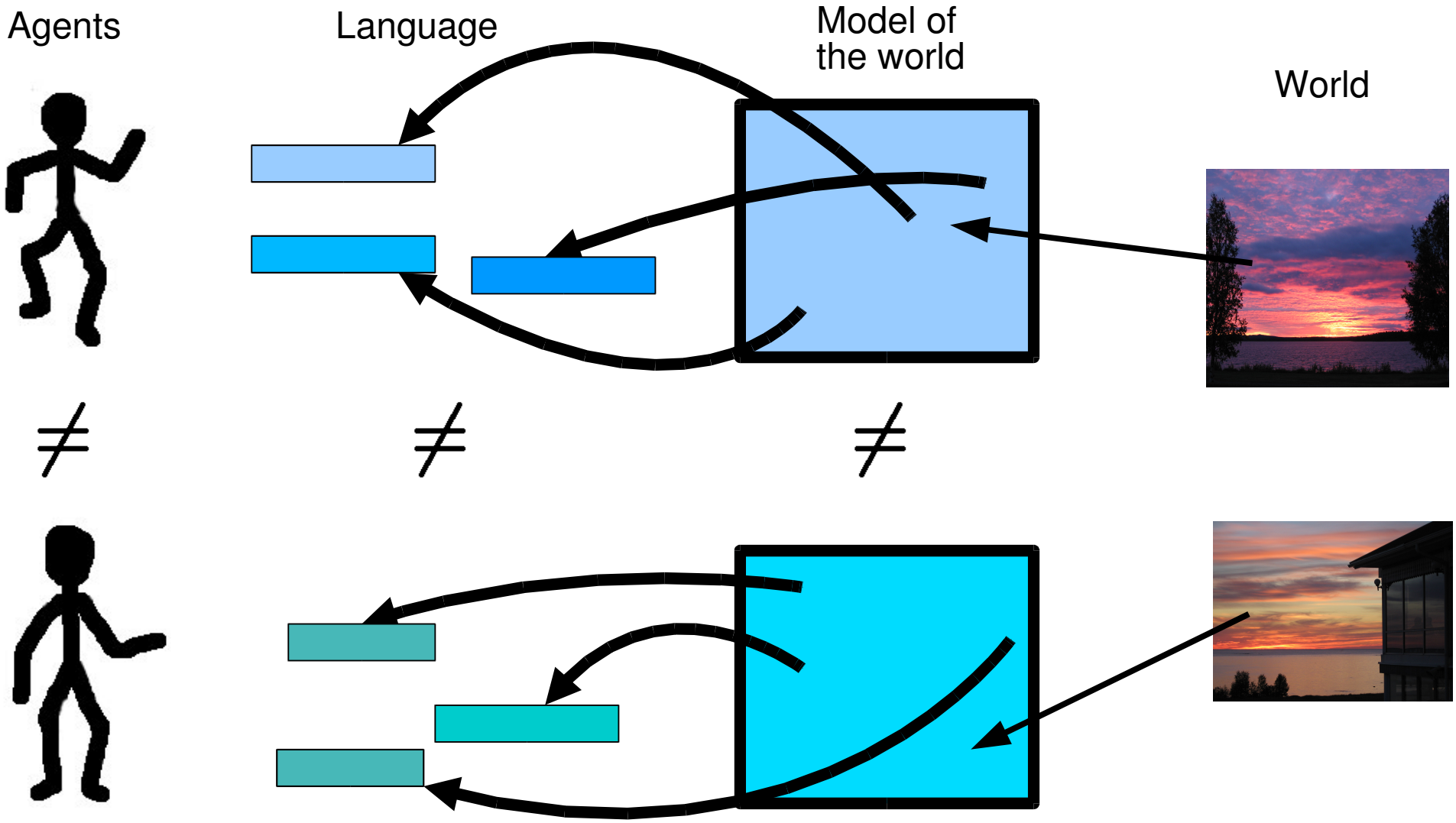
- Almost all formal or computational theories of meaning are based on the assumption that meanings are shared
- However, this does not appear to be empirically true

