# Referentiële onzekerheid, computermodellen en semantische kindertaalcorpora

Barend Beekhuizen[1], Afsaneh Fazly[2], Aida Nematzadeh[2] & Suzanne Stevenston[2]

[1]Universiteit Leiden [2]University of Toronto

TIN-dag, 9 februari 2013

# Vragen

## Onderwerp

Computationele cognitieve modellen van het leren van woord-betekenisparen: data.

## Vraag #1

Hoe komen we aan semantische input voor dergelijke modellen?

## Vraag #2

Hoe kan inzicht hierin gebruikt worden om oude claims te herevalueren? In het bijzonder: kunnen verwervers de betekenis van relationele termen uit de situatie oppikken?

# Data?

- Cross-situationele modellen van lexicale-betekenisverwerving[1]
- Bron van de semantiek: de directe situatie
- Een gemiddeld CHILDES-corpus bevat dat niet.

[1] Siskind 1996, Xu & Tenenbaum 2000, Roy & Pentland 2002, Yu & Ballard 2003, Fazly, Alishahi & Stevenson 2010

# Data?

- Cross-situationele modellen van lexicale-betekenisverwerving[1]
- Bron van de semantiek: de directe situatie
- Een gemiddeld CHILDES-corpus bevat dat niet.
- Dus: kunstmatig genereren van semantiek.
  - Ieder woord is ook een semantisch symbool (Fazly, Alishahi & Stevenson 2010)
  - Gebruik van WordNet e.d. (id., 2008)
- Je kan er grote kwantiteiten data mee maken

[1] Siskind 1996, Xu & Tenenbaum 2000, Roy & Pentland 2002, Yu & Ballard 2003, Fazly, Alishahi & Stevenson 2010

# Data?

- Cross-situationele modellen van lexicale-betekenisverwerving[1]
- Bron van de semantiek: de directe situatie
- Een gemiddeld CHILDES-corpus bevat dat niet.
- Dus: kunstmatig genereren van semantiek.
  - Ieder woord is ook een semantisch symbool (Fazly, Alishahi & Stevenson 2010)
  - Gebruik van WordNet e.d. (id., 2008)
- Je kan er grote kwantiteiten data mee maken
- Maar: kwaliteit van data?
  - Cognitieve beschikbaarheid van de semantiek
  - Situationele beschikbaarheid? (ruis, referentiële onzekerheid)

[1] Siskind 1996, Xu & Tenenbaum 2000, Roy & Pentland 2002, Yu & Ballard 2003, Fazly, Alishahi & Stevenson 2010

# Data?

- Cross-situationele modellen van lexicale-betekenisverwerving[1]
- Bron van de semantiek: de directe situatie
- Een gemiddeld CHILDES-corpus bevat dat niet.
- Dus: kunstmatig genereren van semantiek.
  - Ieder woord is ook een semantisch symbool (Fazly, Alishahi & Stevenson 2010)
  - Gebruik van WordNet e.d. (id., 2008)
- Je kan er grote kwantiteiten data mee maken
- Maar: kwaliteit van data?
  - Cognitieve beschikbaarheid van de semantiek
  - Situationele beschikbaarheid? (ruis, referentiële onzekerheid)
- Recente methode: annoteren video materiaal (Yu, Roy, Frank)
- Maar: hetzij beperkt tot middelgrote objecten of in het pragmatische realisme (expliciete labeling-taken).

[1] Siskind 1996, Xu & Tenenbaum 2000, Roy & Pentland 2002, Yu & Ballard 2003, Fazly, Alishahi & Stevenson 2010

# Data!

## Doel #1

Maken van situationele beschrijvingen (van eigenschappen, dingen, relaties en gedrag) voor een dataset van op video opgenomen ouder-kindinteractie die kan fungeren als een bron van woordbetekenis

# Data!

## Doel #1

Maken van situationele beschrijvingen (van eigenschappen, dingen, relaties en gedrag) voor een dataset van op video opgenomen ouder-kindinteractie die kan fungeren als een bron van woordbetekenis

- Realisaties:
  - Zulke hoge-kwaliteitsdata kan hoge-kwantiteitsdata complementeren en niet vervangen.
  - Weinig beschreven over hoe.
    >> Dit werk: gedocumenteerd en handleiding beschikbaar.

# Het blokkenspelcorpus

- ± 130 90min video's van moeder-dochter (16mo) interactie, verzameld bij Pedagogiek in Leiden
- Elk tweetal speelt een spel: stop samen verschillend gevormde blokjes in verschillend gevormde gaten in de deksel van een emmertje.
- 32 tweetallen (± 5 min. per tweetal) werden door twee codeurs met ELAN gecodeerd en door de eerste auteur getranscribeerd.

