

De informatica als geesteswetenschap

Antal van den Bosch
Centre for Language Studies
Radboud University Nijmegen

Alumnidag 'Informatica voor een interdisciplinaire eeuw', Utrecht, 31 maart 2012



Overzicht

- Geesteswetenschappen
 - Misverstanden
 - Informatica
- Natuurlijke taalverwerking
 - Toepassingen
 - Data
- Zwemmen in tekst
 - Text analytics

Geesteswetenschappen

- **Onderwerp van studie**
 - producten van de menselijke geest
- **Geen eenvoudig formaliseerbare fysieke wereld**
 - een eindeloze zee van abstracties
- **Complexe ruimtes in tijd**
 - Folklore, cultuur, kunst
 - Taal, communicatie, media
 - Ideologie, religie, filosofie, wetenschap

Perceptie van geesteswetenschappen

- “De vergeten wetenschappen” (Bod, 2010)
- De geesteswetenschapper als eenzame wolf
- Onderontwikkelde notie van formalisering en data
- Associatie met stereotypische onderzoeksmethoden
- Scherpe grens met beta en gamma
- Weinig impact



Bestaat het α / β onderscheid?

- Data, informatie, kennis, wijsheid, ...
 - Taal is een datamedium
 - Logica, kennisrepresentatie, redeneren
- Middelbare school
 - Latijn – taal als systeem
 - Economie – menselijk handelen als systeem
- Geschiedenis van de kunstmatige intelligentie
 - Turing test
 - NLP vast onderdeel van AI-conferenties



Turing

This last statement needs some explanation. It may appear rather startling, but with some reservations it appears to be an inescapable fact. It can be shown to follow from a characteristic property of digital computers, which I will call their universality. A digital computer is a universal machine in the sense that it can be made to replace any machine of a certain very wide class. It will not replace a bulldozer or a steam-engine or a telescope, but it will replace any rival design of calculating machine, that is to say any machine into which one can feed data and which will later print out results. In order to arrange for our computer to imitate a given machine it is only necessary to programme the computer to calculate what the machine in question would do under given circumstances, and in particular what answers it would print out. The computer can then be made to print out the same answers.

Turing

In view of this it seems that the wisest ground on which to criticise the description of digital computers as 'mechanical brains' or 'electronic brains' is that, although they might be programmed to behave like brains, we do not at present know how this should be done. With this outlook I am in full agreement. It leaves open the question as to whether we will or will not eventually succeed in finding such a programme. I, personally, am inclined to believe that such a programme will be found. I think it is probable for instance that at the end of the century it will be possible to programme a machine to answer questions in such a way that it will be

Turing

extremely difficult to guess whether the answers are being given by a man or by the machine. I am imagining something like a viva-voce examination, but with the questions and answers all typewritten in order that we need not consider such irrelevant matters as the faithfulness with which the human voice can be imitated. This only represents my opinion; there is plenty of room for others.

There are still some difficulties
One difficulty appears to remain. To behave like a brain seems to involve free will, but the behaviour of a digital computer, when it has been programmed, is completely determined. These two facts must somehow be reconciled, but to do so seems to involve us in an age-old controversy that of 'free will and determinism'. There are two ways out. It

