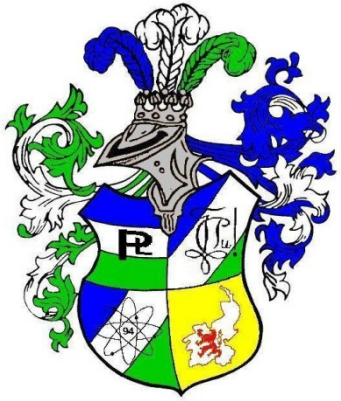


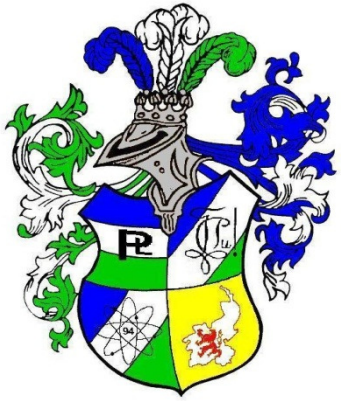
Academia Plutonicana

Een inleiding tot taaltechnologie

Peter Dirix

15 december 2010

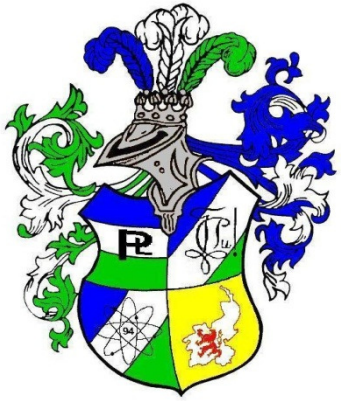
Lange Trappen, Leuven



Academia Plutonicana

Overzicht

- Overzicht & terminologie
- Formele taaltheorie
- Voorbeelden uit automatische vertaling en spraakherkenning

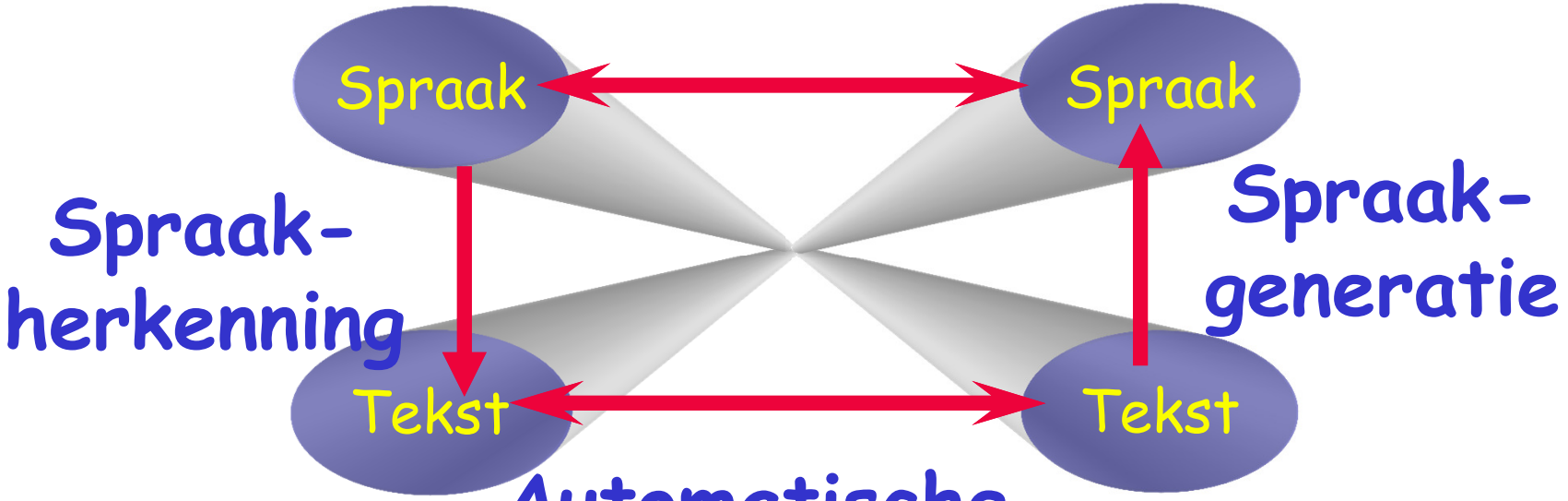


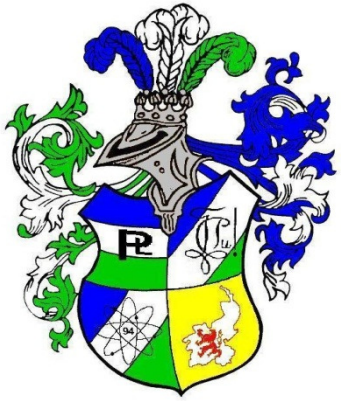
Taaltechnologie

- Verschillende benamingen:
  - Taaltechnologie (LT)
  - Computertaalkunde (CL)
  - Natuurlijketaalverwerking (NLP)
  - Spraak- en taalverwerking (SLP)

