Agent-Based Modeling in Evolutionary Linguistics

Remi van Trijp

Sony Computer Science Laboratory Paris
remi@csl.sony.fr

Guest Lecture: Trends in AI
Brussels, 15 February 2013
“Trend”

- **Noun; (I)** A general direction in which something is developing or changing: 
  
  *an upwards trend in our company’s sales*

- **(II)** A fashion: *the latest trends* (~ *trendy*)

- **Origins:** Germanic, e.g. Old English *trendan* (‘to turn in a general direction’)
  
  > Became metaphor in 19th Century
  > Spatial sense disappeared
Trends in AI

• “Automaton”
  > Greek origins, ‘acting of one’s own will’
  > autonomous systems

• “Andr-oid”
  > Greek origins, ‘like a (hu)man’

• “Robot”
  > Introduced by Czech writer Karel Čapek in a theater play in 1920
  > stems from *robota* (‘serf labor’, ‘corvée’)

Autonomous Agents

• 250 BC: Chinese *Lie Zi* book describes human-like automaton

• 50 BC: Greek mathematician *Hero of Alexandria* invents vending machines, machines for pouring wine, ...

• 13th Century: *Al-Jazari* invents a musical robot band, drink-serving waitress, ...

• ...

...
Autonomous Agents
Hypothesis

- **Language** is a key ingredient to bootstrap the intelligence of autonomous agents to human-like cognition

- We must understand how language...
  > is processed and learned
  > emerges and evolves
How can agents learn to use language?

But does this approach work??

language model

agent

statistical analysis

data

Growing By A Factor Of 44

2009: 0.8 Zb

2020: 35.2 Zettabytes
Dominant approach to language processing
Dominant approach: probabilistic

**Mostly solved:**
- Spam detection
- Parts-of-speech tagging
- Named entity recognition

**Progress being made:**
- Sentiment analysis
- Word sense disambiguation
- Information extraction

*buy VI aGrA...*

*Einstein met the UN officials.*

*The waiter ignored us for 20 minutes.*

*I need a new mouse for my pc.*

*You’re invited to our party on Friday, March 23.*
Dominant approach: probabilistic

**Still problematic:**
- Question-answering
- Summarization and paraphrasing
- Dialog

**Really, really hard**
- Human-machine interaction
- Linguistic creativity

How effective is ibuprofen in reducing fever in patients with acute febrile illness?
Some Observations in Language
Crazy English

- Your nose can run and your feet can smell.
- Most bathrooms do not have a bath.
- Apartments are actually together.
- Women can man a station, but men can’t women one
- A writer writes and a singer sings. But fingers don’t fing, hammers don’t ham and grocers don’t groce
Crazy English

• Old news
• Awfully good
• A small fortune
• Loose tights
• Open secret
• Recorded live
• act naturally
• Microsoft Works
Crazy English

• Fatal attack leaves no survivors
• I’m going to proceed ahead
• 100 percent pure
• A new innovation
• He had no peers or equals
• Talk about your past experiences
• At this point in time
Crazy Dutch/Flemish

• Hij zit in de kelder.
  *He is in the basement.*
  *(lit.: he sits in the basement)*

• Hij kwam op mij af gelopen.
  *He came up to me.*
  *(lit.: He came up me down running)*

• Ze is volslank.
  *She’s a bit chubby.*
  *(lit.: she is full-skinny)*
Crazy French

• Est-ce que c’est vrai qu’elle est française?
  *Is it true that she’s French?*
  (lit: *Is it that it is true that she is French?*)

• Au jour d’aujourd’hui.
  *The day today.*
  (lit.: *At the day of at the day of today*).

• C’est terrible!
  *That’s great!*
  (lit.: *it is terrible*)
Language is embodied

You find it at the **back**.
The building **faces** south.
**Face** your demons.
It is **behind** you.
It is in **front** of you.
The **heart** of the problem.
I’m **knee**-deep in trouble.
The **eye** of the storm.
The **head** of the company.
I have a **gut** feeling.
Language is embodied

- Artificial
  > from Latin *artificium* (‘made by man’)
  > *ars* (‘art/craft’) + *facere* (‘do’)
  > *ars* is related to *arm*

- Intelligence
  ~ Latin *intelligere* (‘to understand’)
  ~ *inter-* (‘between’) + *legere* (‘choose/pick’)
  (related to *lecture*)

- Guest lecture
  = things selected by a stranger
Language is cultural

- English: The cat is in front of the tree.
- Hausa: The cat is behind the tree.
- Tonga: The cat is at the seaside of the tree.
Language is cultural
Language is cultural

(Davidoff, Roberson, et.al. *Nature*, 1999)
Language is cultural

“Now! That should clear up a few things around here!”
Evolutionary Linguistics

Origins and evolution of language

*The Fall of Adam, 1470*
Hugo Van der Goes

*Tower of Babel, 1563*
Pieter Breughel The Elder
Cultural Linguistic Selection

- Variation
- Testing variants against selection criteria
  - Communicative success
  - Cognitive effort
  - Expressivity
  - Social conformity
  - ...

