Bachelorproject
Multi-agent model van taalverandering

Gert van Valkenhoef
Begeleider: Bart de Boer

30 januari 2007

Inhoud presentatie

Inleiding

- Er is al multi-agent onderzoek naar taalverandering
- Huidig onderzoek verklaart veelal *convergentie*
- We willen ook het continu veranderende karakter van taal verklaren
- En: harde, geleidelijke of verplaatsende taalgrenzen

D. Barr [Barr, 2004]

- Spatieel en non-spatieel model
- Agents leren word-meaning paren
- Leren gebeurt slechts door lokale interacties
- Toch is er sprake van globale convergentie
- Conclusie: common knowledge niet nodig

R. Axelrod [Axelrod, 1997]

- Spatieel model
- Cultuur bestaat uit n features, met m traits
- Iedere agent krijgt een random cultuur
- Interacties met de directe buren
- Cultuur (-verschillen) bepalen interactiekans
- Lokale convergentie en globale polarisatie
- Conclusie: alleen lokale interacties voldoende voor ontstaan homogene taalgebieden

Het 'threshold' probleem

- De voorgaande modellen zijn convergent
- Diversiteit introduceren d.m.v. random mutaties werkt niet
- Kleine variaties 'middelen uit', een eenmaal geconvergeerd model is dus zeer star
- Het threshold probleem: hoe zorg je ervoor dat een verandering groot genoeg is om invloed te hebben?

Andere problemen

- Hoe verklaar je dat taal voortdurend verandert?
- Hoe verklaar je het ontstaan en verdwijnen van taalgebieden?
- Hoe verklaar je scherpe (Vlaanderen-Wallonië) en geleidelijke (Groningen-Limburg) taalovergangen?
- Hoe verklaar je verplaatsende taalgrenzen?
- Convergente modellen leveren vaak 'kluizenaars' op: agents die niet met de rest van hun omgeving om kunnen gaan

D. Nettle [Nettle, 1999]

- Sociale structuur, met 'hyperinvloedrijke' personen
- Invloed van leeftijd, afstand, status, aantal sprekers
- Lost threshold problem op
- Maar: zeer eenvoudige (binaire) taalrepresentatie
- Pieter vertelt hier meer over

Doel onderzoek

- Binnen een convergent model verklaring bieden voor het volgende:
- Het voortdurend veranderende karakter van taal, ook binnen een gebied dat dezelfde taal spreekt
- Het ontstaan of verdwijnen van (sub-)culturen
- Geleidelijke of juist scherpe grenzen
- Geen 'kluizenaars'

De wereld

- Een spatieel model: een $m \times n$ grid
- In ieder vakje leeft een *agent*
- De wereld is toroïdaal: iedere agent heeft evenveel buren
- Iedere agent heeft een eigen *taal* (of: cultuur)
- Iedere agent heeft een 'neighborhood' van N agents

Taalrepresentatie

- Een taal bestaat uit f *features*
- Iedere feature kan t *traits* (waarden) hebben
- Dit is dezelfde representatie als [Axelrod, 1997]
- Anders dan Axelrod, neem ik verschillen euclidisch, niet alles of niets
- Dus de trait 1 verschilt meer van trait 5 dan van 3, maar voor het 'overnemen' van traits dat niet van belang

Bepalen eigen 'groep'

Hoe ziet de agent zijn omgeving

- Lokale complete-link clustering, met een bepaalde cutoff τ
- Clustert de neighborhood (inclusief de agent zelf)
- Levert groepen met maximaal τ verschil tussen agents
- Clustergrootte t.o.v. neighborhood geeft comfort: c
- Maximale afstand t.o.v. τ geeft spread: s
- Boredom: $b = (1 - s) \times c$

Gedragingen

- Random mutatie (vaste kans): een feature krijgt random trait
- Paniek (afhankelijk van c): neem een taal over van andere groep
- Convergentie (afhankelijk van b): neem feature over van gelijkende groepsgenoot
- Divergentie (afhankelijk van b): neem zeldzame feature over van groepsgenoot ('novelty drive')

Meten is moeilijk

- Hoe visualiseer je een $m \times n$ wereld, met $f \times t$ mogelijke talen per agent?
- Welke statistieken zijn zinnig om iets te zeggen over taaldiversiteit?
- Hoe identificeer je taalgebieden?