Tekst en spraak

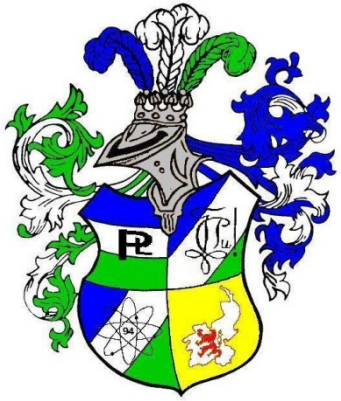
Spraakcompressie





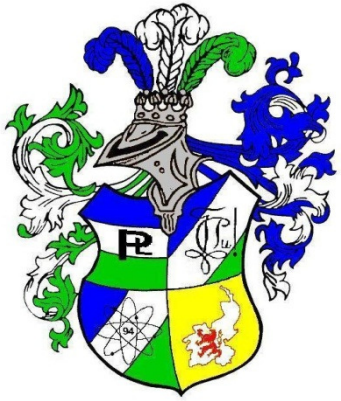
Domeinen

- Spraakherkenning (ASR)
- Spraakgeneratie / spraaksynthese (TTS)
- Automatische vertaling (MT)
- Automatisch samenvatten
- Dialoogsystemen
- Informatieontsluiting en –extractie (IR/IE)
- Natuurlijketaalbegrip (NLU)
- Spelling- en grammaticacontrole
- Vertaalgeheugens



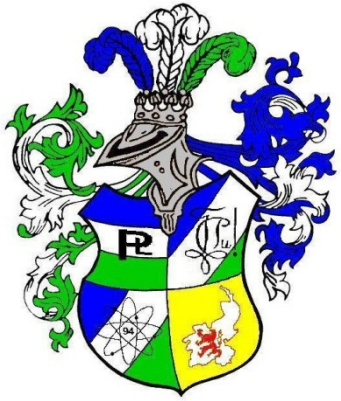
Taal

- Klanken (fonetiek & fonologie)
- Woordenschat / lexicon
- Syntaxis / grammatica
- Semantiek / betekenis
- Pragmatiek / context



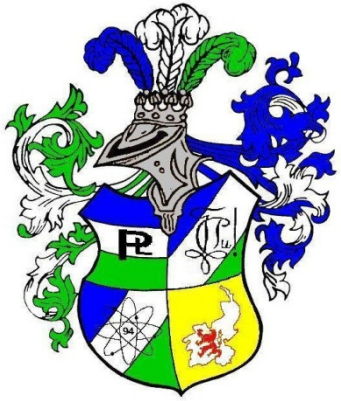
Problemen met taal

- Ambiguïteit
  - woordenschat
  - homofonie
  - syntactische ambiguïteit
- Woordenschat is productief



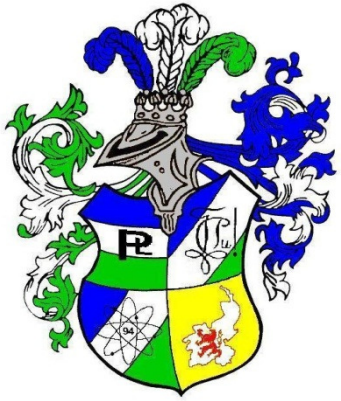
Filosofie

- Wanneer is een systeem 'intelligent'?
- Turingtest (1950)
- ELIZA (1966)
- Kunstmatige intelligentie



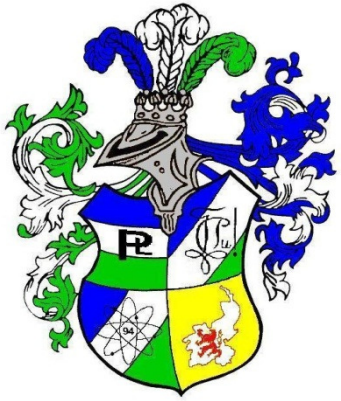
Korte geschiedenis

- 1939: Eerste spraaksynthese (Bell Labs)
- Jaren '40-'50: theoretische beschouwingen – Turing, Shannon
- Formeletaaltheorie (Chomsky, 1956)
- Vanaf jaren '50: automatische vertaling
- 1952: Eerste spraakherkenner (Bell Labs)
- 1976: Statistische aanpak
- Jaren '90: commerciële softwarepakketten & hybride systemen



Tools (1)

- Tokeniseren: tekst terugbrengen naar tokens - atomaire eenheden (club, ., New York)
- Morfologische analyse (terugbrengen van een token tot woordenboekvorm, maar ook stemming, herkennen van samenstellingen, afleidingen, voor- en achtervoegsels, …)



Tools (2)

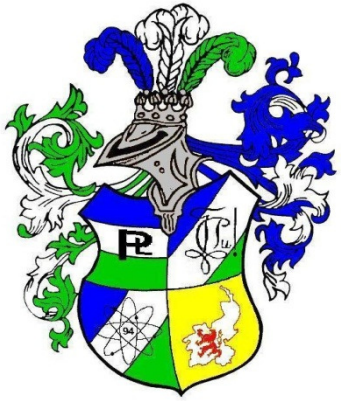
- Syntactische analyse (herkenning van woordklasse, functie in de zin)
- Semantische analyse (zoekt bv. het antecedent van een betrekkelijk voornaamwoord)
- Pragmatische analyse (analyseert bv. of een tekst formeel of niet-formeel is)

Formele grammatica's (1)