Tekst-naar-tekstverwerking

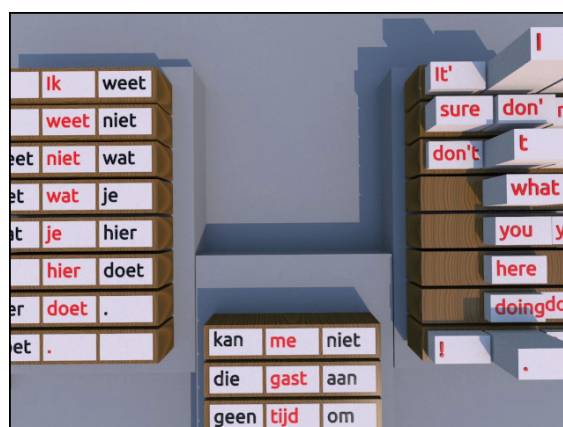
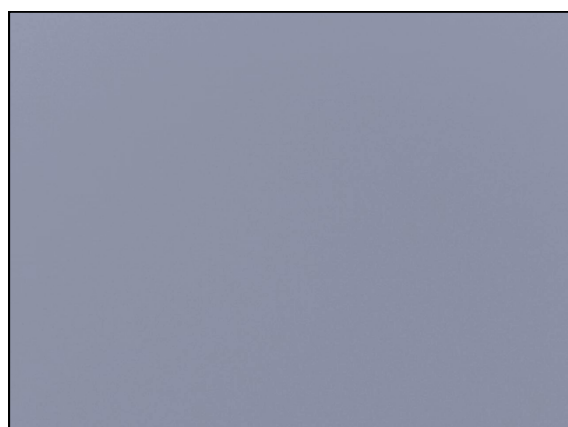
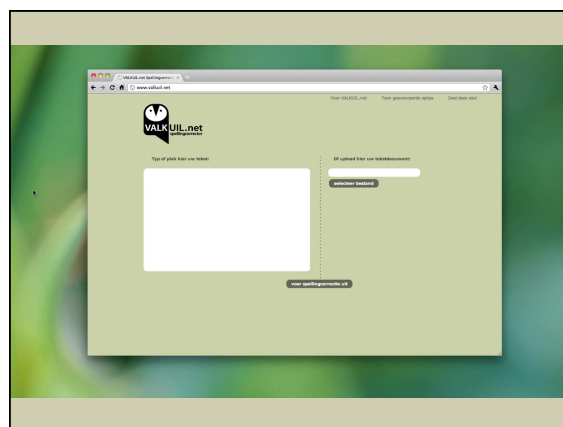
- Automatisch vertalen
- Automatische vraagbeantwoording
- Mens-machine dialoogvoering
- Automatisch samenvatten, parafraseren, simplificeren
- Spraakherkenning
- Spraaksynthese
- Zoekmachines
- Recommender systems
- Spelling- en grammaticacorrectie
- Text completion

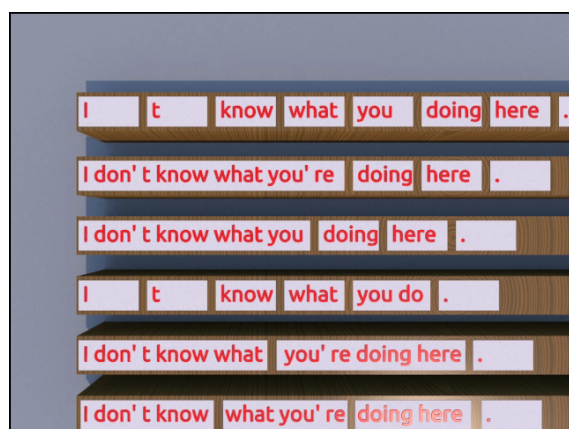
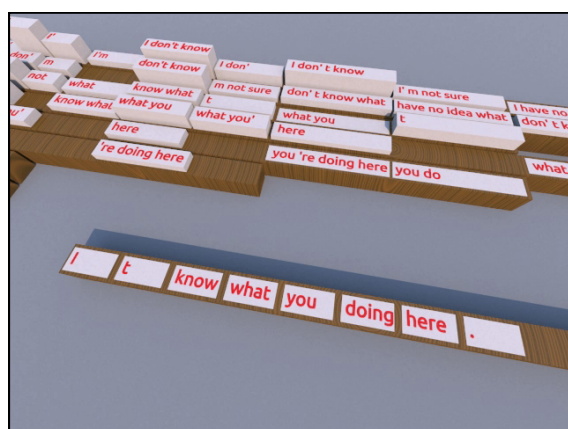
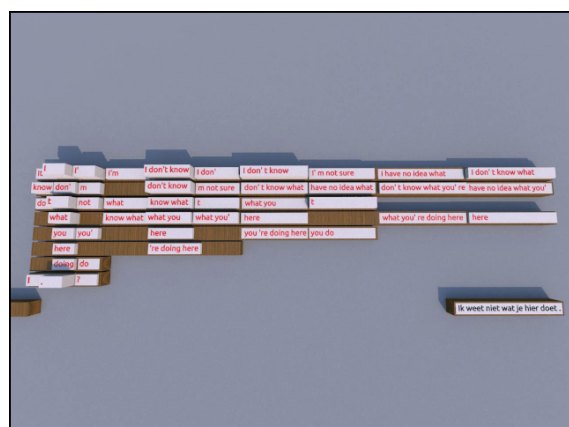