# Het blokkenspelcorpus

- $\pm$ 130 90min video's van moeder-dochter (16mo) interactie, verzameld bij Pedagogiek in Leiden
- Elk tweetal speelt een spel: stop samen verschillend gevormde blokjes in verschillend gevormde gaten in de deksel van een emmertje.
- 32 tweetallen ($\pm$ 5 min. per tweetal) werden door twee codeurs met ELAN gecodeerd en door de eerste auteur getranscribeerd.
- 175 minuten materiaal, 7842 tokens, 2492 uitingen.
- Situationele codering. Voor elk interval (3sec), codeer:
  - simpel gedrag (`grab`,`move`,`position`,`letgo`),
  - veranderingen in ruimtelijke relaties (`in`,`on`,`out`,`off`,`match`),
  - objecten (`block`,`bucket`,`mother`,`table`)
  - eigenschappen (`triangular`,`square`,`red`,`blue`)
- Gestructureerd: `grab(mother,(red,square,block))`
- Hoge intra- & interannotator agreement (bijna alle $\kappa > 0.8$)

# Voorbeeld

Table: Een voorbeeld van de dataset. De afkortingen geven eigenschappen van blokken en gaten aan (kleur & vorm)

| tijd | type | codering/transcriptie |
|---|---|---|
| 0m0s | **situatie** | |
| | **taal** | *een. nou jij een.* |
| 0m3s | **situatie** | position(mother, toy, on(toy, floor)) grab(child, b-ye-tr) move(child, b-ye-tr, on(b-ye-tr, floor), near(b-ye-tr, ho-ro)), mismatch(b-ye-tr, ho-ro) |
| | **taal** | *nee daar.* |
| 0m6s | **situatie** | point(mother, ho-tr, child) position(child, b-ye-tr, near(b-ye-tr, ho-ro)) mismatch(b-ye-tr, ho-ro) |
| | **taal** | *nee lieverd hier past ie niet.* |

# Verwerven van lexicale betekenis

- Hoe leer je de betekenis van een woord
  - Cross-situationeel waarnemen van objecten, eigenschappen, relaties
  - Lijkt onvoldoende (i.h.b. voor relationele termen: werkwoorden, voorzetsels)
    - Aantal mogelijkheden is enorm (Gentner 1978)
    - Veel acties en relaties die benoemd worden, vallen niet samen met uiting (Gleitman 1990)
  - 'Bootstrapping' met syntactische structuur (Gleitman 1990), intentionaliteit (Tomasello 2003), ...

# Verwerven van lexicale betekenis

- Hoe leer je de betekenis van een woord
  - Cross-situationeel waarnemen van objecten, eigenschappen, relaties
  - Lijkt onvoldoende (i.h.b. voor relationele termen: werkwoorden, voorzetsels)
    - Aantal mogelijkheden is enorm (Gentner 1978)
    - Veel acties en relaties die benoemd worden, vallen niet samen met uiting (Gleitman 1990)
  - 'Bootstrapping' met syntactische structuur (Gleitman 1990), intentionaliteit (Tomasello 2003), ...

**Doel #2**

De data gebruiken om de claim te herevalueren: is betekenis echt moeilijk uit de situatie te halen, en waar ligt dat aan?

# Het model

- Fazly, Alishahi & Stevenson (2010) incrementeel, probabilistisch model dat woord-betekenisassociaties leert.
- Woorden worden aan betekeniselementen gekoppeld met een gewicht dat afhangt van eerdere ervaring en de eerdere ervaringen met de andere woorden en betekenissen
- De ervaring wordt vervolgens geüpdated met dat gewicht.
- Voor elk woord kan je dan een voorwaardelijke waarschijnlijkheid van verschillende betekenissen bepalen

# Voorbereiding data

- Representaties zijn gestructureerd, model neemt verzamelingen van primitieven, dus we maken ze plat: `grab(mother,(red,square,block))` → `{grab,mother,red,square,block}`
- De beschikbare betekenis is de verzameling van de platgemaakte representaties in het interval waarin de uiting begint
- Woorden als lemma's

# Evaluatie

- Geen 'gouden lexicon', dus zelf maken voor 'betekenisvolle' woorden ($n = 41$):
  - Objectlabels: *blok* betekent `block`
  - Eigenschappen: *rood* betekent `red`
  - Ruimtelijke relaties: *op* betekent `on`
  - Acties: *pakken* betekent `grab`, *stoppen* betekent {`move`,`in`}

# Evaluatie

- Geen 'gouden lexicon', dus zelf maken voor 'betekenisvolle' woorden ($n = 41$):
  - Objectlabels: *blok* betekent `block`
  - Eigenschappen: *rood* betekent `red`
  - Ruimtelijke relaties: *op* betekent `on`
  - Acties: *pakken* betekent `grab`, *stoppen* betekent {`move`,`in`}
- Maat
  - Average Precision ($AP$): Hoe zijn de juiste betekenissen gerangschikt m.b.t. hun associatiescore t.o.v. de onjuiste betekenissen?