- Symbolen: terminale en niet-terminale
- Herschrijfregels
- 5 types:
  - constituentenstructuurgrammatica's (PSGs, 0)
  - contextgevoelige grammatica's (CSGs, 1)
  - contextvrije grammatica's (CFGs, 2)
  - eindigetoestandsgrammatica's (FSGs, 3)
  - eindigekeuzegrammatica's (FCGs, 4)

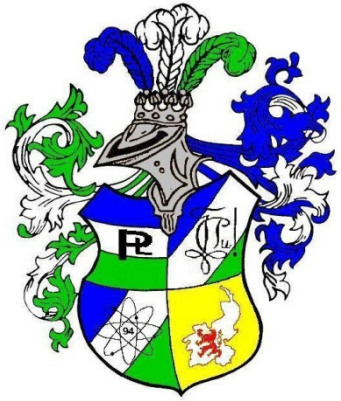
Formele grammatica's (2)

- S -> NP VP
- NP -> Pronoun
- VP -> V
- V -> wandel | loop
- Pronoun -> ik

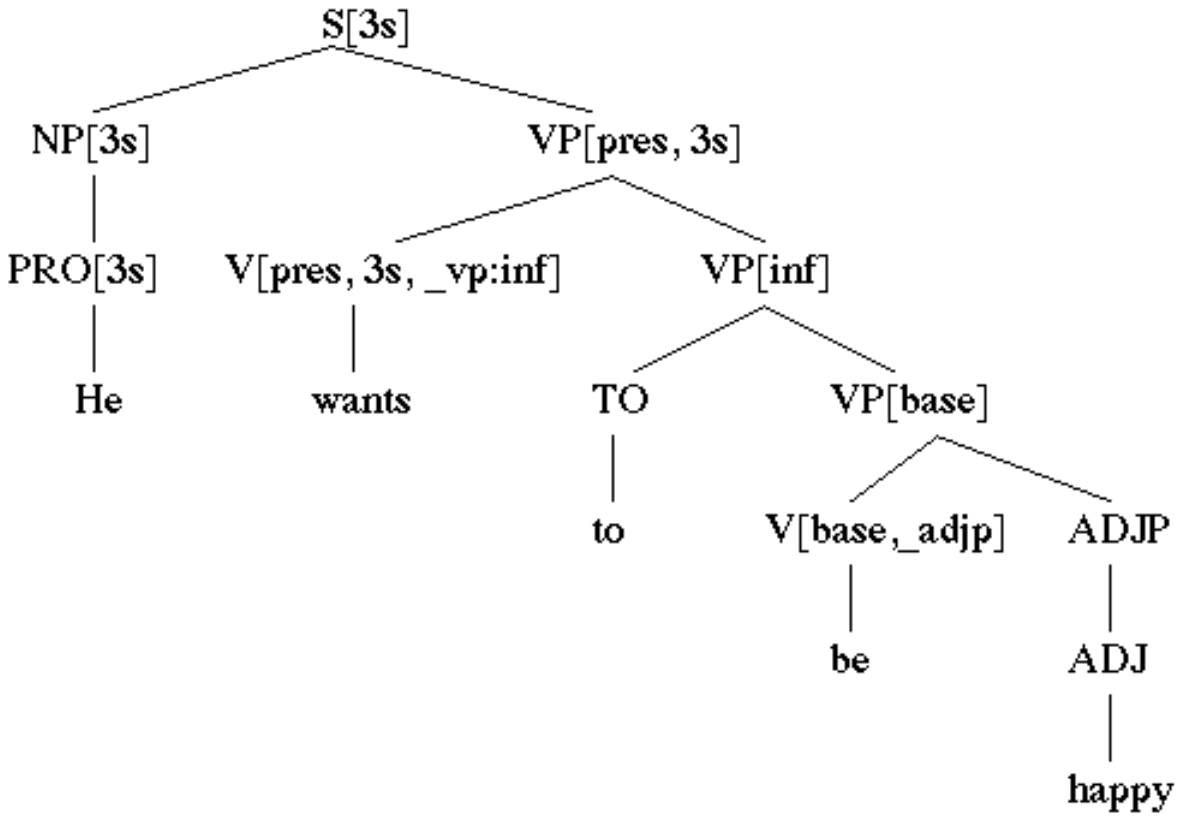


Analyse: parsers

- Parser of ontleder is programma dat deze analyse doet
- Verschillende benaderingen:
  - top-down / bottom-up
  - breedte-eerst / diepte-eerst
- Creëert een ontleedboom die de analyse bevat

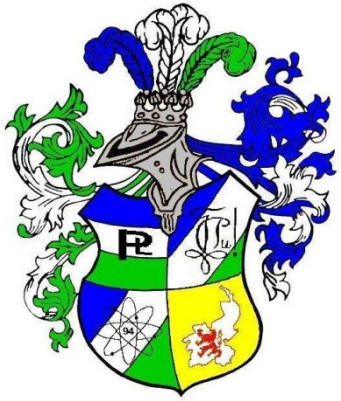


Ontleedboom



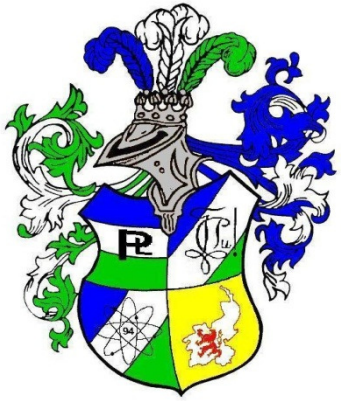
Regelgebaseerde automatische vertaling

- Analyse van brontaal, minimaal morfologisch en syntactisch
- Gebruik van vertaalwoordenboek
- Gebruik van regels, bv. Spa-Eng: zelfst.+bijv. nw. -> bijv.+zelfst. nw.
- Eventueel nog herschrijfregels (postprocessor) in doeltaal
- Bv. SYSTRAN, EUROTRA, METEO