Language Games
Language games

- **Routinized interactions** between two or more agents in a *locally situated* environment
- Integrate the various activities required for dialogue
- Provide a communicative context
Let agents play language games

Sony Qrio

Myon
(Humboldt U. Berlin)
A “Whole Systems” Approach

Embodiment and vision

Interaction with the real world

Conceptualization

Language
Semiotic Cycle

Speaker

- world model
- conceptualization
- semantic structure
- verbalization
- syntactic structure

- utterance

Hearer

- world model
- interpretation
- semantic structure
- parsing
- syntactic structure
Problem-solving: two-level processing system

Meta-layer
- diagnostics
- repairs
- consolidation

Routine layer
Language technologies

Babel2
- Sensori-motor processing
- Rich semantics
- Fluid grammars

www.fcg-net.org
Sensori-motor processing

- **Grounding**
  Connecting representations to reality through the robot’s sensori-motor apparatus

- **Embodiment**

  Open-ended: new situations, noisy perception, ...
Rich semantics

• **Conceptualization**
  Planning what to say.

• **Interpretation**
  Plan “execution”
  (“apply” the meaning to the real world)

Open-ended: new meanings, categories, ...
Incremental Recruitment Language (IRL)

- Speaker uses language to achieve a communicative effect in the hearer’s mind
- Procedural semantics: meaning is like a semantic program that the speaker wants the hearer to execute
The speaker wishes to draw the hearer’s attention to this object using language.
Referential game

get-context

- Red circle
- Green square
- Blue triangle
- Red square
Referential game

get-context

filter-by-prototype [SQUARE]
Referential game
Referential game

- get-context
- filter-by-prototype [SQUARE]
- filter-by-color [GREEN]
- select-unique-object
IRL networks:
cognitive operations + bind statements

(get-context ?context)

(filter-set-class ?set-1 ?context ?class)

(bind object-class ?class square)


(bind color ?color green)

(select-entity ?ref ?set-2 ?selector)

(bind selector ?selector unique)
Fluid grammars

- **Production**
  Verbalizing a meaning into an utterance

- **Parsing**
  Parsing an utterance into a meaning

  Open-ended: new words, grammatical constructions, personal preferences, ...
Fluid Construction Grammar (FCG)

- Grammar expresses conceptualizations through meaning-form mappings
- Construction grammar

**Meaning**

(bind color ?color red)

**Form**

“rouge”
Fluid Construction Grammar (FCG)

- Intermediary layer of semantic and syntactic categories

```
Meaning
filter-by-color, filter-by-shape, filter-by-spatial-category, ...
```

```
Form
"red", "red one", "redden"
```

```
Sem-cat
modifier, object, ...
```

```
Syn-cat
adjective, noun, ...
```

grammatical constructions
Example 1

The Naming Game

Power of Self-Organization

Termite Mound (Cape York, Australia)

Sagrada Familia (Barcelona)
Example II
The Color Game

Detecting Color
Example III
The Emergence of Grammar

Model

• design language game in which colour is functionally relevant

• each use of a colour category is operationalised as a different language strategy

  • full colour space: all dimensions of the colour space are used

  • brightness: only brightness ($L^*$) dimension of colour space is used

• add layer of linguistic selection of language strategies

Colour Naming Game

Full Colour Space Strategy: production & interpretation

Full Colour Space Strategy: production & interpretation

Full Colour Space Strategy: production & interpretation

Full Colour Space Strategy: production & interpretation

Full Colour Space Strategy: production & interpretation

Full Colour Space Strategy: production & interpretation

Full Colour Space Strategy: production & interpretation

Full Colour Space Strategy: production & interpretation

Full Colour Space Strategy: production & interpretation

Full Colour Space Strategy: adoption & alignment

“pugami”?

Full Colour Space Strategy: adoption & alignment

"pugami"?