Traditional formal semantics revisited



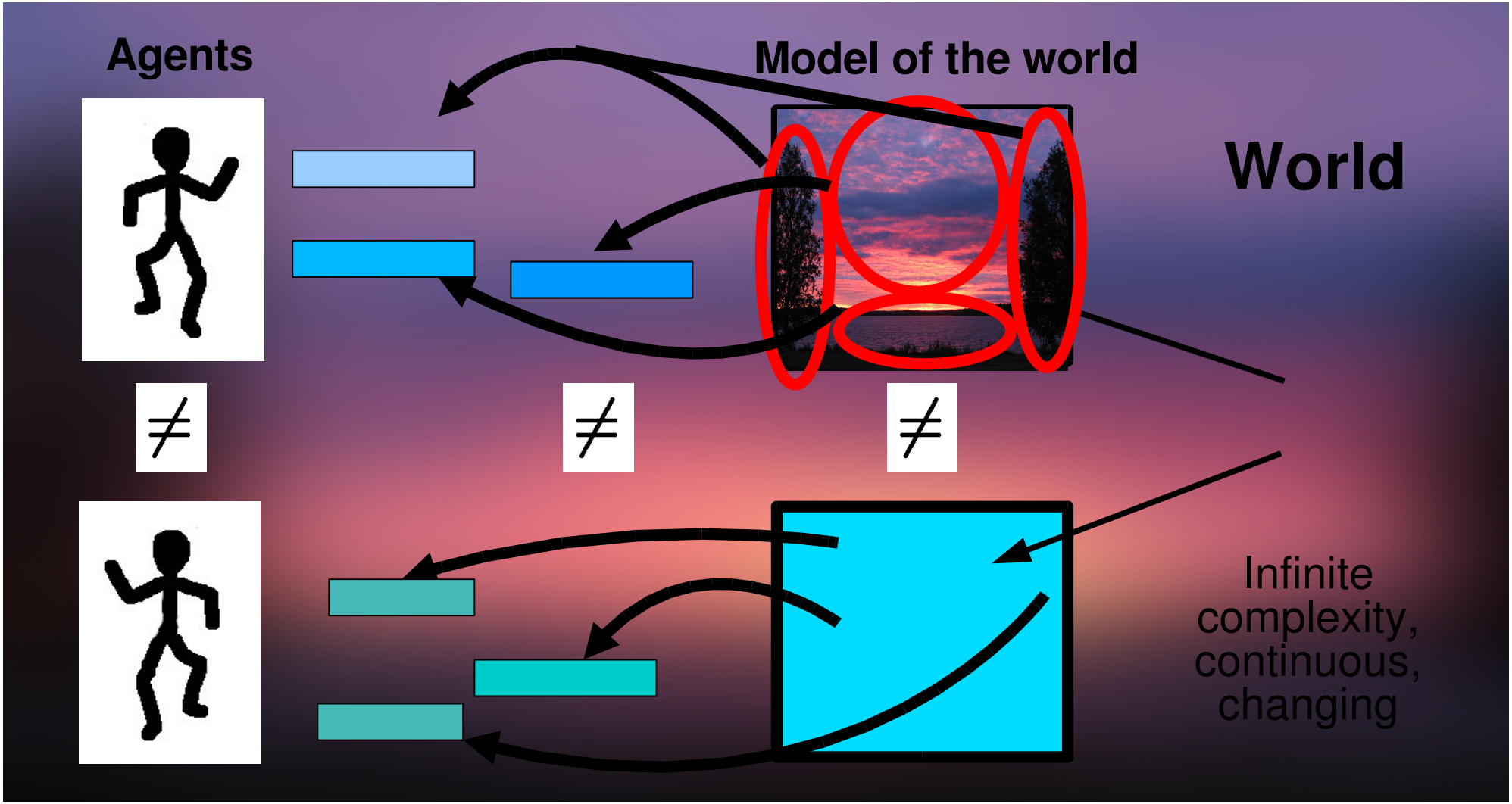
Emergentist viewpoint

(importance of pattern recognition and learning)