Spraakherkenning: geschiedenis (1)

- 1952: Eerste spraakherkenner (Bell Labs)
- 1969: Darpa begint financiering
- 1976: Eerste statistische modellen (IBM)
- Jaren '80: Eerste toepassingen (Dragon Systems, IBM, Kurzweil)
- 1990: Eerste dicteertoepassing (Dragon)
- 1997: IBM ViaVoice (eerste PC-toepassing)
- 1997: DragonNaturallySpeaking
- 1998: VoiceXpress (Lernout & Hauspie)



Spraakherkenning: geschiedenis (2)

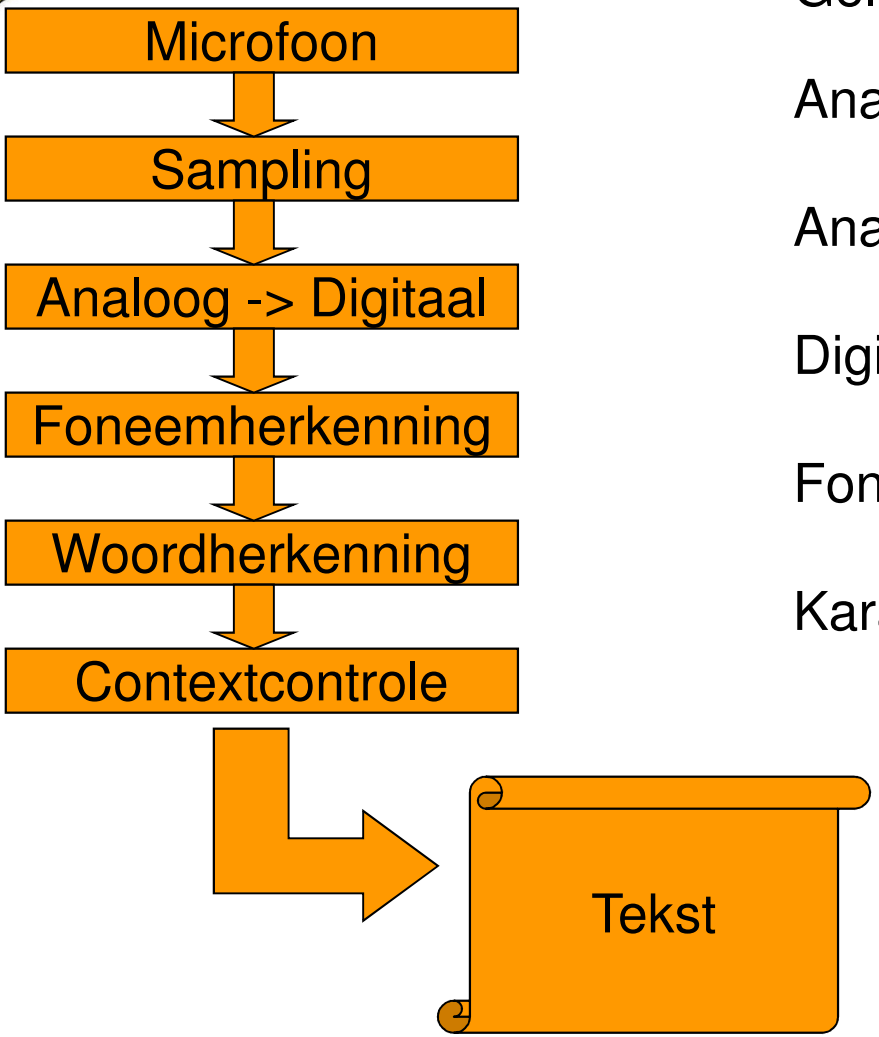
- 1997: L&H koopt Kurzweil
- 2000: L&H koopt Dragon en Dictaphone
- 2001: faillissement L&H; gekocht door Scansoft
- 2003: Scansoft koopt Speechworks en Philips Speech
- 2005: Scansoft koopt Nuance, nieuwe naam
- 2006: Nuance koopt Dictaphone
- 2009: Nuance koopt IBM-patenten
- 2010: 3 grote spelers – Nuance, Google, Microsoft

Types spraakherkenning

- Sprekeronafhankelijk
  - geen training
  - beperkte woordenschat
  - meestal gebruikt in mobiele toepassingen
- Sprekerafhankelijk
  - systeem wordt getraind per gebruiker
  - algemeen akoestisch model
  - taalmodel per gebruiker
  - uitgebreide woordenschat
  - gebruikt in dicteertoepassingen

Gesproken taal

Spraakherkenning: proces



Geluidssignaal

Analoog Signaal

Analoog Signaal

Digitaal Signaal

Fonemen

Karakters / woord

Spraakherkenning: aanpak (1)

- Statistische benadering
- Spraakherkenningspakket bestaat uit:
  - akoestisch model (AM)
  - taalmodel (LM)
  - herkenner: zoekalgoritme
    * analyse van signaal
    * transformeert signaal naar een woord

Spraakherkenning: aanpak (2)

- Stochastische aanpak maakt gebruik van de inherente statistische eigenschappen van het samen en gecombineerd voorkomen van individuele spraaksegmenten

woordvolgorde: $W = w_1, w_2, \ldots, w_n \in V$

foneemvolgorde: $P = p_1, p_2, \ldots, p_m \in Z$

akoestiek (spraakspectrumvectoren): $Y = y_1, y_2, \ldots, y_t$.