Full Colour Space Strategy: adoption & alignment

“pugami”

Full Colour Space Strategy: adoption & alignment

Full Colour Space Strategy: adoption & alignment

Full Colour Space Strategy: expansion

Full Colour Space Strategy: expansion

Full Colour Space Strategy: evolving category system

Full Colour Space Strategy: evolving category system

Full Colour Space Strategy: resulting dynamics

Full Colour Space Strategy: evolving colour system

Brightness Strategy

• instead of using 3 dimensions ($L^*u^*v^*$) of colour space, the prototypes only specify value on lightness channel ($L^*$)

• similar production/interpretation functions

• similar adoption/alignment functions

• similar expansion function

Brightness Strategy: resulting dynamics

Brightness Strategy: evolving language system

1000 3000 4000
colour lexicon after X interactions

Model

• design language game in which colour is functionally relevant

• each use of a colour category is operationalised as a different language strategy
  
  • brightness: only brightness ($L^*$) dimension of colour space is used
  
  • full colour space: all dimensions of the colour space are used

• add layer of linguistic selection of language strategies

Linguistic Selection of Language Strategies

- higher expressive power
- shared in population
- lower cognitive effort

selective advantage

context

language strategies

language system

communicative success

context

Linguistic Selection: implementation

- production/interpretation
  - strategy preferred by linguistic item
  - otherwise re-interpretation using strategies sorted based on their fitness
- alignment
  - use strategy that was used during production/interpretation
- adoption/expansion
  - use strategy that is most fit and that is sufficient for current context

Linguistic Selection: one strategy prevails
Linguistic Selection: co-existence of strategies

![Graph showing communicative success, strategy coherence, fitness, and usage over number of interactions/agent.](image-url)
Current work: combination of strategies

Figure 2. Aggregate data on sitá-, the red-focused category. Columns are rearranged from Figure 1, and the irrelevant columns 8 through 32 are deleted. (a) The root sitá-, combining all modifiers, aggregated from all 15 informants. (b) Sitakame (very red), aggregated from 9 informants. (c) Sitáname (somewhat red), aggregated from 9 informants. (d) Sitánanti (only slightly red), aggregated from 2 informants.

Conclusion

- Language strategies provide a methodology to study certain observations in the evolution of language.

- Coordination of language strategies can be orchestrated by communicative success.
References


Argument Structure

*English*

The boy walks to the girl.

The left scene.
Argument Structure

Riau Indonesian (David Gil, pers. comm.)

Cowok lari cewek.

???
The emergence of case and argument structure

- FCG needs to be capable to handle language systems that go from:

  - Participant Structure
    - event-specific participant roles
      - (giver, gift, receiver, ...)
  - Argument Realization
    - forms
      - (give, send, ...)

  ???

The emergence of case and argument structure

- To:

**Participant Structure**

- Event-specific participant roles (giver, gift, receiver, ...)
  - Depending on context

- Semantic roles (Agent, Patient, Beneficiary, ...)
  - Depending on context

**Argument Realization**

- Case markers (der, die, das, ...)
  - Depending on context

- Case (categories) (nominative, accusative, ...)
  - Depending on context

**Argument structure constructions**
The emergence of case and argument structure

- Without predefined semantic or syntactic categories...
- With massive variation in the speech community...
- With great uncertainty about meaning/function
- With various degrees of entrenchment
- ...

The Case Experiments

Description Games about real-world events

- Case Strategy
  - Innovation
  - Learning
  - Alignment

- Case Marking System
  - Lexical and grammatical constructions

- Linguistic Behavior
  - Utterances

- Communicative Outcome

Feedback
An artificial language

Population

Agent-1
Agent-2
Agent-3
Agent-4
Agent-5
Agent-6
Agent-7
Agent-8
Agent-9
Agent-10
An artificial language

Population

Agent-1
Agent-2
Agent-3
Agent-4
Agent-5
Agent-6
Agent-7
Agent-8
Agent-9
Agent-10

Jack-cui walk-to Jill-ge.

Yes.
An artificial language

• *Jack-cui* walk-to *Jill-ge*
  jack-sem-role-6 walk-to Jill-sem-role-26
  ‘Jack walks to Jill.’

• *house-lu* move-inside *boy-cui*
  house-sem-role-10 move-inside boy-sem-role-6
  ‘The boy moves inside the house.’
Case marking for event structure

(world model)

(boy ?puppet-x)
?puppet-x = ?walker

(walk-to ?ev)
(walk-to-1 ?ev ?puppet-x)
(walk-to-2 ?ev ?puppet-y)

(boy jack)
(girl jill)
(walk-to ev-1)
(walk-to-1 ev-1 jack)
(walk-to-2 ev-1 jill)

(boy -cui)
walk-to
girl -ge

?puppet-y = ?walked-to
Experimental set-up

- Population of agents
- Given lexicon, but no grammar
- Play description games
- Problem-solving approach:
  - Diagnostics
  - Repair strategies
  - Consolidation strategies
Comparative experiments