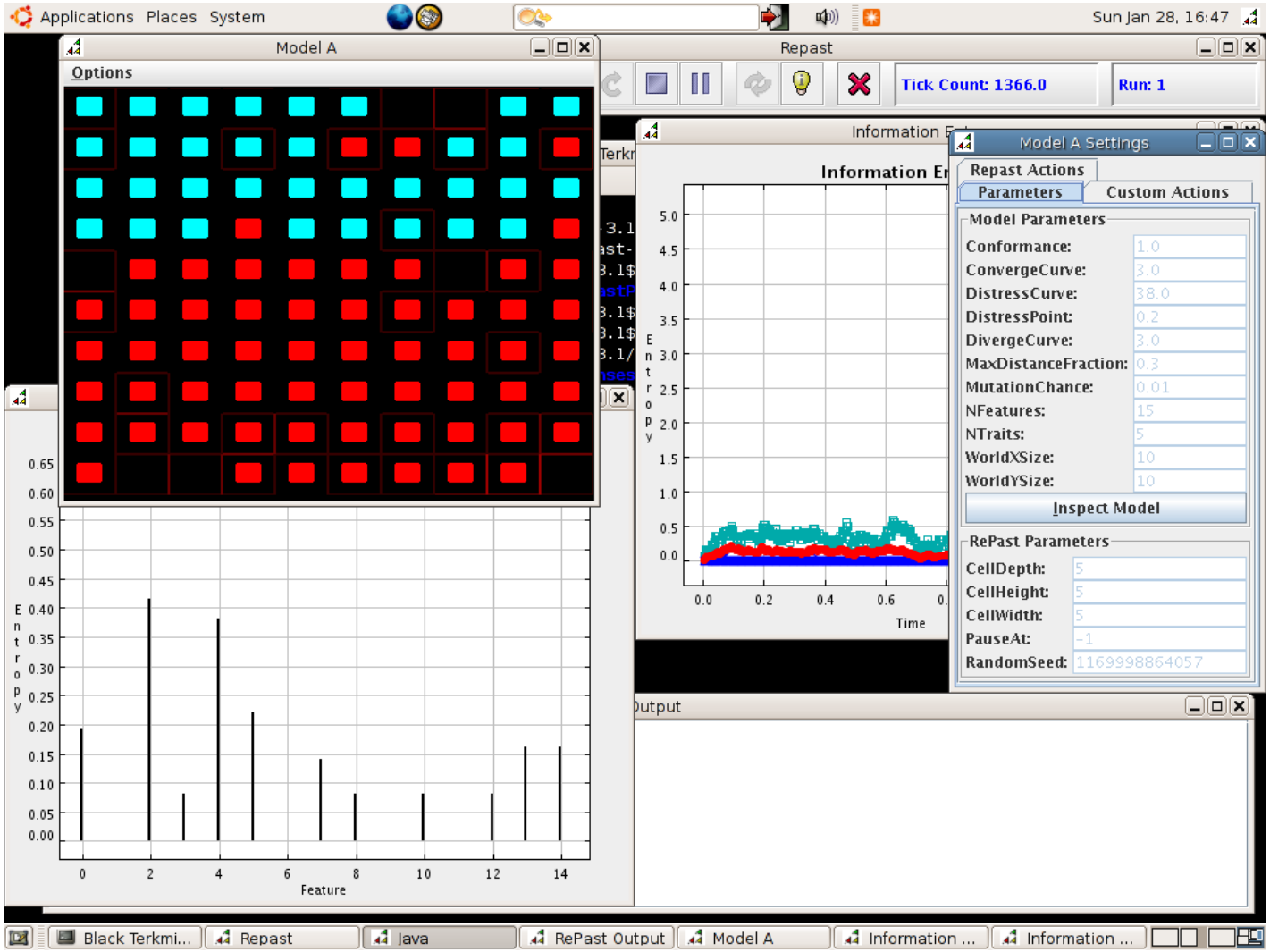
Entropie

- Entropie is een maat voor 'informatie', ofwel de hoeveelheid variatie die een bepaalde variabele heeft
- Entropie per feature, dit geeft een bar chart waarin convergentie of divergentie per feature bekeken kan worden
- Daarnaast ook een plot van de ontwikkeling van minimale, maximale en gemiddelde entropie in het model tegen de tijd