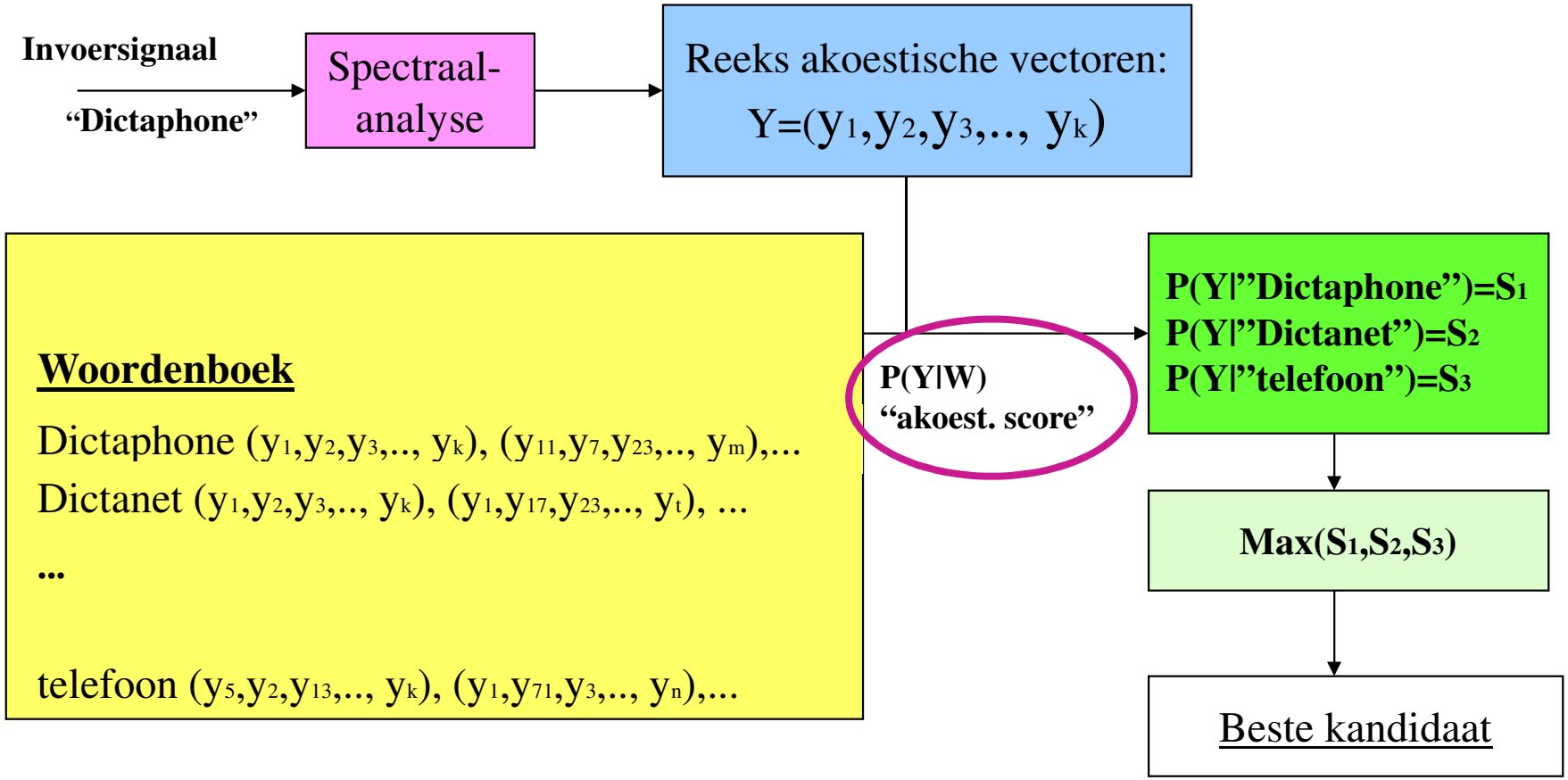
W' = argmax P(W | Y)

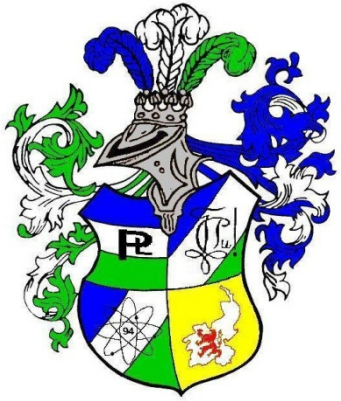
W'= argmax P(W) P(Y | W) / P(Y)

Akoestisch model (1)

W'= argmax P(W) P(Y | W)