<table>
<thead>
<tr>
<th>Stage</th>
<th>Diagnostics and repairs</th>
<th>Detecting and resolving variable equalities</th>
<th>Invention and adoption of new markers</th>
<th>Reuse and generalisation of existing markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1</td>
<td>Baseline experiment 1</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Stage 2</td>
<td>Baseline experiment 2</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Stage 3</td>
<td>Baseline experiment 3</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
Baseline: a lexical language

Baseline experiment 1: population size = 2, context size = 5

The graph shows the average communicative success and average cognitive effort over language games. The x-axis represents the number of language games, ranging from 0 to 500. The y-axis represents the communicative success/cognitive effort ratio, ranging from 0 to 1. The graph includes error bars indicating the variability in the data.
No marking in natural languages

- Riau Indonesian
  - ayam   makan
    chicken   eat

- English
  - the shooting of the hunters
Baseline: a lexical language

• Why would grammar be useful for communication?
  • reducing cognitive effort
  • avoiding ambiguity
  • increasing expressiveness (generalization)
The experiments: stage 2

- **Stage 2: specific marking**

  - **diagnostic:** re-entrance for diagnosing cognitive effort and ambiguity
    > “boy walk”
    > cognitive effort needed

  - **repair strategy:** invention and adoption strategy
    > “boy -bo walk”

  - **alignment strategy:** lateral inhibition
    > walk-to-1 <--0.6--> “-bo”
    > walk-to-1 <--0.4--> “-ka”
The experiments: stage 2

Stage 2: Verb-specific marking

- conceptualisation
  - semantic structure
  - production
  - repair needed?
    - yes
      - utterance
    - no
      - speaker interpretation
        - semantic structure
        - speaker parsing
        - re-entrance
- transmission to hearer
The experiments: stage 2

Baseline experiment 2: population size = 10, context size = 5, equal frequency

Stage 2: Verb-specific marking
The experiments: stage 2

Baseline experiment 2: population size = 10, context size = 5, equal frequency

average number of specific markers

innovation +
learning  
alignment
The experiments: stage 2

Baseline experiment 2: population size = 10, context size = 5, equal frequency

markers for `walk-to-1’ known by agent 1
The experiments: stage 2

- Stage 2: specific marking in natural languages

  - thăm cà bin maakrunghêep (Thai)
    he will fly come Tokyo
    “He will fly to Tokyo.”

  - A man comes pulling a goat.
The experiments: stage 2

- Why isn’t this strategy enough?
  - no generalization
  - explosion of inventory size
The experiments: stage 3

• Stage 3: extension through analogy

  • **diagnostic**: re-entrance for diagnosing cognitive effort
    > “boy walk”
    > cognitive effort needed

  • **repair strategy**: try to reuse existing marker, else invent or adopt a new one
    > “boy -bo walk”
    > “boy -bo move”

  • **alignment strategy**: lateral inhibition
    > walk-to-1 <--0.6--> “-bo”
    > walk-to-1 <--0.4--> “-ka”
The experiments: stage 3
The experiments: stage 3

• Analogy
  
  • First try to find an existing and analogous marker
  
  • Always try markers with highest type frequency first
  
  • If no analogy could be found, invent an existing marker
The experiments: stage 3

Baseline experiment 3b: population size = 10, context size = 5, equal frequency

Stage 3: Analogy
The experiments: stage 3

Baseline experiment 3b: population size = 10, context size = 5, equal frequency

innovation
learning
alignment fails

language games
The experiments: stage 3

• Why does it fail?

  • **polysemy:** many-to-one mappings of semantic roles

  • move-1 \[<-0.6->\] -bo
    move-inside-1

    move-inside-1 \[<-0.5->\] -ka
    move-outside-1

  • The alignment strategy is not fine-grained enough!
The experiments: stage 3

Baseline experiment 3c: population size = 10, context size = 5, equal frequency

average number of markers for semantic roles
average number of specific markers
average total number of participant roles covered
The experiments: stage 3

- Convergence: yes, but (almost) no generalization
  - Fine-grained strategy does not prefer more general roles
  - Each new role has to be negotiated as if it were a new lexical item
- New strategy:
  - token frequency during processing
  - type frequency during semantic extension
The experiments: stage 3

Baseline experiment 3d: population size = 10, context size = 5, equal frequency

Stage 3: Analogy
The experiments: stage 3

Baseline experiment 3d: population size = 10, context size = 5, equal frequency

average number of markers for semantic roles
average number of specific markers
average total number of participant roles covered

Stage 3: Analogy
The experiments: stage 3

Case Strategy with Analogy and Multi-Level Alignment, 10 agents
The experiments: stage 3

