

LUONNOLLINEN KIELI JA TEKOÄLYN KOGNITIO

Mathias Creutz
kieliteknologian yliopistonlehtori

Nykykielten laitos
Humanistinen tiedekunta
Helsingin yliopisto

4.5.2017



KIELITIEDE JA KIELITEKNOLOGIA HELSINGIN YLIOPISTOSSA

Professori Jörg Tiedemann (kielitekniologia)

Professori Matti Miestamo (yleinen kielitiede)

Professori Timo Honkela (digitaaliset ihmistieteet)

Tutkimusjohtaja Krister Lindén (FINCLARIN)





LUONNOLLISEN KIELEN KÄSITTELY KÄÄNTÄMISENÄ TAI MUUNNOKSENA

- Konekääntäminen
- Automaattinen puheentunnistus
- Puhesynteesi
- Tekstintunnistus (OCR – optical character recognition)
- Älykäs tekstinsyöttö
- Puheen kääntäminen

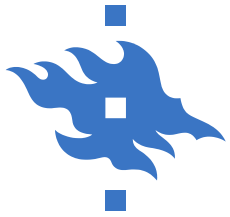
Tuotetaan kieltä.
Kielen sekventiaalinen luonne olennainen.



LUONNOLLISEN KIELEN KÄSITTELY TIEDONLOUHINTANA

- Hakukoneet
- Sentimenttianalyysi
- Erisnimien tunnistus
- “Nousussa olevien sanojen tunnistus” (trending words and concepts)
- Kuvien automaattinen annotointi

Tuotetaan “tietoa”, joka voi olla ei-kielellistä.
Kielen sekventiaalinen luonne ei välttämättä yhtä olennainen.



KIELITIETEEN OSA-ALUEET

Pragmatiikka

Semantiikka

Syntaksi

Morfologia

Fonologia

Fonetiikka



KIELITIETEEN OSA-ALUEET

Pragmatiikka

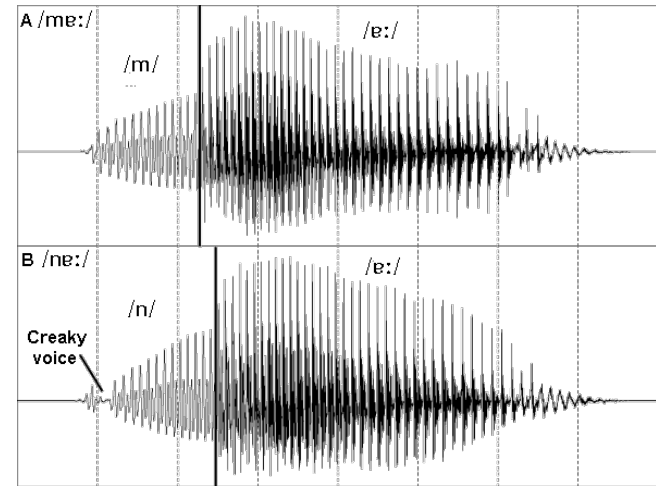
Semantiikka

Syntaksi

Morfologia

Fonologia

Fonetiikka





KIELITIETEEN OSA-ALUEET

Pragmatiikka

Semantiikka

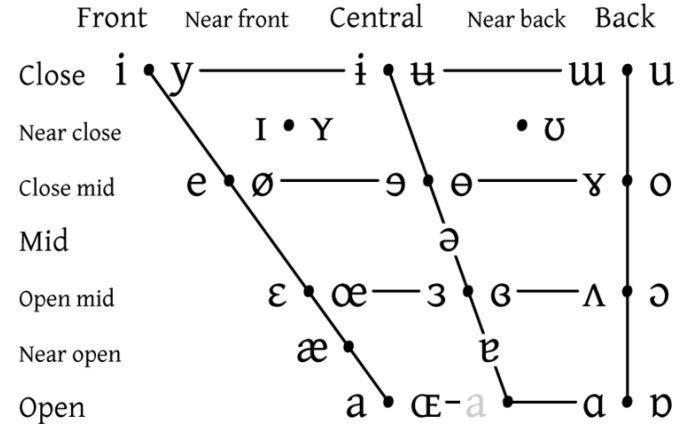
Syntaksi

Morfologia

Fonologia

Fonetiikka

VOWELS



Vowels at right & left of bullets are rounded & unrounded.



KIELITIETEEN OSA-ALUEET

Pragmatiikka

Semantiikka

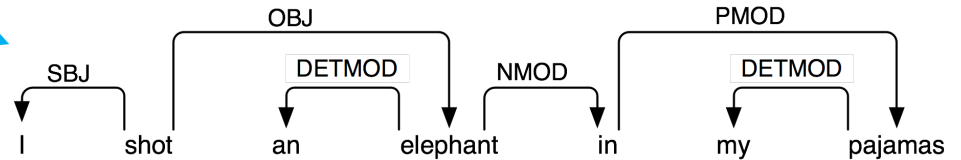
Syntaksi

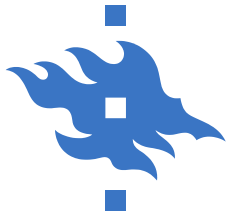
Morfologia

Fonologia

Fonetiikka

Steven Bird, Ewan Klein, Edward Loper: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit





KIELITIETEEN OSA-ALUEET

Pragmatiikka

Semantiikka

Syntaksi

Morfologia

Fonologia

Fonetiikka

lautasiltammeko

lauta#silta N Gen Sg PxPI1 Foc_kO

lauta#silta N Nom PI PxPI1 Foc_kO

lauta#silta N Nom Sg PxPI1 Foc_kO

lautanen#ilta N Nom Sg PxPI1 Foc_kO

lautanen N Abl PI PxPI1 Foc_kO



KIELITIETEEN OSA-ALUEET



x	y
Angus(x)	
dog(y)	
own(x,y)	

Angus owns a dog



x	y	u	z
Angus(x)			
dog(y)			
own(x,y)			
	u = y		
Irene(z)			
bite(u,z)			

*Angus owns a dog.
It bit Irene*

Steven Bird, Ewan Klein, Edward Loper: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.