P(Y|W) – akoestisch model (waarschijnlijkheid voor een willek. vectorvolgorde Y, gegeven woord W)



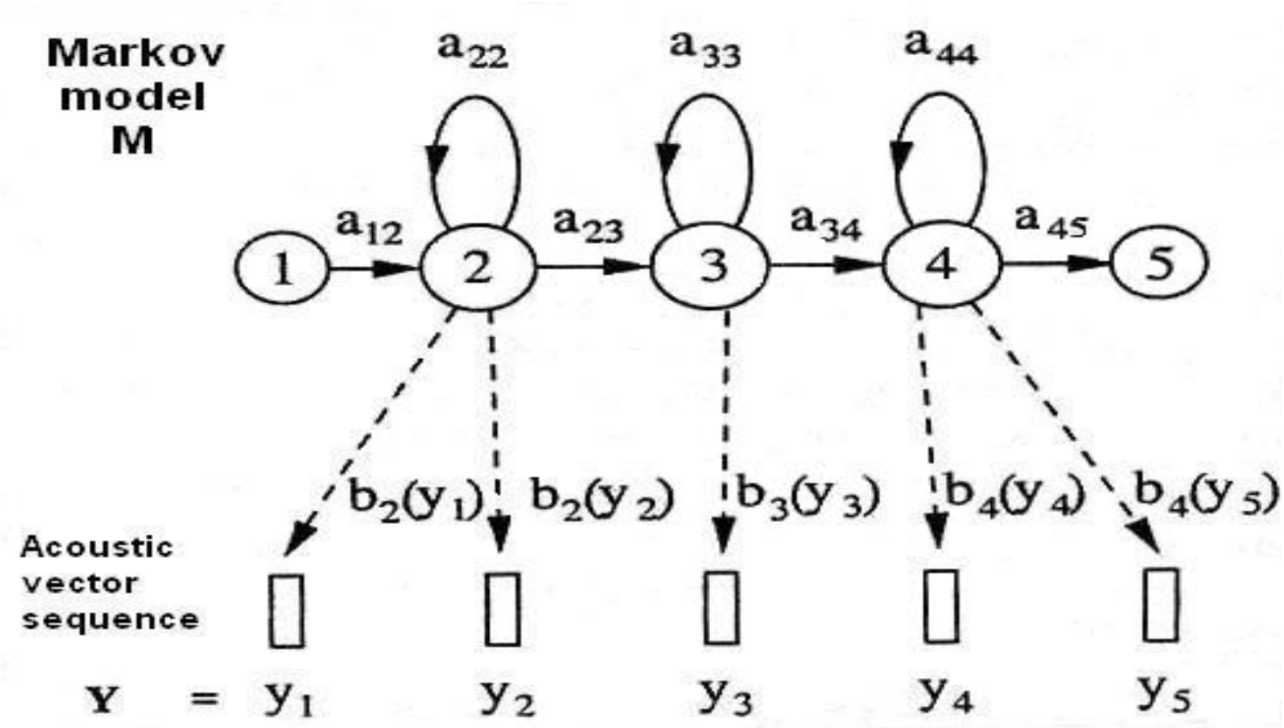


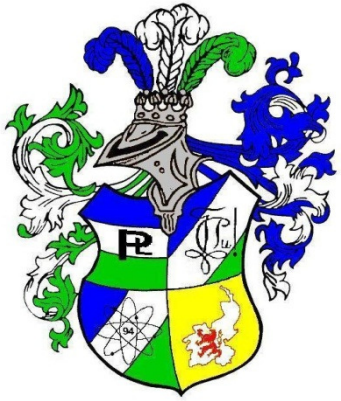
Akoestisch model (2)

- Woorden ontbinden in fonemen (30-70 per taal)
- Toevoegen pauzevullers, geluiden, etc.
- Fonemen groeperen in trifonen:
  test → t-e-s en e-s-t
- Elke trifoon wordt voorgesteld door een reeks toestanden (2 à 3)
- Gemodelleerd door verborgen markovmodel (HMM)
- Overgang tussen toestanden volgens viterbialgoritme

Akoestisch model (3): HMM

- Probabilistische overgang tussen toestanden (a_{ij})
- Iedere toestand produceert akoestische vector y_t met waarschijnlijkheid $b_j(y_t)$
- $P(Y) = a_{12}\, b_2(y_1)\, a_{22}\, b_2(y_2)\, a_{23}\, b_3(y_3)\ldots$





Akoestisch model (4): training

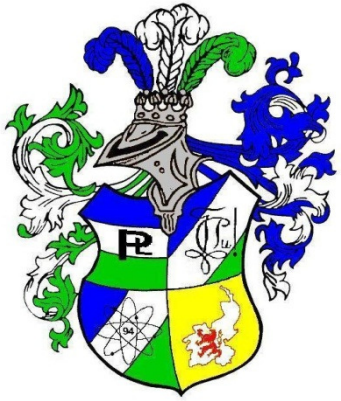
- Selecteer verzameling fonemen
- Transcribeer alle woorden fonetisch volgens deze verzameling
- Definieer het aantal toestanden/foneem
- Representatieve datacollectie: man/vrouw, leeftijd, opleiding, regio, …
- Train het HMM door voor elke trifoon te berekenen:
  - overgangswaarschijnlijkheden tussen de toestanden (a_{ij})
  - productiewaarschijnlijkheden voor een bepaalde akoestische vector $b_j(y_t)$