Stage 3: Analogy

-visible-entity
-visible-entity
-mover
(stand)

-approached
(approach)

-approached
(leave)

-grasped
(grasp)

-move-inside-container
(move-inside)

-move-outside-container
(move-outside)

-touched
(touch)

-touch
(touch)

-walk-to-goal
(walk-to)

-get-closer-location
(get-closer)

-get-closer-location
(go-away)

-get-destination
(get-closer)

-get-destination
(go-away)

-get-closer-actor
(go-away)

-get-closer-actor
(get-closer)

-get-closer-actor
(disappear)

-grasper
(grasp)

-apprearer
(apppear)

-faller
(fall)

-faller
(fall)

-toucher
(touch)

-toucher
(separate)

-walker
(walk-to)

-hider
(hide)

-put-undergoer
(put)

-giver
(give)

-gift
(give)

-receiver
(give)

-taker
(take)

-taken
(take)

-take-source
(take)
The experiments: stage 3

Stage 3: Analogy
The experiments: stage 3

<table>
<thead>
<tr>
<th>distance-decreasing-1</th>
<th>sem-role-99</th>
<th>‘-pui’</th>
</tr>
</thead>
<tbody>
<tr>
<td>take-1</td>
<td>sem-role-105</td>
<td>syn-role-21</td>
</tr>
<tr>
<td>touch-2</td>
<td>sem-role-83</td>
<td></td>
</tr>
<tr>
<td>grasp-1</td>
<td>sem-role-14</td>
<td></td>
</tr>
<tr>
<td>cause-move-on-1</td>
<td>sem-role-15</td>
<td></td>
</tr>
</tbody>
</table>

| take-2                 | sem-role-171|
| visible-1              | sem-role-77 |
| cause-move-on-3        | sem-role-73 |
| hide-2                 | sem-role-58 |
| object-1               | sem-role-51 |
| move-inside-1          | sem-role-40 |
| touch-1                | sem-role-8  |
| approach-1             | sem-role-8  |
| move-outside-2         | sem-role-8  |
| walk-to-1              | sem-role-8  |
| fall-2                 | sem-role-8  |
| grasp-2                | sem-role-8  |

| take-2                 | sem-role-155|
| move-outside-1         | sem-role-155|
| walk-to-2              | sem-role-33 |
| approach-2             | sem-role-33 |
| move-inside-1          | sem-role-84 |
| distance-decreasing-2  | sem-role-15 |
| fall-1                 | sem-role-15 |
| hide-1                 | sem-role-15 |
The experiments: stage 3

- **jack** -fuitap  walk-to  **jill**  -ginah
  Jack sem-role-7  walk-to  Jill  sem-role-24
  ‘Jack walks to Jill.’

- **touch**  **jill**  -fuitap  **house**  -payis
  touch  Jill  sem-role-7  house  sem-role-29
  ‘Jill touches the house.’

- **house**  -woechen  move-inside  **boy**  -fuitap
  house  sem-role-56  move-inside  boy  sem-role-7
  ‘The boy moves inside the house.’
Example IV
Explaining Real Language Change

## German definite articles

<table>
<thead>
<tr>
<th>Case</th>
<th>SG-M</th>
<th>SG-F</th>
<th>SG-N</th>
<th>PL-M</th>
<th>PL-F</th>
<th>PL-N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nom</td>
<td>dër</td>
<td>diu</td>
<td>daZ₂</td>
<td>die</td>
<td>deo</td>
<td>diu</td>
</tr>
<tr>
<td>Acc</td>
<td>dën</td>
<td>die</td>
<td>daZ₂</td>
<td>die</td>
<td>deo</td>
<td>diu</td>
</tr>
<tr>
<td>Dat</td>
<td>dëmu</td>
<td>dëru</td>
<td>dëmu</td>
<td>dēm</td>
<td>dēm</td>
<td>dēm</td>
</tr>
<tr>
<td>Gen</td>
<td>dës</td>
<td>dëra</td>
<td>dës</td>
<td>dëro</td>
<td>dëro</td>
<td>dëro</td>
</tr>
</tbody>
</table>

Old High German demonstratives (900-1100 AD; Wright 1906)
German definite articles