# Resultaten

Table: Resultaten van experiment 1. Gegeven zijn gemiddelde *AP* waarden per klasse

| | eigenschap | object | ruimtelijk | actie | **totaal** |
|---|---|---|---|---|---|
| *AP* | 0.81 | 0.25 | 0.19 | 0.15 | **0.31** |

- Rangschikking gaat goed voor eigenschappen (kleur, vorm), maar vrij slecht voor andere klassen.

# Interpretatie

| | eigenschap | object | ruimtelijk | actie | **totaal** |
|---|---|---|---|---|---|
| *AP* | 0.81 | 0.25 | 0.19 | 0.15 | **0.31** |

- Herevaluatie bevestigt Gleitmans bevinding:
  Eigenschappen $>$ objectlabels $>$ beide relationele klassen
- Waarom zijn de laatste drie zo moeilijk te leren?
  1. Juiste betekenis is afwezig in de situatie $S$ (ruis)
  2. Onjuiste betekenissen zijn structureel aanwezig in $S$ (referentiële onzekerheid)
  3. Juiste betekenis is ook aanwezig in veel andere $S$s (?)
- Combinatie! Voor eigenschappen gelden 2) en 3) ook.

# Experiment 2

- Ruis: afwezige, juiste betekenissen
- Misschien is de reikwijdte van het interval te beperkt?
- Leerders letten ook op gebeurtenissen rondom de uitingen, vooral erna (Tomasello & Kruger 1992)

vooruit Verwerver let op situatie op moment van uiting $U_i$ tot volgende uiting $U_{i+1}$

achteruit Verwerver let op situatie op moment van vorige uiting $U_{i-1}$ tot huidige uiting $U_i$.

# Experiment 2

| $W$ | eigensch. | object | ruimtelijk | actie | **totaal** |
|---|---|---|---|---|---|
| $U_i$:$U_i$ | 0.81 | 0.25 | 0.19 | 0.15 | **0.31** |
| $U_{i-1}$:$U_i$ (**achteruit**) | 0.80 | 0.17 | 0.20 | 0.14 | **0.31** |
| $U_i$:$U_{i+1}$ (**vooruit**) | 0.79 | 0.41 | 0.22 | 0.20 | **0.39** |

- $U_i : U_{i+1}$ geeft kleine verbetering voor drie categorieën
  - Vooruit kijken is informatief: meer juiste betekenissen gevonden
  - maar: referentiële onzekerheid wordt niet veel groter
- Achteruit kijken heeft geen positief effect

- Goede data is schaars
- Maar je kan handmatig coderen, als het met een methode doet, om
  1. te kijken of de aannames achter het 'synthetiseren' van semantische data kloppen.
  2. kleinschalige experimenten te draaien

- Goede data is schaars
- Maar je kan handmatig coderen, als het met een methode doet, om
  1. te kijken of de aannames achter het 'synthetiseren' van semantische data kloppen.
  2. kleinschalige experimenten te draaien
- Werkwoords- en voorzetselbetekenissen zijn inderdaad moeilijker te verwerven dan nominale betekenissen, die op hun beurt weer lastiger zijn dan de kleur- en vormbetekenissen.
- Een vooruitkijkend breder interval van situaties helpt een beetje
- Maar ook 'gestructureerd leren' (cf. Gleitman) moet een rol spelen – taak voor modelleurs om dit te formaliseren

- Goede data is schaars
- Maar je kan handmatig coderen, als het met een methode doet, om
  1. te kijken of de aannames achter het 'synthetiseren' van semantische data kloppen.
  2. kleinschalige experimenten te draaien
- Werkwoords- en voorzetselbetekenissen zijn inderdaad moeilijker te verwerven dan nominale betekenissen, die op hun beurt weer lastiger zijn dan de kleur- en vormbetekenissen.
- Een vooruitkijkend breder interval van situaties helpt een beetje
- Maar ook 'gestructureerd leren' (cf. Gleitman) moet een rol spelen – taak voor modelleurs om dit te formaliseren
- Natuurlijke, observationele data geeft complementair inzicht

Dank u!