KIELITIETEEN OSA-ALUEET

Pragmatiikka

Semantiikka

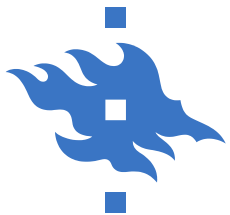
Syntaksi

Morfologia

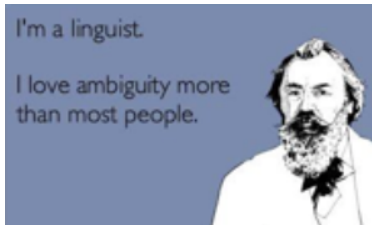
Fonologia

Fonetiikka

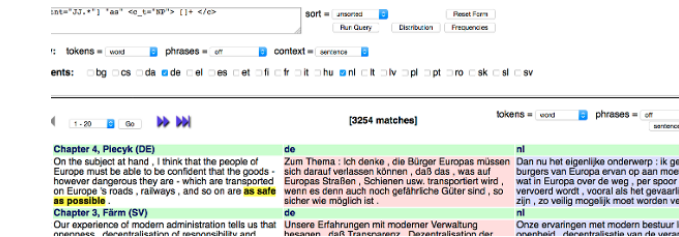




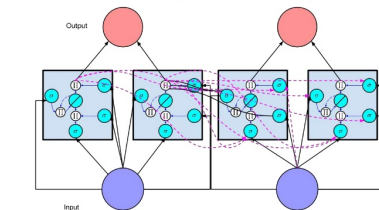
ALAN KEHITYKSESTÄ



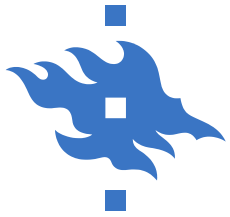
asiantuntijoiden
laatimat säännöt
ja ohjelmat
(1900-luku)



tilastolliset
järjestelmät
(2000-luku)



takaisinkytketyt
LSTM-neuroverkot
(2010-luku)



KUINKA PALJON MERKITYSTÄ TARVITAAN KIELEN KÄÄNTÄMISESSÄ?

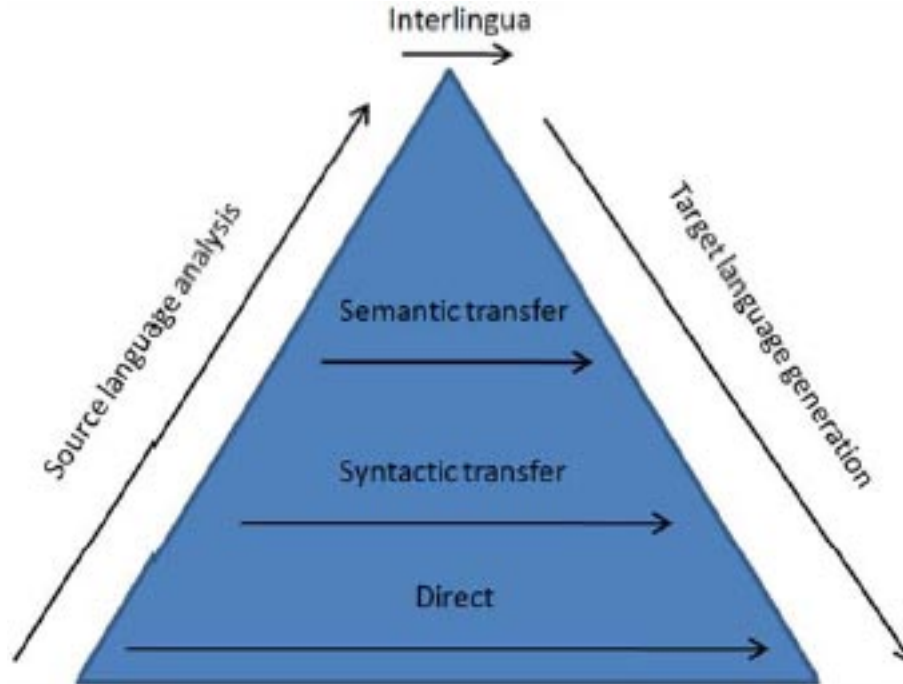


Figure 1: The Vauquois triangle



KUINKA PALJON MERKITYSTÄ TARVITAAN KIELEN KÄÄNTÄMISESSÄ?

tanska:

Dronningens fødselsdag kulminerer med middag torsdag aften, og gæsterne forventer en sjov, løssluppen fest.