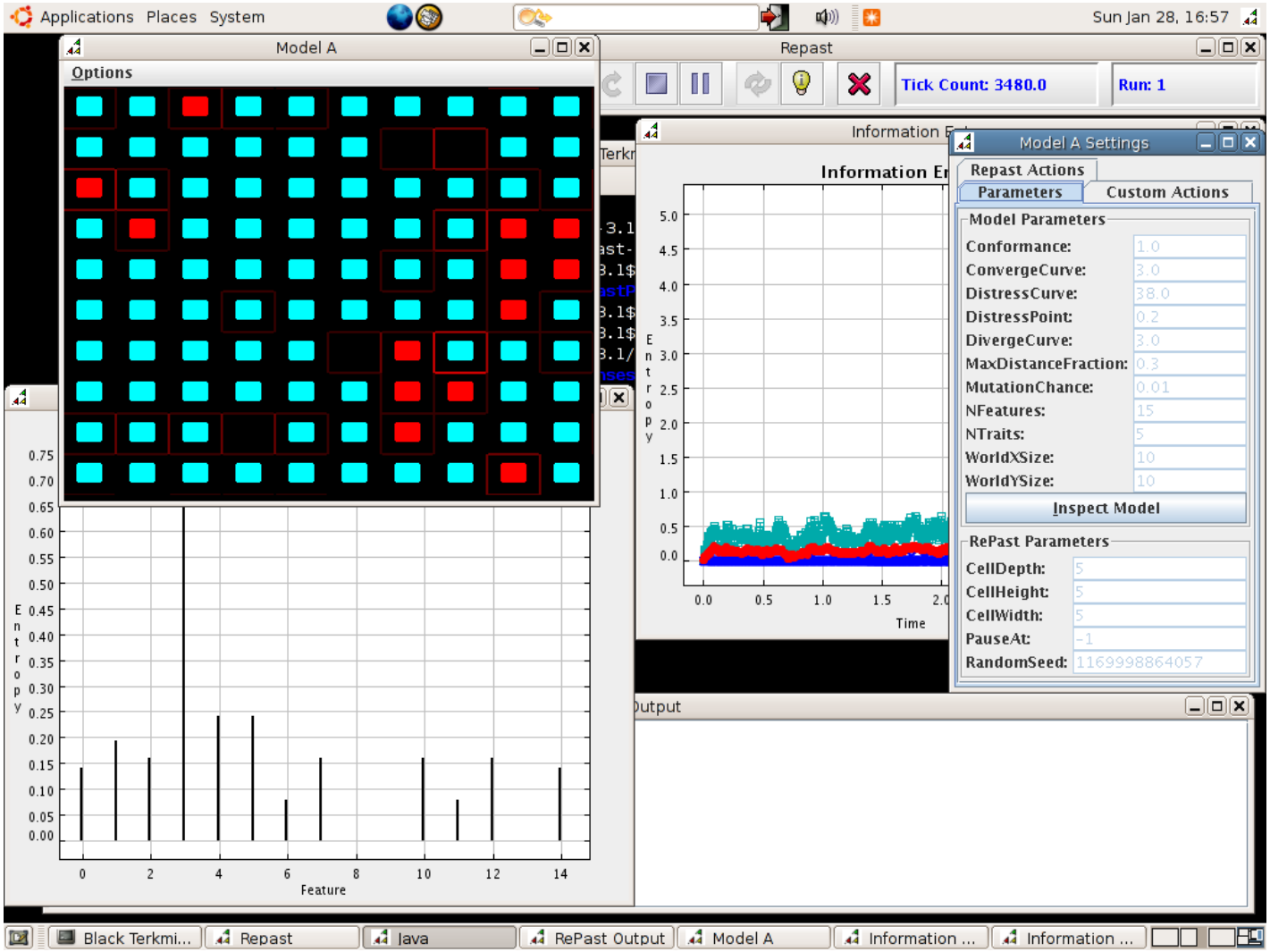
Clustering

- Een aangepaste ISODATA die het CAIC (Consistent Akaike Information Criterion) gebruikt om het aantal clusters k en goede seed points voor een K-Means clustering te bepalen [Ball & Hall, 1965] [Jain et al., 1999] [Carman & Merickel, 1990]
- Veel interessante statistiek, die ik zal bewaren voor de scriptie
- Levert een vrij stabiele clustering op, zonder parameters
- Geef maximaal X clusters met minimaal Y members een kleurtje
- Ook hier zijn statistieken uit af te leiden: aantal clusters, cluster spread, etc

Visualisatie: voorbeeld 1/2



Visualisatie: voorbeeld 2/2



Resultaten

- Ik ben nog niet helemaal klaar met de experimenten en analyse..
- Dus enigszins terughoudend, maar het blijkt:
  - Het convergente gedrag houdt de diversiteit beperkt
  - en: brengt een 'random' situatie naar een homogenere
  - In een vrij homogeen gebied ontstaan en verdwijnen 'subculturen'
  - Het taalgebied als geheel verandert voortdurend
  - Ik ben nog op zoek naar een manier om dit statistisch verantwoord te onderbouwen

Conclusie/discussie

- Multi-agent systems lijken inderdaad de gewenste verschijnselen te kunnen modelleren, maar:
- De analyse en visualisatie van dergelijke modellen is een kunst op zich
- De vraag blijft in hoeverre de systemen de werkelijkheid benaderen
- Wel helpt het, om te weten wat voor mechanisme er achter de verschijnselen die we bij taalverandering veronderstellen *kan* veroorzaken


Axelrod, R. (1997).
The dissemination of culture: A model with local convergence and global polarization.
The Journal of Conflict Resolution, 41(2):203–226.


Ball, G. H. en Hall, D. J. (1965).
Isodata, a novel method of data analysis and pattern classification.
Technical report, Stanford Research Institute, Springfield, USA.


Barr, D. J. (2004).
Establishing conventional communication systems: Is common knowledge necessary?
Cognitive Science, 28(6):937–962.


Carman, C. en Merickel, M. (1990).

Supervising iso-data with an information theoretic stopping rule.
Pattern Recognition, 23(1/2):185–197.


Jain, A., Murty, M., en Flynn, P. (1999).
Data clustering: A review.
ACM Computing Surveys, 31(3):264–323.


Nettle, D. (1999).
Using social impact theory to simulate language change.
Lingua, 108(2-3):95–117.