<table>
<thead>
<tr>
<th>Case</th>
<th>SG-M</th>
<th>SG-F</th>
<th>SG-N</th>
<th>PL-M</th>
<th>PL-F</th>
<th>PL-N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nom</td>
<td>dër</td>
<td>diu</td>
<td>daɐ̃</td>
<td>die</td>
<td>diu</td>
<td>die</td>
</tr>
<tr>
<td>Acc</td>
<td>dën</td>
<td>die</td>
<td>daɐ̃</td>
<td>die</td>
<td>diu</td>
<td>die</td>
</tr>
<tr>
<td>Dat</td>
<td>dëm</td>
<td>dër</td>
<td>dëm</td>
<td>dën</td>
<td>den</td>
<td>dën</td>
</tr>
<tr>
<td>Gen</td>
<td>dës</td>
<td>dër</td>
<td>dës</td>
<td>dër</td>
<td>dër</td>
<td>dër</td>
</tr>
</tbody>
</table>

Middle High German definite articles (1100-1500 AD; Wright 1916)
German definite articles

<table>
<thead>
<tr>
<th>Case</th>
<th>SG-M</th>
<th>SG-F</th>
<th>SG-N</th>
<th>PL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nom</td>
<td>der</td>
<td>die</td>
<td>das</td>
<td>die</td>
</tr>
<tr>
<td>Acc</td>
<td>den</td>
<td>die</td>
<td>das</td>
<td>die</td>
</tr>
<tr>
<td>Dat</td>
<td>dem</td>
<td>der</td>
<td>dem</td>
<td>den</td>
</tr>
<tr>
<td>Gen</td>
<td>des</td>
<td>der</td>
<td>des</td>
<td>der</td>
</tr>
</tbody>
</table>

New High German definite articles
# German definite articles

<table>
<thead>
<tr>
<th></th>
<th>Masc</th>
<th>Neut</th>
<th>Fem</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Singular</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOM</td>
<td>dër</td>
<td>daz</td>
<td>diu</td>
</tr>
<tr>
<td>ACC</td>
<td>dën</td>
<td>daz</td>
<td>die</td>
</tr>
<tr>
<td>DAT</td>
<td>dém</td>
<td>dém</td>
<td>dém</td>
</tr>
<tr>
<td>GEN</td>
<td>dës</td>
<td>dës</td>
<td>dëru</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Masc</th>
<th>Neut</th>
<th>Fem</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Plural</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOM</td>
<td>die</td>
<td>diu</td>
<td>deo</td>
</tr>
<tr>
<td>ACC</td>
<td>die</td>
<td>diu</td>
<td>deo</td>
</tr>
<tr>
<td>DAT</td>
<td>dën</td>
<td>dën</td>
<td>dën</td>
</tr>
<tr>
<td>GEN</td>
<td>dëro</td>
<td>dëro</td>
<td>dëro</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Masc</th>
<th>Neut</th>
<th>Fem</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Singular</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOM</td>
<td>dër</td>
<td>daz</td>
<td>diu</td>
</tr>
<tr>
<td>ACC</td>
<td>dën</td>
<td>daz</td>
<td>die</td>
</tr>
<tr>
<td>DAT</td>
<td>dém</td>
<td>dém</td>
<td>dém</td>
</tr>
<tr>
<td>GEN</td>
<td>dës</td>
<td>dës</td>
<td>dër</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Masc</th>
<th>Neut</th>
<th>Fem</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Plural</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOM</td>
<td>die</td>
<td>diu</td>
<td>die</td>
</tr>
<tr>
<td>ACC</td>
<td>die</td>
<td>diu</td>
<td>die</td>
</tr>
<tr>
<td>DAT</td>
<td>dën</td>
<td>dën</td>
<td>dën</td>
</tr>
<tr>
<td>GEN</td>
<td>dër</td>
<td>dër</td>
<td>dër</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Masc</th>
<th>Neut</th>
<th>Fem</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Singular</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOM</td>
<td>der</td>
<td>das</td>
<td>die</td>
</tr>
<tr>
<td>ACC</td>
<td>den</td>
<td>das</td>
<td>der</td>
</tr>
<tr>
<td>DAT</td>
<td>dem</td>
<td>dem</td>
<td>der</td>
</tr>
<tr>
<td>GEN</td>
<td>des</td>
<td>des</td>
<td>der</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Masc</th>
<th>Neut</th>
<th>Fem</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Plural</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOM</td>
<td>die</td>
<td>die</td>
<td>die</td>
</tr>
<tr>
<td>ACC</td>
<td>die</td>
<td>die</td>
<td>die</td>
</tr>
<tr>
<td>DAT</td>
<td>den</td>
<td>den</td>
<td>den</td>
</tr>
<tr>
<td>GEN</td>
<td>der</td>
<td>der</td>
<td>der</td>
</tr>
</tbody>
</table>

**Languages:**
- Old High German
- Middle High German
- New High German
Why?
What do experts say?

- Formal perspective: Non-systematic (accidental) syncretism
- Historical perspective: historical accident