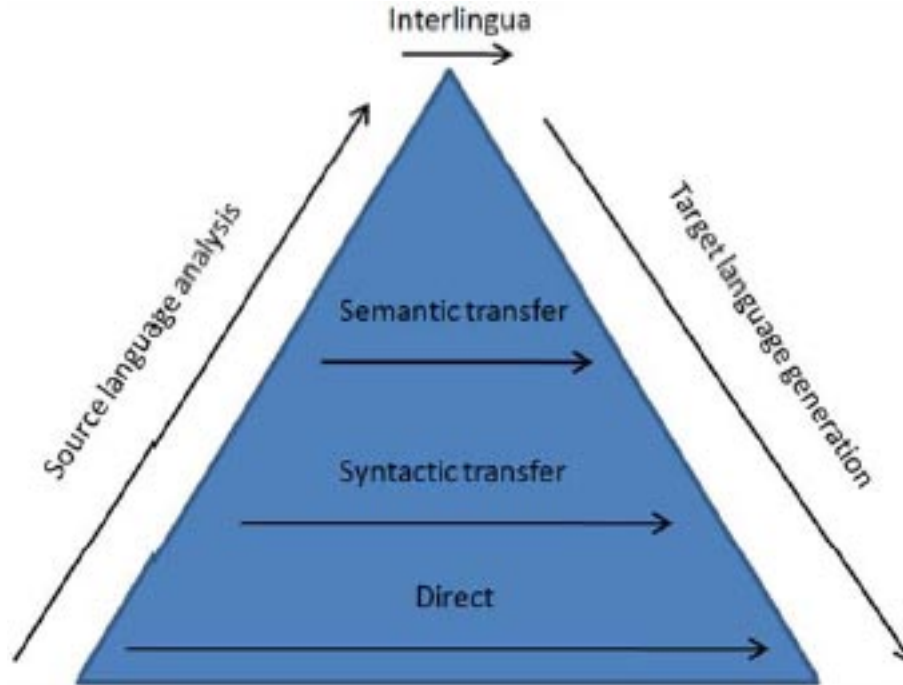
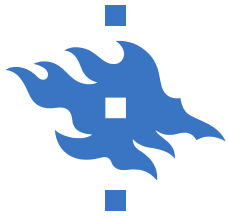


Figure 1: The Vauquois triangle

ruotsi:

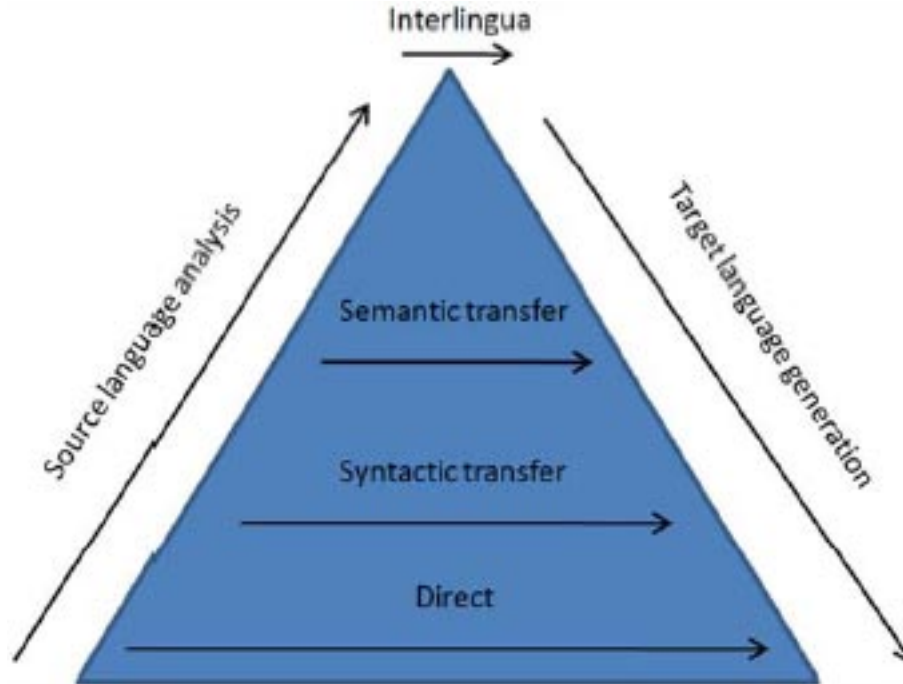
Drottningens födselsdag kulminerar i en middag torsdag kväll, och gästerna förväntar sig en rolig, lössläppt fest.



KUINKA PALJON MERKITYSTÄ TARVITAAN KIELEN KÄÄNTÄMISESSÄ?

mandariinikiina:

“Although north wind howls, but sky still very clear.”



englanti:

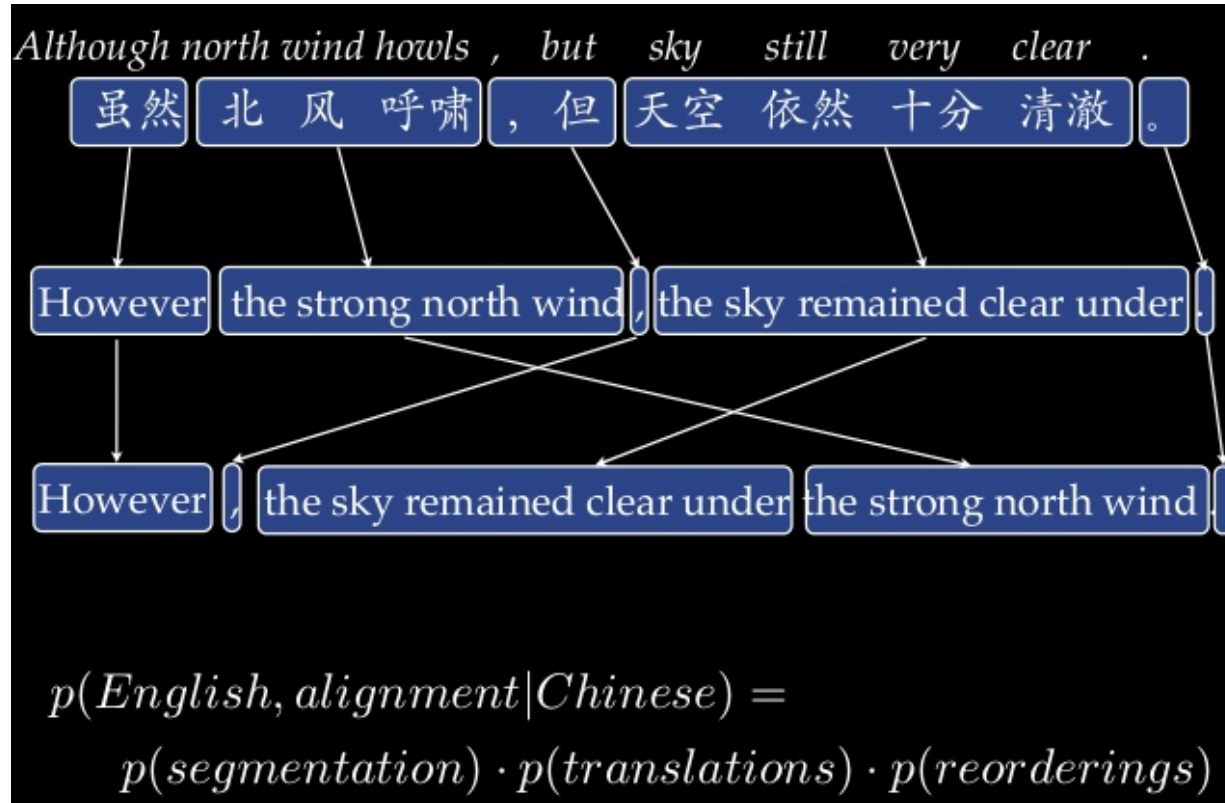
However, the sky remained clear under the strong north wind.

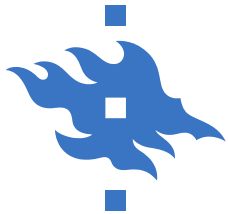
Figure 1: The Vauquois triangle



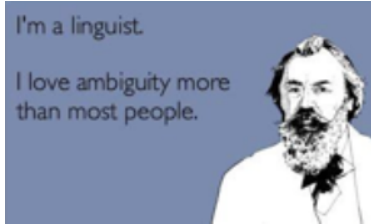
TILASTOLLINEN KONEKÄÄNTÄMINEN

Adam Lopez:
Phrase-based
Models. MTM 2012,
Seventh Machine
Translation Marathon,
Edinburgh,
September 3rd-8th,
2012.





ALAN KEHITYKSESTÄ



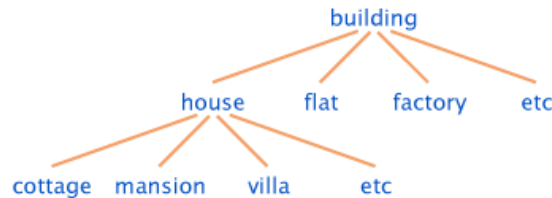
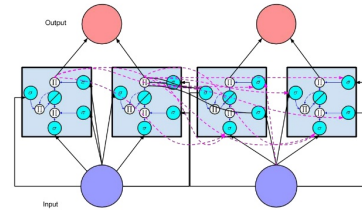
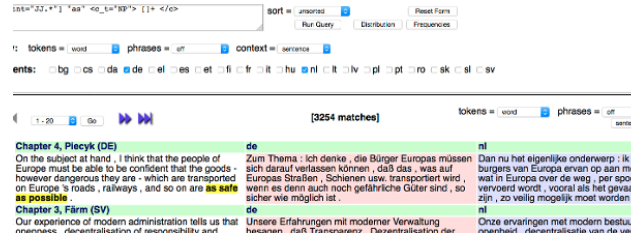
asiantuntijoiden
laatimat säännöt
ja ohjelmat
(1900-luku)



tilastolliset
järjestelmät
(2000-luku)



takaisinkytketyt
LSTM-neuroverkot
(2010-luku)

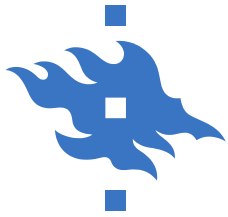


symboliset / kategoriset / diskreetit
representaatiot



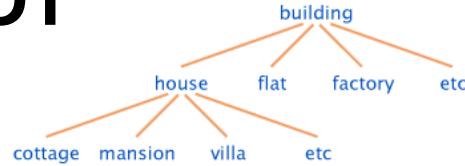
jatkuvat
representaatiot





SYMBOLISET JA JATKUVAT REPRESENTAATIO

- Pragmatiikka
- Semantiikka
- Syntaksi
- Morfologia
- Fonologia
- Fonetiikka

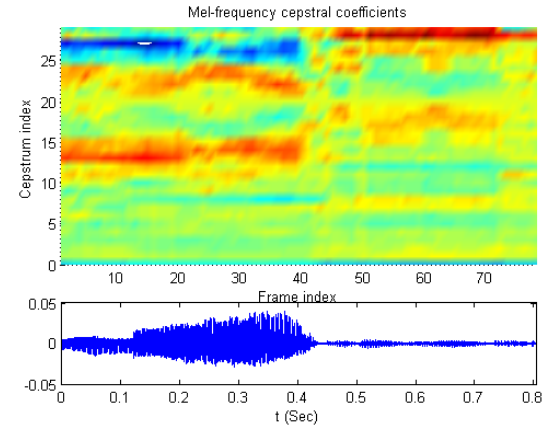


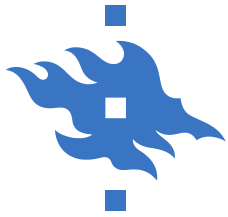
x	y	u	z
Angus(x)			
dog(y)			
own(x,y)			
u = y			
Irene(z)			
bite(u,z)			