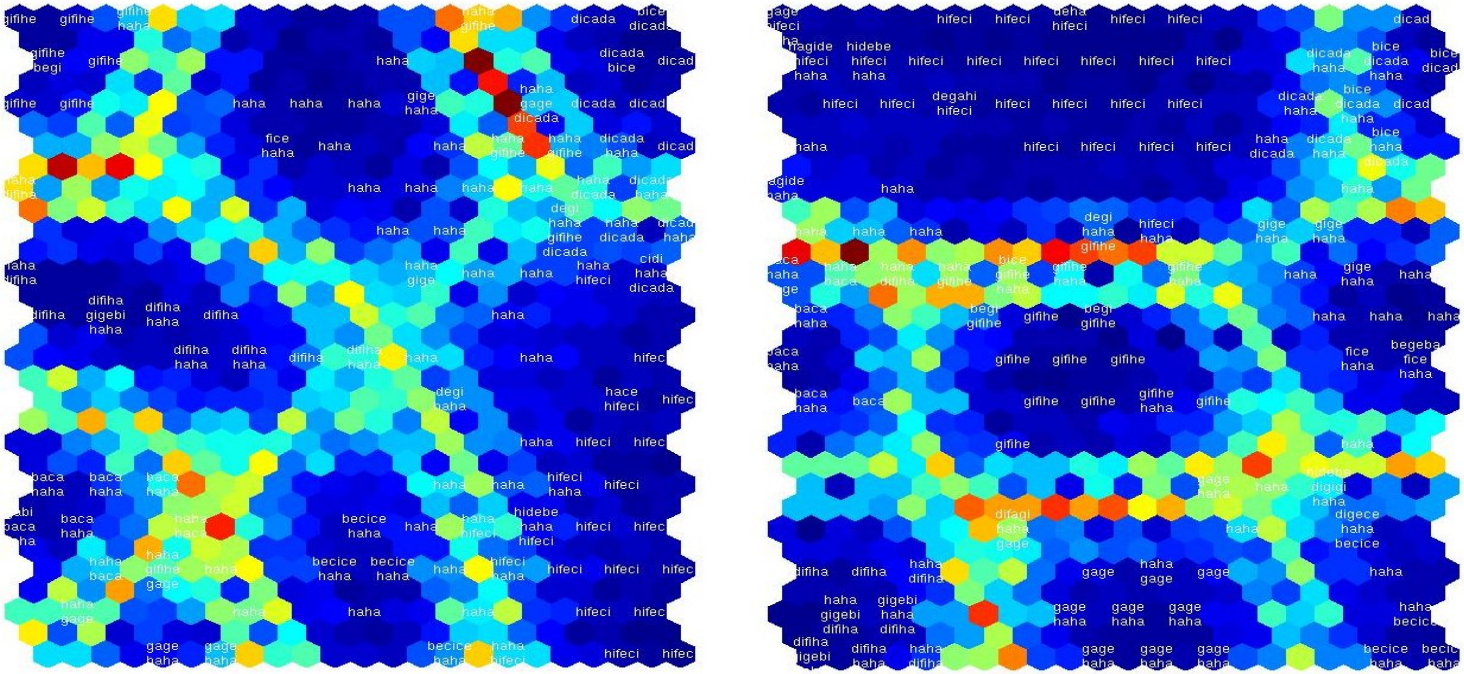
Emergentist viewpoint

(emphasis on constructivism)