“Accidental syncretism occurs as a by-product of phonological or morphological changes. For example, the collapse of case forms due to phonological reduction or loss is a familiar theme in the history of Indo-European languages, as in Middle High German, where the reduction of unstressed vowels in final syllables to schwa (graphically e) in Middle High German led to widespread mergers throughout the inflectional system” (Baerman 2009)
Problems for the accidental hypothesis

• The “accidental” paradigm has survived for several centuries despite huge structural variation & language contact

• “Explanations” only say what has happened, but not why

(Shrier 1965)
Case loss in Germanic

- Phonological erosion
- Change from synthetic to analytic
- Change from free to fixed word order
- Change from lexical to structural case
- Development of the definite article
- Merger of (partially synonymous) argument structure constructions & attraction by high type frequency constructions

(Barðdal 2009)
Evolutionary Linguistics

- Linguistic selectionism
- Processes that cause variation
- Selection of variants because of advantages for linguistic communication
Can we demonstrate the validity of our hypothesis?

“Old High German field work is no option.”

True... But we can “bring back alive” Old High German!

(Fleischer 2012)
Research Plan

• Step 1: Bring back “alive” a language through a processing model

Implementation details of the grammar:
R. van Trijp (subm). Handling Structural Variation in Feature-Structure Grammars.

>> Online demo of the approach: www.fcg-net.org
>> Demo of OHG online as soon as paper is accepted
>> All language technologies are open-source
Research Plan

- **Step 1:** Bring back “alive” a language through a processing model

- **Step 2:** Linguistic Assessment Criteria
  (this presentation)
Populate speech communities of artificial agents
Provide agents with “reconstructed” grammar

Old High German *(Wright 1906)*

New High German

Texas German *(Boas 2009)*
Let agents play language games

Der Junge geht zu dem Mädchen.

Ja.
Let agents play language games

processing effort
articulatory effort
usage of memory
social conformity
communicative success

....

processing effort
cognitive effort
auditory precision
usage of memory
social conformity
communicative success

....
Comparative Reconstruction Experiment

• Three utterance types, 360 subtypes
  • 216 subtypes for ditransitive
  • 72 subtypes for Nom-Acc and 72 for Nom-Dat

• Symbolic, deep language processing model in Fluid Construction Grammar (www.fcg-net.org)
Results & Discussion
Cue Reliability & Disambiguation Power

• The **man** crossed the street.
• The fish **were** biting well that day.
• The antelope ran away when John tried to approach **them**.
• The antelope ran away.
• ...

A Language consists of a set of a set of language systems \( \{L_i, L_{i+1}, \ldots, L_n\} \)

- **Cue reliability** = \( \frac{UD|L_i}{U} \)
- **Disambiguation power** = \( \frac{UD|L_i, L_{i+1}, \ldots, L_n}{U} \)

\( U = \text{total number of utterances} \)
\( UD = \text{number of disambiguated utterances} \)
**Utterance Disambiguation Given Language Systems**

<table>
<thead>
<tr>
<th>System</th>
<th>Number of Utterances</th>
</tr>
</thead>
<tbody>
<tr>
<td>( U_D</td>
<td>L_1 )</td>
</tr>
<tr>
<td>( U_D</td>
<td>L_1, L_2 )</td>
</tr>
<tr>
<td>( U_D</td>
<td>L_1, L_3 )</td>
</tr>
<tr>
<td>( U_D</td>
<td>L_1, L_2, L_3 )</td>
</tr>
</tbody>
</table>

\( L_1 = \) determiners + nouns / \( L_2 = \) subject-verb agreement / \( L_3 = \) selection restrictions

Old High German □ New High German
Problematic utterance types

• 9 out of 10 remaining ambiguities in OHG involve plural nominative vs. accusative distinctions (die vs. diu vs. deo)

• This three-way distinction has disappeared from the language
Problematic utterance types

- **OHG:**
  Die Man fundun deo Friuntinnā
  the.NOM/ACC.PL the.NOM/ACC.PL
  ‘The men found the (female) friends.’
  ‘The (female) friends found the men.’