Akoestisch model (5): aanpassing aan spreker

- Gebruiker leest aantal teksten
- Voor elk audio-/tekstpaar gaan we de tekst herkennen en vergelijken met wat er herkend zou moeten zijn
- Aligneren
- Indien er een goede herkenning is en consistente verschillen, wordt het AM aangepast

Akoestisch model (6): aanpassing aan spreker

- Aanpassing aan spreker verbetert herkenning bij:
  - consequent gewestelijk accent
  - consequente omgevingsgeluiden
  - consequente dicteersnelheid
- Aanpassing aan spreker lost niets op bij:
  - consequent verkeerd uitspreken van individuele woorden
  - gebruik van woorden die niet in woordenboek zitten
  - plotse veranderingen in omgevingsgeluiden
  - inconsequent dicteergedrag



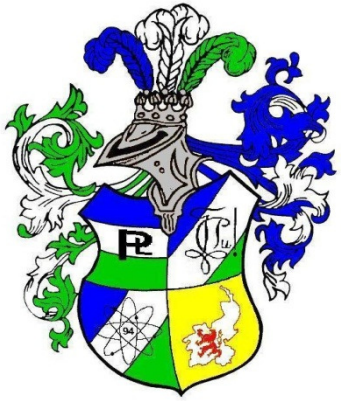
Taalmodel (1)

- **W'= argmax P(W) P(Y | W)**

  P(Y|W) – Akoestisch model

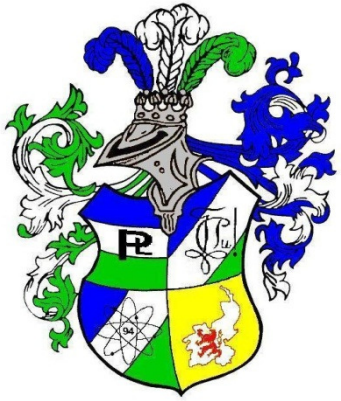
  P(W) – Taalmodel

- Modelleert syntaxis en semantiek
- Modelleert lokale afhankelijkheid
- Akoestisch model genereert hypothesen
- Taalmodel geeft een score aan deze hypothesen: $P(W = w_1, w_2, \ldots, w_n) = P(w_n \mid w_1, w_2, \ldots, w_{n-1})$
- Akoestisch model per taal, taalmodel per domein/gebruiker



Taalmodel (2)

- Spraakherkenners gebruiken eenvoudig taalmodel: n-grammen
  - bigram: $P(W) = P(w_n \mid w_{n-1})$
  - trigram: $P(W) = P(w_n \mid w_{n-2}, w_{n-1})$
  - 4-gram: $P(W) = P(w_n \mid w_{n-3}, w_{n-2,} w_{n-1})$
- Probleem: extreem groot aantal mogelijke tri-/4-grammen (*data sparsity*)
- Opvangen door terug te vallen op bigrammen of woordklassemodellen

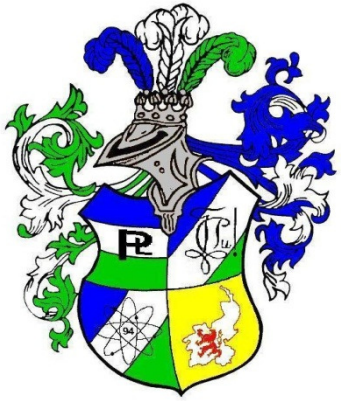


Taalmodel (3)

- Sommige trigrammen veel frequenter dan andere: `geachte heer ,\komma` is veel waarschijnlijker dan `geachte meer bonen`
- Tegenwoordig ongeveer 150.000 unigrammen en een paar tientallen miljoenen bi-, tri- en 4-grammen, bv.

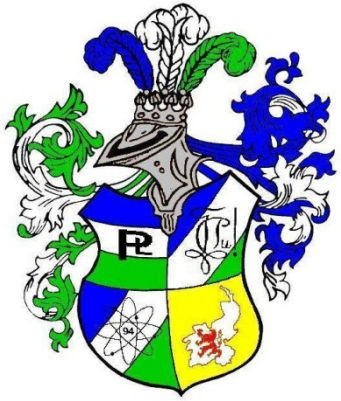
« De vrouw zou ongeveer dertig jaar oud zijn. »

```
de vrouw zou 1
vrouw zou ongeveer 1
zou ongeveer dertig 1
ongeveer dertig jaar 1
dertig jaar oud 1
jaar oud zijn 1
oud zijn .\punt 1
```



Taalmodel (4): training

- Zeer veel data nodig: *There's no better data than more data*
- Representatieve datacollectie: veel auteurs, types tekst, regio, opleiding, …
- Terugbrengen naar standaardformaat
- Maken fonetische transcripties voor nieuwe woorden (*pronning*), bv.
  `spraakherkenning 'sprak-hEr-kE-nIN`
- Nieuw taalmodel berekenen