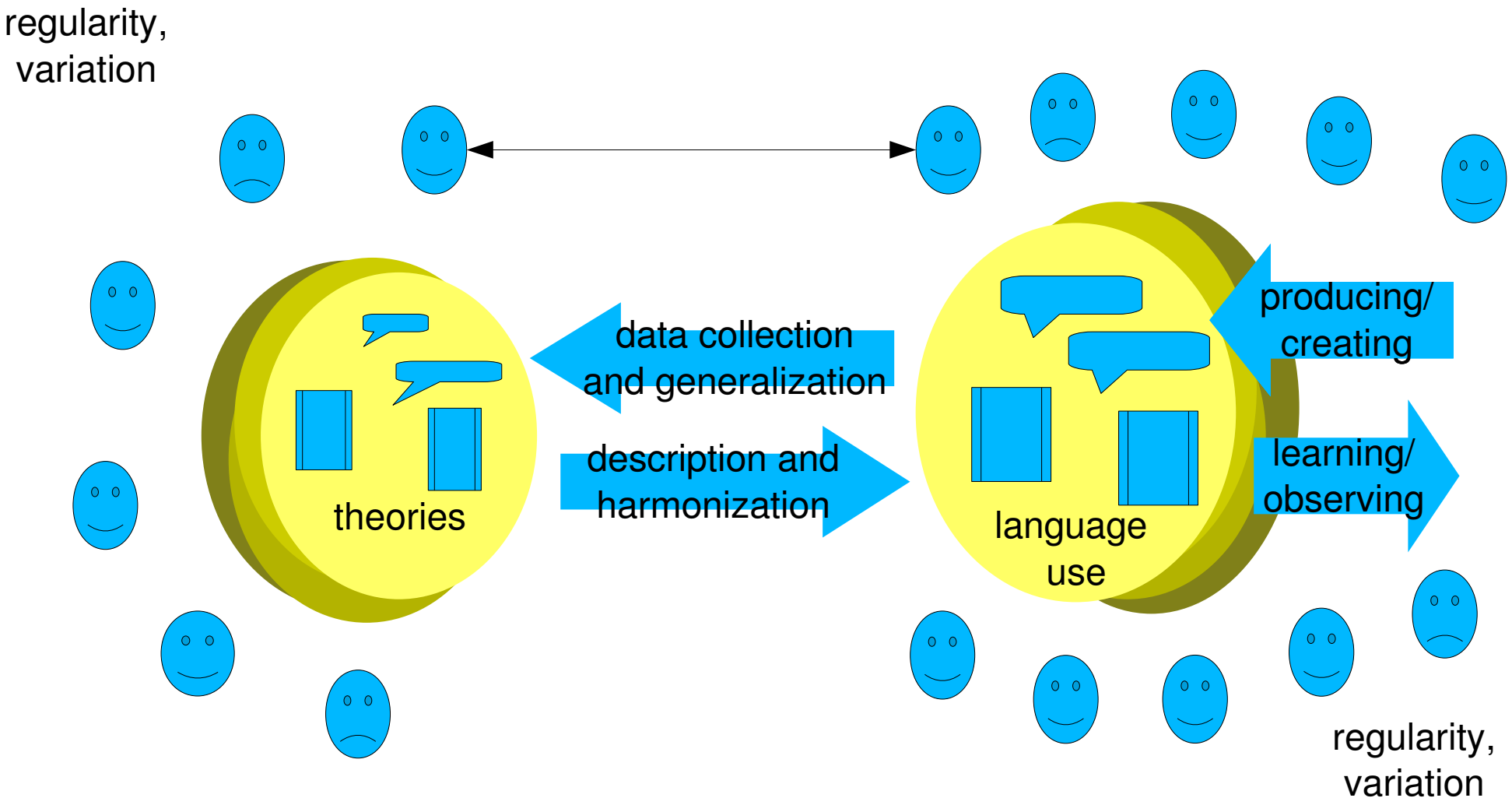
Example: concept of computation

Simulating language emergence and meaning negotiations

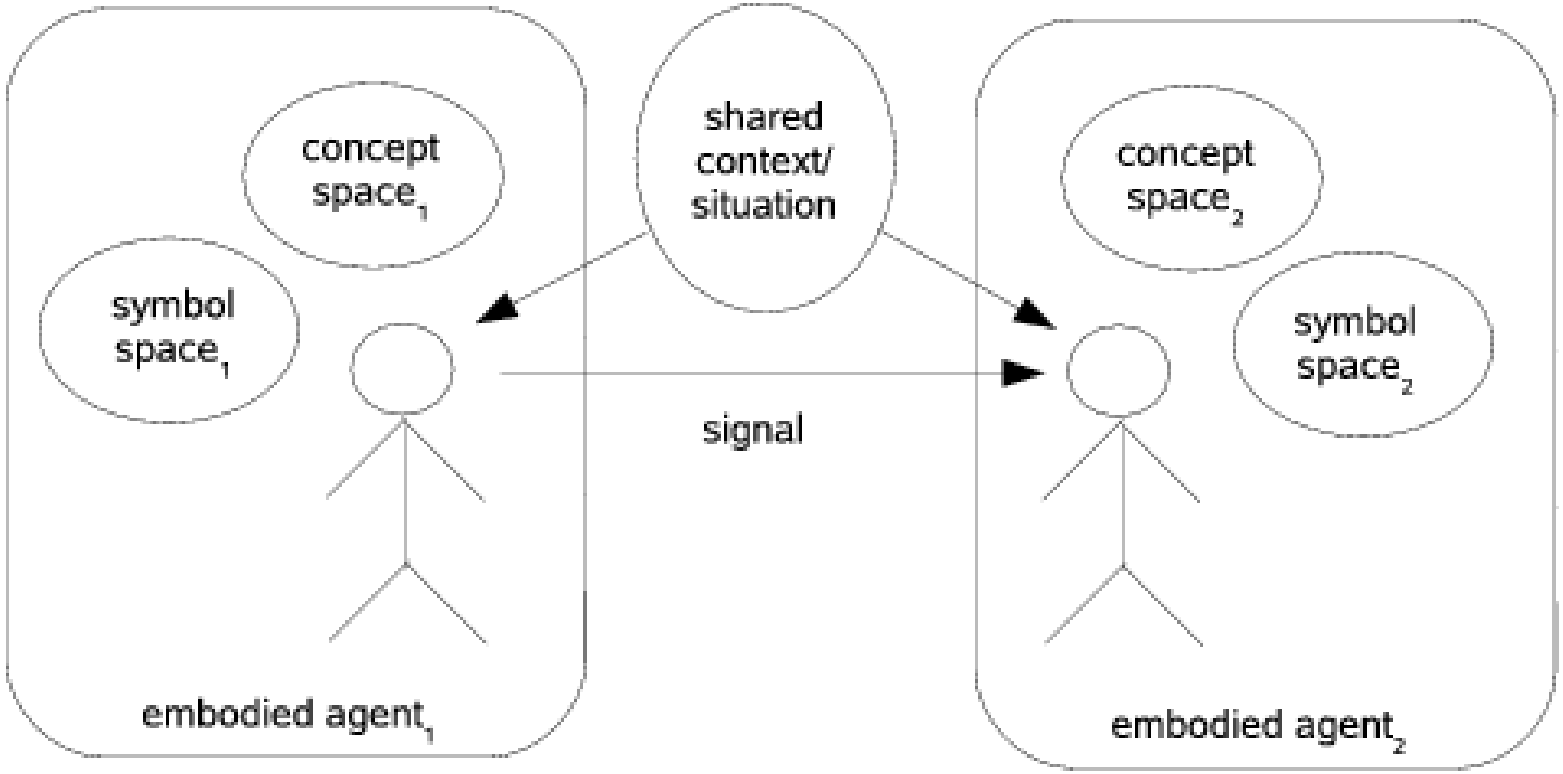


Emergence of a coherent lexicon in
a community of interacting SOM-based agents
(Lindh-Knuutila, Lagus & Honkela, 2006)

Language use and linguistic theory formation as social phenomena



Not enough time to go into the details in the Theory of Intersubjective Meaning Spaces (IMS)



Thank you!