lautasiltammeko: lautanen N Abl PI PxPI1 Foc_kO

Perinteisesti symbolinen

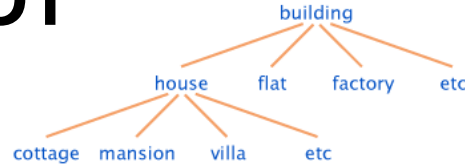
Jatkuva





SYMBOLISET JA JATKUVAT REPRESENTAATIO

- Pragmatiikka
- Semantiikka
- Syntaksi
- Morfologia
- Fonologia
- Fonetiikka

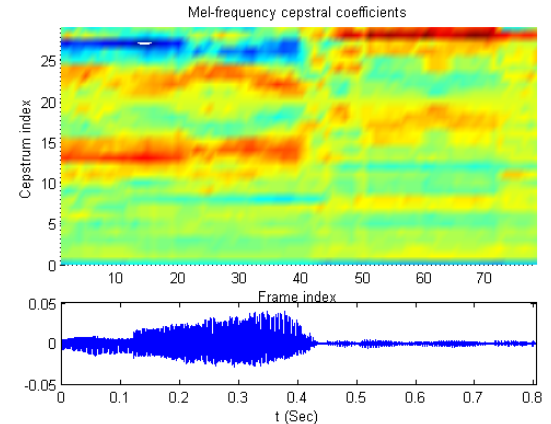


x	y	u	z
Angus(x)			
dog(y)			
own(x,y)			
u = y			
Irene(z)			
bite(u,z)			

lautasiltammeko: lautanen N Abl PI PxPI1 Foc_kO

Perinteisesti symbolinen

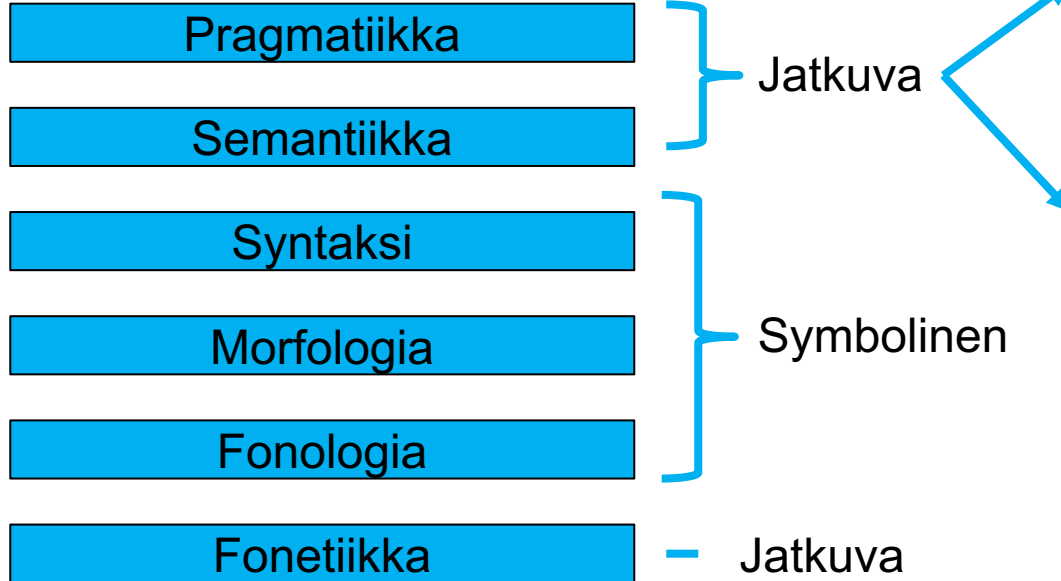
Jatkuva



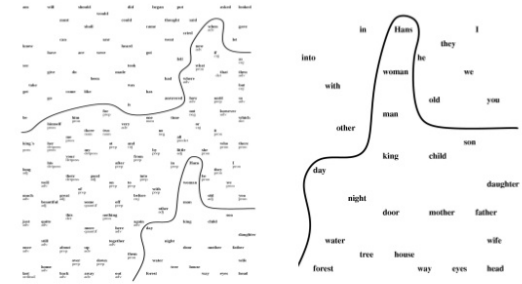
YLEISTYSKYKY? ADAPTIIVISUUS?