- Extreem data-intensieve toepassingen
- Beste methoden werken met veel data en weinig tot geen expliciete taalkundige kennis



Het is taalkunde, Jim, maar niet zoals we het kennen.







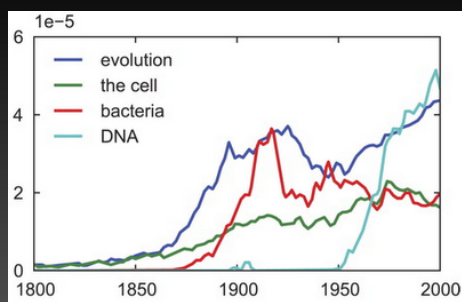
There is no data like more data

Gebruikte corpora

aantal woorden

• Alice In Wonderland	33,361
• Bijbel	783,137
• Corpus Gesproken NL	10,089,346
• Reuters RCV1	130,396,703
• Twente-ILK News Corpus	566,350,874
• LDC Gigaword	2,080,268,230
• ValkuilCorpus	3,491,856,359
• Google 1Tb 5-grams	5,882,353,315

Culturomics



Voorbeeld bestaande software

- Frog: automatische morfosyntactische analyse
 - 10.000 regels, 138.689 woorden
 - Zonder grammaticale relaties:
 - 82 seconden, **1,678** woorden per seconde
 - 529 Mb voetafdruk op 64-bit arch
 - Met grammaticale relaties:
 - 623 seconden, **223** woorden per seconde
 - 1944 Mb voetafdruk
- Beschikbaar als cloud webservice

Hoe goed werkt het?

Module	Wat	% correct
MBT	Bekende woorden	96.8 (98.7)
	Onbekende	76.4 (84.3)
	Alle woorden	96.5 (98.6)
MBLEM	Onbekende	73.9
MBMA	Onbekende	79.0
MBDP	Gramm. relaties	81.9

F911.3 Animal swallows man (not fatally)

1.	"	"	LETI	0	ROOT
2.	Nauwelijks	nauwelijks	BWJ	3	mod
3.	had	hebben	WW(pv,ver,ev)	0	ROOT
4.	de	de	LID(bep,stan,rest)	5	det
5.	woof	woof	N(soort,ev,basis,zijd,stan)	3	su
6.	dat	dat	VNW(aanw,pron,stan,vol,3o,ev)	7	obj1
7.	gezegd	gezegd	WW(vd,vrij,zonder)	3	vc
8.	,	,	LETI	7	punct
9.	of	of	VG(onder)	7	mod
10.	hij	hij	VNW(pers,pron,nomin,vol,3,ev,masc)	11	su
11.	sprong	springen	WW(pv,ver,ev)	9	cnj
12.	uit	uit	VZ(init)	3	vc
13.	bed	bed	N(soort,ev,basis,onz,stan)	12	obj1
14.	en	en	VG(reven)	0	ROOT
15.	slokte	slokken	WW(pv,ver,ev)	14	cnj
16.	Roodkapje	Roodkapje	SPEC(deeleigen)	15	su
17.	in	in	VZ(init)	15	mod
18.	den	den	TW(isoofd,prenom,stan)	19	det
19.	hap	hap	N(soort,ev,basis,zijd,stan)	17	obj1
20.	op	op	VZ(fin)	15	svp
21.	.	.	LETI	20	punct