Postprocessor

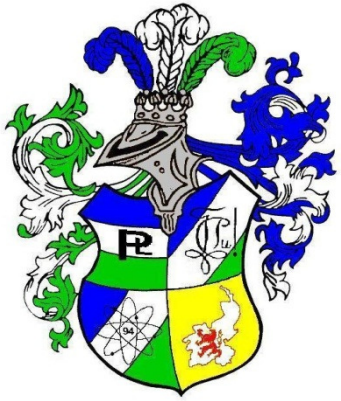
- Regelgebaseerd
- Herschrijft getallen, data, tijden, telefoonnummers etc. naar een door de gebruiker gekozen formaat, bv.
  15/12/2010, 15 december 2010, 2010-12-15, …
- Zorgt er ook voor dat leestekens juist geplaatst worden, bv. een punt wordt tegen het vorige woord geplakt

Uitdagingen in spraakherkenning

- Veranderlijkheid tussen sprekers: lengte stemkanaal, taalsociologische factoren (moedertaal, accent, opleiding)
- Veranderlijkheid bij dezelfde spreker: verkoudheid, emoties
- Veranderlijkheid in de omgeving
- Coarticulatie, prosodie, etc.
- Dicteergewoonten van de gebruiker
- Verwachtingspatroon van de gebruiker

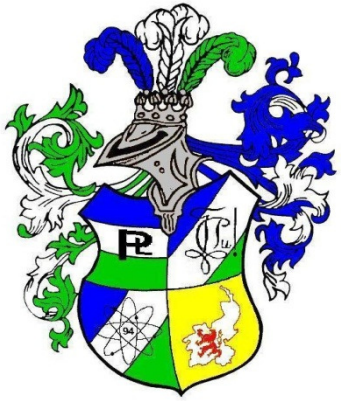
Terug naar automatische vertaling

- Jaren '50: Heilige Graal
- 1966: vernietigend rapport
- Problemen:
  - rekenkracht computers
  - ontwikkeling van regels en woordenboeken is heel arbeidsintensief



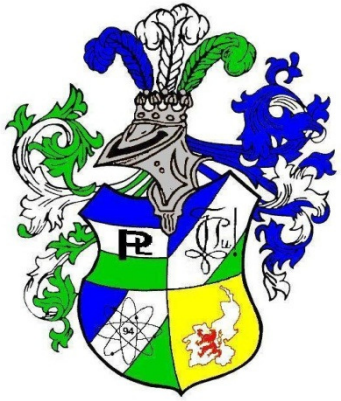
Spraakherkenning en SMT

- Jaren '80: statistische modellen o.i.v. spraaktechnologen
- Statistische automatische vertaling (SMT)
- Nood aan parallelle corpora
- Taalmodellen:
  - taalmodel brontaal
  - vertaalmodel
  - taalmodel doeltaal
- Voorloper: IBM
- Zelfde n-grammodel als spraakherkenning



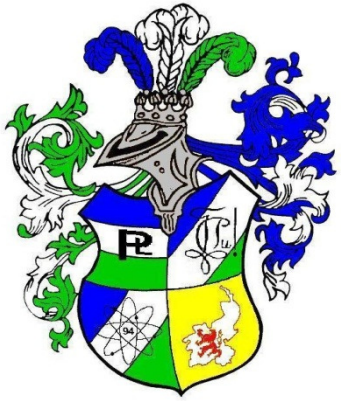
SMT (2)

- Ander type problemen: data sparsity, er zijn 'geen' regels
- Voordeel: snel en goedkoop te ontwikkelen, geen linguïstische kennis nodig
- Belangrijkste commerciële ontwikkelaars: Google, Microsoft (intern)



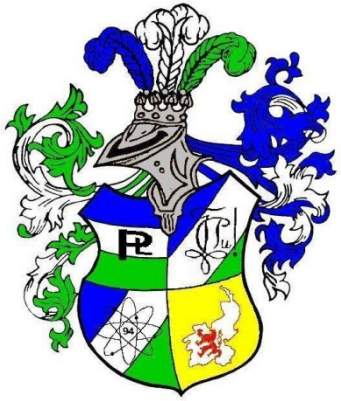
EBMT

- Variant: voorbeeldgebaseerde automatische vertaling (EBMT)
- Gedeeltelijk statistisch
- Zoekt voorbeelden in vertaling en probeert die samen te plakken
- Lijkt op wat vertaalgeheugens doen



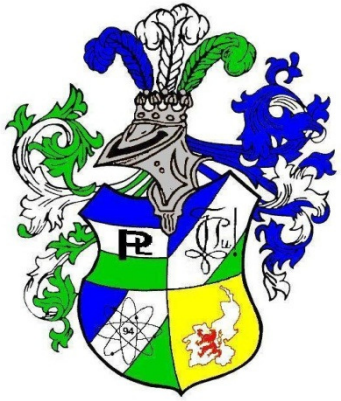
Nu

- Hybride systemen: SMT en EBMT met regelgebaseerde componenten
- Gebruik van taalkundige informatie in SMT en EBMT
- Gebruik van gelijkende i.p.v. parallelle corpora



De toekomst

- Hybride systemen
- Ontwikkeling van benodigde data (cfr. Stevinproject van de NTU)
- Combinatie van domeinen: automatische vertaling van spraak, NLU in dialoogsystemen, automatische ondertiteling, information retrieval in telefoongesprekken (spionage)



Meer info

- Bijvoorbeeld:

  Daniel Jurafsky & James H. Martin, *Speech and Language Processing*, 2nd edition, Pearson Education, 2008