SYMBOLISET JA JATKUVAT REPRESENTAATIO, VAIHE 2

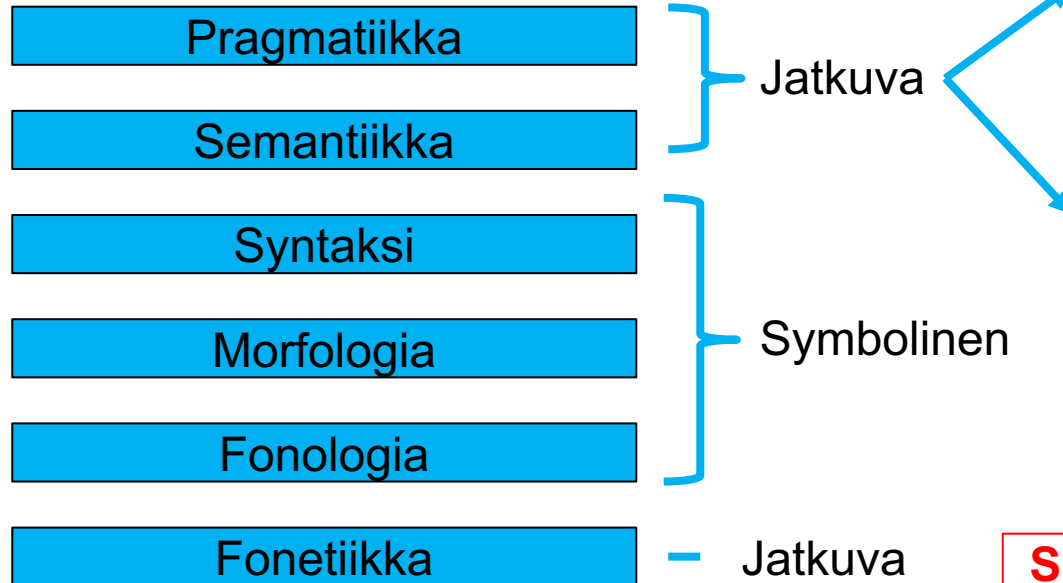


Classical example: A map of words (vector-space model) in Grimm fairy tales

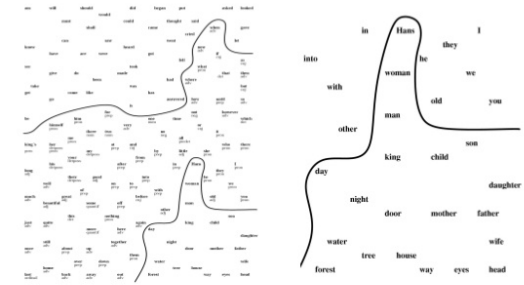




SYMBOLISET JA JATKUVAT REPRESENTAATIO, VAIHE 2



Classical example: A map of words (vector-space model) in Grimm fairy tales



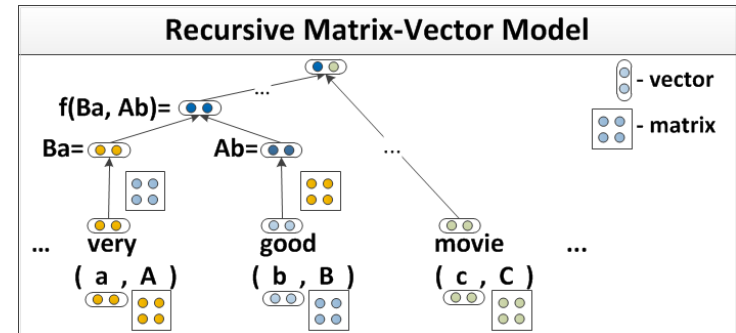
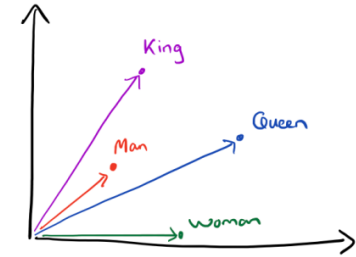
SEKVENTIAALISUUS?
KOMPOSITIONAALISUUS?