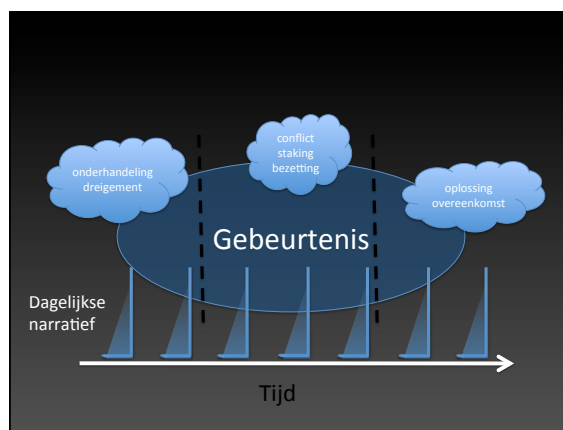
HiTIME

1	Luitjes	Luitjes	[Luitjes]	SPEC(deeleigen)	2	su
2	belandde	belanden	[be][land][de]	WW(pv,verl,ev)	0	ROOT
3	al	al	[al]	BW()	4	mod
4	jong	jong	[jong]	ADJ(vrij,basis,zonder)	2	predc
5	in	in	[in]	VZ(init)	2	mod
6	de	de	[de]	LID(bep,stan,rest)	8	det
7	socialistische	socialistisch	[socialist][isch][e]	ADJ(prenom,basis)	8	mod
8	beweging	beweging	[beweeg][ing]	N(soort,ev,basis)	5	obj1
9	in	in	[in]	VZ(init)	8	mod
10	de	de	[de]	LID(bep,stan,rest)	11	det
11	stad	stad	[stad]	N(soort,ev,basis)	9	obj1
12	Groningen	Groningen	[Groningen]	SPEC(deeleigen)	11	app
13	.	.	[.]	LET()	12	punct

Text analytics

- Assisteer – automatiseer:
 - Assisteer de expert met zoeken, browsen, serendipiteit
 - Geef de expert suggesties door te filteren, classificeren, matchen
 - Doe het monnikenwerk: informatie-extractie, aggregatie, tellen, berekenen

Verrijken van bestaande informatie

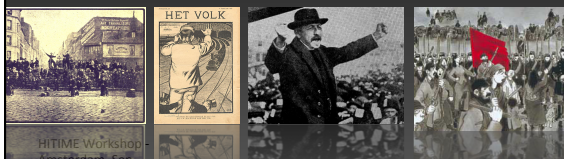


Historische teksten: HiTIME

Geschiedenis van de sociale beweging in Nederland en Europa, 1870-1940

- O1: Primair brontekstmateriaal
- O2: Secundaire bronnen
- O3: Verbinden van de informatie
- O4: Visualiseren met tijdlijn, netwerk

Persoon
Beroep
Organisatie
Locatie
Tijd
Gebeurtenissen



Stakingendatabase

- Bijgebracht door Sjaak van der Velden
- Bevat alle 16+ duizend stakingen in Nederland



Nationaal Instituut voor Sociale Geschiedenis




[Index zoeken](#) / [Database stakingen in Nederland](#)

< 5 | 16 >

[Resultatenlijst](#) | [Zoekopdracht aanpassen](#)

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

Krantenarchieven

- Koninklijke Bibliotheek (Den Haag)
 - Selectie kranten van 1618 tot 1995
 - Op naar 8 miljoen krantenpagina's online
 - Zoekbaar op titel, datum woorden



Zoekresultaten

[illegible]

Discriminatieve woorden



HiTiME

- Vind alle concepten in alle teksten
- Verbind ze met betekenisvolle relaties
- Vul een kennisbank



Toen **Domela Nieuwenhuis** in 1887 na een jaar gevangenschap wegens vermeende majesteits-schennis vrij kwam , organiseerde **Baye** voor hem in **Walhalla** een groots onthaal .

When Domela Nieuwenhuis was released in 1887 after a year of imprisonment for alleged Majesty-sacrilege , Baye organized a grand welcome for him in Walhalla .