- **NHG:**
  Die Männer fanden die Freundinnen.
Processing Efficiency

- First indicator of processing efficiency: search tree length
- Counterintuitively, NHG performs equally well as OHG
- Increased syncretism does not lead to increased search (given the right representation)
Processing Efficiency

- Second indicator: how many primitive operations are needed for processing case-number-gender specifications?

\[ p = 2.2 \times 10^{-16} < 0.01 \]
Processing Efficiency

• The NHG-system is twice the size of the OHG-system, so it can be processed faster

\[
PC(FM) = 2 \times (|FM| + \sum_{i=1}^{\frac{|FM|}{|FV_i|}})
\]

• System size has reduced without harming disambiguation power
Ease of Articulation

- Speech = balance between ease of articulation and intelligibility (ease of perception)

<table>
<thead>
<tr>
<th>Phonemes</th>
<th>die</th>
<th></th>
<th>das</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>d</td>
<td>i</td>
<td>d</td>
<td>a</td>
</tr>
<tr>
<td>Syllabic</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Continuant</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Sonorant</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Nasal</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Voice</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Anterior</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Coronal</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Lateral</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>High</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Low</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Back</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Rounded</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Long</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Ease of Articulation

- Articulatory effort = cost of moving from one articulator to another (~ edit distance)

<table>
<thead>
<tr>
<th>Phonemes</th>
<th>die</th>
<th>das</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>d</td>
<td>d</td>
</tr>
<tr>
<td>Syllabic</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Continuant</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Sonorant</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Nasal</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Voice</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Anterior</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Coronal</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Lateral</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>High</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>Low</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Back</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Rounded</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Long</td>
<td>+</td>
<td>−</td>
</tr>
</tbody>
</table>
Ease of Articulation

- Articulatory effort = cost of moving from one articulator to another (~ edit distance)

### Old High German

<table>
<thead>
<tr>
<th></th>
<th>Singular</th>
<th></th>
<th></th>
<th>Plural</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NOM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>der</td>
<td>24 0.6757</td>
<td>daz 0.6486</td>
<td>20 0.7297</td>
<td>die 0.7568</td>
<td>20 0.7297</td>
<td>18 0.7568</td>
</tr>
<tr>
<td>ACC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dümu</td>
<td>24 0.6757</td>
<td>dêmu 0.4595</td>
<td>20 0.7568</td>
<td>dênu 0.6757</td>
<td>24 0.6757</td>
<td>24 0.6757</td>
</tr>
<tr>
<td>DAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### New High German

<table>
<thead>
<tr>
<th></th>
<th>Singular</th>
<th></th>
<th></th>
<th>Plural</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NOM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>der</td>
<td>24 0.6757</td>
<td>das 0.6486</td>
<td>14 0.8108</td>
<td>die 0.8108</td>
<td>14 0.8108</td>
<td>14 0.8108</td>
</tr>
<tr>
<td>ACC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dem</td>
<td>24 0.6757</td>
<td>dem 0.6757</td>
<td>14 0.8108</td>
<td>den 0.6757</td>
<td>24 0.6757</td>
<td>24 0.6757</td>
</tr>
<tr>
<td>DAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Ease of Articulation

• “Expensive” articles in OHG: dëmu and dëru have been shortened in NHG (dem and der)

• Diphthongs (die, diu, deo) have been simplified to less costly long vowel (die)
Auditory Efficiency

- How easily can articles be distinguished from each other in perception?
- Similarly measured as an edit distance
- Distance between closest neighbors

```
  d  i:  -
  d a s
```
Auditory Efficiency

• Results: In NHG, the distance between articles that are hard to distinguish from each other has increased...

• But only if disambiguation power remained unharmed!
Auditory Efficiency

Distance from OHG "die"
Auditory Efficiency

Distance from OHG "die"

Distance from NHG "die"
Auditory Efficiency

Distance from OHG "die"

Distance from NHG "die"

Distance from OHG "dēru"
Auditory Efficiency

Distance from OHG "die"

Distance from NHG "die"

Distance from OHG "dēru"

Distance from NHG "der"
Summary

- **Old High German**
  - Cue Reliability
  - Disambiguation Power
  - Processing Efficiency
  - Ease of Articulation
  - Auditory Distinctiveness

- **New High German**
  - Cue Reliability
  - Disambiguation Power
  - Processing Efficiency
  - Ease of Articulation
  - Auditory Distinctiveness
That Awful German Language

“Surely there is not another language that is so slipshod and systemless, and so slippery and elusive to the grasp.”

Mark Twain, 1835-1910
Not that awful after all!

- Evolution of German definite articles was no accident...
- ... it was motivated by needs and constraints of language users.
- But what happened precisely? And why?
Formulating a more precise hypothesis

• **Variation** is inevitable in a speech population

• The German “linguistic ecosystem” changed
  > Most syncretisms emerged by the Middle Ages when the demonstrative had grammaticalized into an article...
  > ... so an additional cue became available in the NP’s head noun

• Syncretic forms were **selected** because they had become more advantageous for language usage (in the new ecosystem)
Background reading


• http://ai.vub.ac.be

• http://www.emergent-languages.org

• http://www.fcg-net.org
Thank you!

Questions? Comments?
remi@csl.sony.fr