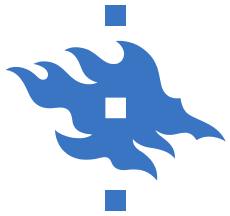


SYMBOLISET JA JATKUVAT REPRESENTAATIOIOT, VAIHE 3

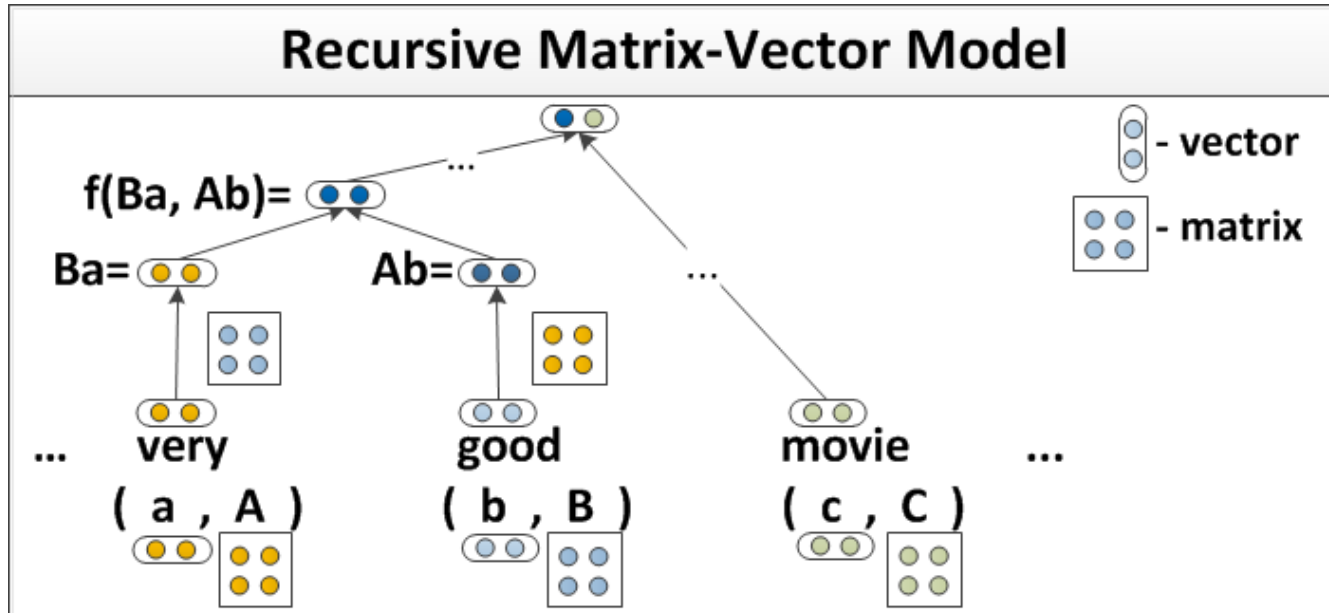


Jatkuva

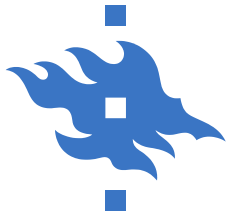




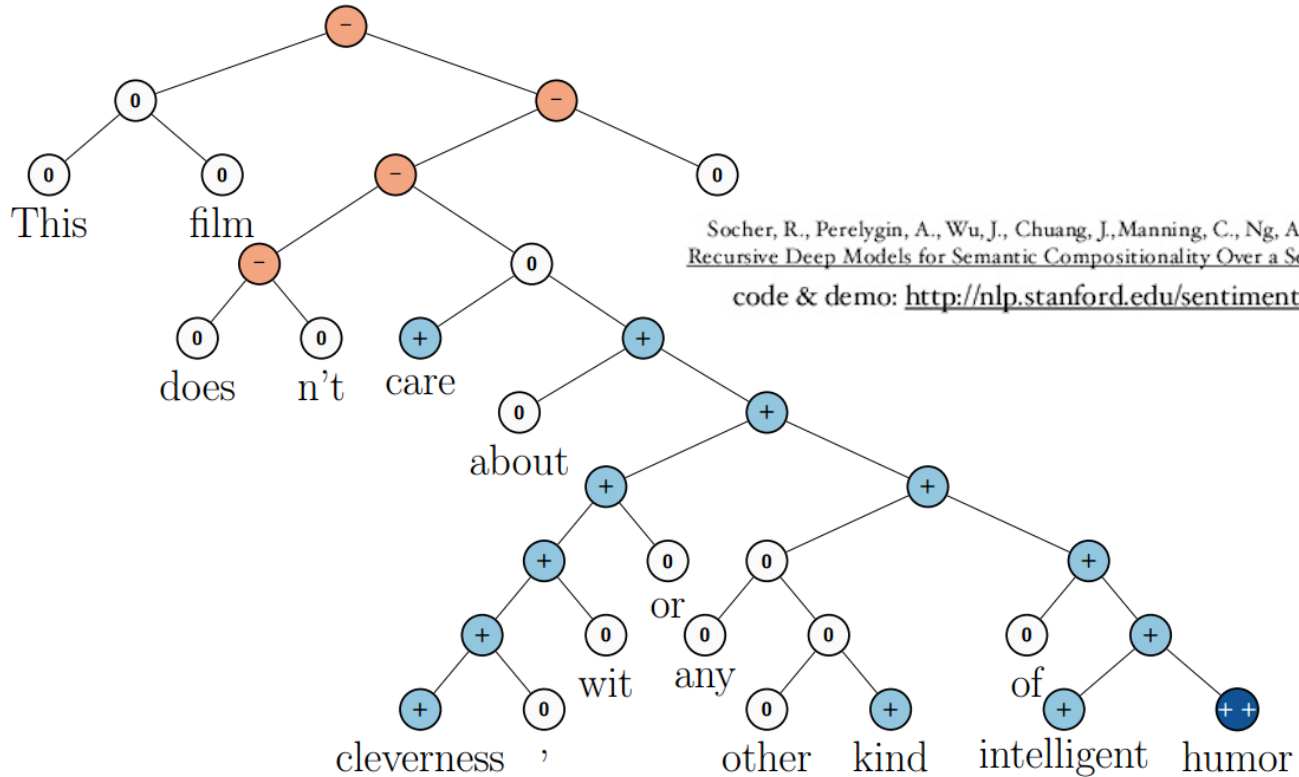
MERKITYKSEN KOMPOSITIONAALISUUS



Richard Socher, Brody Huval, Christopher D. Manning, Andrew Y. Ng: Semantic Compositionality through Recursive Matrix-Vector Spaces. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1201–1211, Jeju Island, Korea, 12–14 July 2012.

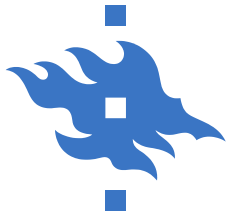


SOVELLETTUNA SENTIMENTTIANALYYSIIN



Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., Potts, C. (2013)
 Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank.

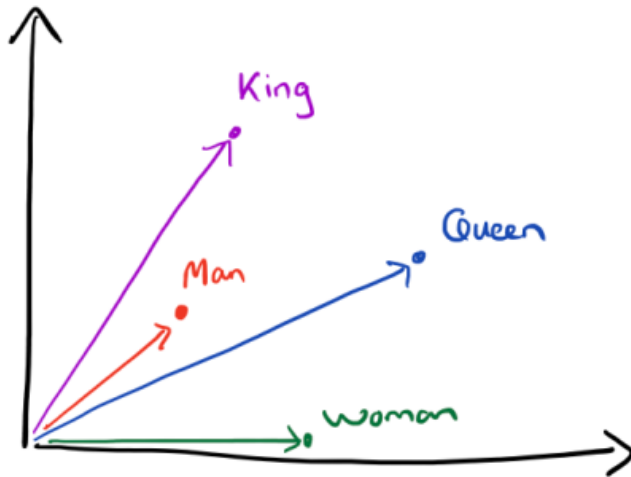
code & demo: <http://nlp.stanford.edu/sentiment/index.html>



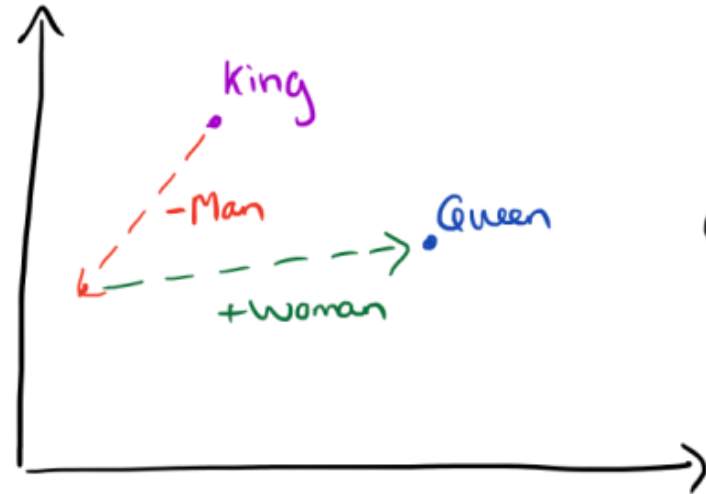
MERKITYKSET VEKTORIAVARUUDESSA

Adrian Colyer: The amazing power of word vectors

<https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/>



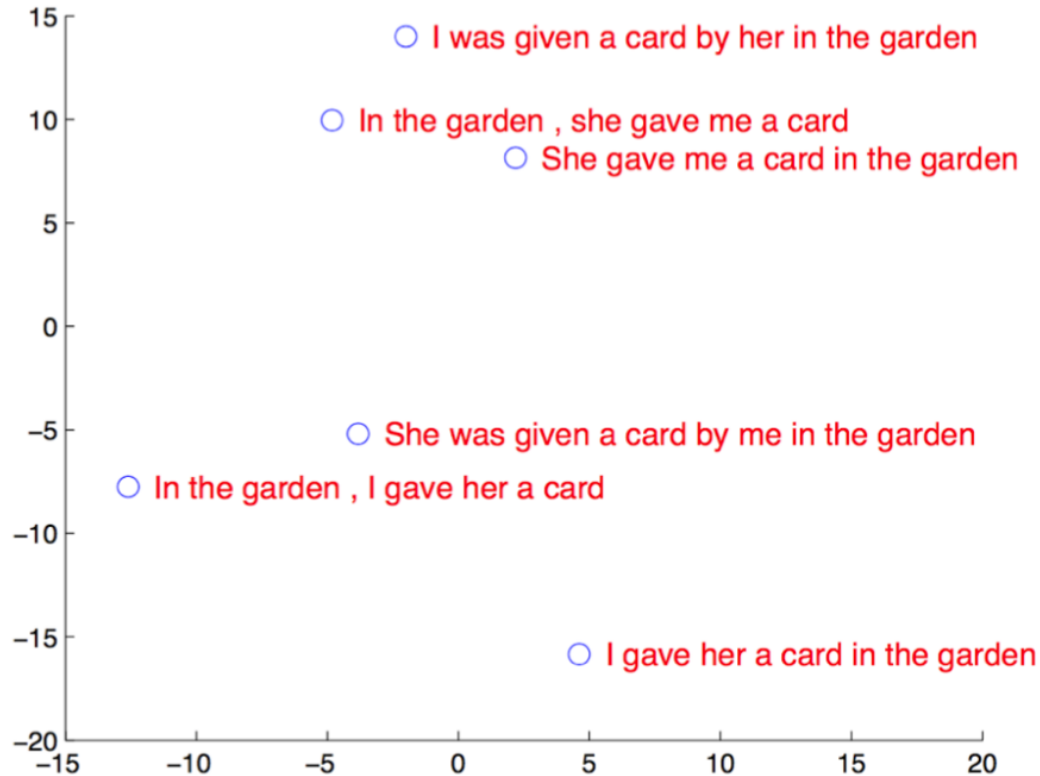
Sanavektorit



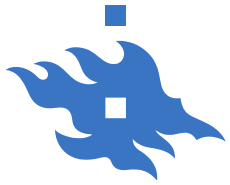
Sanavektorien yhdistelmiä



MERKITYKSET VEKTORIAVARUUESSA



Ilya Sutskever,
Oriol Vinyals,
Quoc V. Le:
Sequence to
Sequence
Learning with
Neural Networks.
NIPS 2014.



NMT-KÄÄNNÖSESIMERKKEJÄ (EN -> FI)

moreover , the employment of approximately 80 Asian temporary workers will end when the routes will be outsourced . => lisäksi noin 80 aasialaisten työntekijöiden työelämä päättyy , kun reittejä tulee ulkoistaa .

however , the EU support can be granted also in the future for investments that improve municipalities ' economy prerequisites . => EU : n tukea voidaan myöntää myös jatkossa investoinneille , jotka parantaisivat kuntien talouden edellytyksiä .

researcher : this is what Russia aims at in Ukraine => tutkija : näin Venäjä tähtää Ukrainaan

the plan is to start the outsourcing during the first three months of next year . => suunnitelma on tarkoitus aloittaa ensi vuoden ensimmäisellä puoliskolla .

according to the company the co - operation negotiations shall continue also with other partner candidates . => yhtiön mukaan yhteistoimintaneuvottelut jatkuvat myös muiden yhteistyökumppanien ehdokkaalla .



Kiitos!