Historical Linguistics and AI

Dirk Pijpops, QLVL research unit, University of Leuven
Overview

- Historical Linguistics
- Agent-based modelling
- Example: own research
Historical Linguistics

- Subfield of linguistics which asks how and why languages change
- Languages change
Historical Linguistics

- Cycles
- Drifts
Historical Linguistics

- **Cycles: grammaticalization of the French future tense**

<table>
<thead>
<tr>
<th>Latin-Old French</th>
<th>Middle French</th>
<th>Modern French</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ego cantabo</em></td>
<td><em>je chanterai</em></td>
<td><em>je chanterai</em></td>
</tr>
<tr>
<td>I sing-FUT</td>
<td>I sing-FUT</td>
<td>I sing-FUT</td>
</tr>
<tr>
<td><em>je chanter ai</em></td>
<td>→</td>
<td>→</td>
</tr>
<tr>
<td>I to sing have</td>
<td><em>je chanterai</em></td>
<td><em>je chanterai</em></td>
</tr>
<tr>
<td><em>je vais chanter</em></td>
<td>→</td>
<td>→</td>
</tr>
<tr>
<td>I go to sing</td>
<td>*je vais chanter</td>
<td>I sing-FUT</td>
</tr>
</tbody>
</table>
Drifts: deflexion, e.g. loss of case in the Germanic and Romance languages

\[ \exists x, y: \text{ girls}(y) \& \text{ women}(x) \& \text{ caressing}(x, y) \]

\[ \text{ puell-ae } \quad \text{ femin-ae } \quad \text{ permulcent} \]

De vrouwen aaien de meisjes.
The women caress the girls
Dat zijn de meisjes die de vrouwen aaien.

...de meisjes die zij aaien

...de meisjes die hun aaien

<table>
<thead>
<tr>
<th>subject</th>
<th>object</th>
</tr>
</thead>
<tbody>
<tr>
<td>ik</td>
<td>mij</td>
</tr>
<tr>
<td>jij</td>
<td>jou</td>
</tr>
<tr>
<td>hij/zij</td>
<td>hem/haar</td>
</tr>
<tr>
<td>wij</td>
<td>ons</td>
</tr>
<tr>
<td>jullie</td>
<td>jullie</td>
</tr>
<tr>
<td>zij</td>
<td>hun</td>
</tr>
</tbody>
</table>
Historical Linguistics

- **Nou, hun zeggen dat...**
  - Well, **them** say that...

- **Dan zegt hem weer...**
  - Then says **him** again

- **Dirks fiets → Dirk zijn fiets**
  - Dirk’s bike → Dirk his bike

- **iets leuk → iets leuk**
  - something fun → something fun
Historical Linguistics

- Why is historical linguistics of interest to researchers in AI?
  - Interest in language: Turing test, robot communication, learning, the emergence of language
  - How do (rules in) languages come into being?
    - Grammaticalization (e.g. French future tense)
    - Exaptation (e.g. Dutch adjectival inflection, Dutch gender)
    - Reanalysis (e.g. Dutch z’n-construction)
    - ...
Historical Linguistics

- Why is Artificial Intelligence of interest to historical linguists?
Historical Linguistics

Corpus research

Data

Mechanisms of language change
Historical Linguistics

<table>
<thead>
<tr>
<th>Sanskrit</th>
<th>Zend</th>
<th>Gr. Dor.</th>
<th>Latin</th>
<th>Gothic</th>
<th>Lithuanian</th>
<th>Old Slavonic</th>
</tr>
</thead>
<tbody>
<tr>
<td>prathama</td>
<td>frathema</td>
<td>πρώτα</td>
<td>prima</td>
<td>fruma</td>
<td>pirmà</td>
<td>perva-ya</td>
</tr>
<tr>
<td>dwitiya</td>
<td>bitya</td>
<td>δευτέρα</td>
<td>altera</td>
<td>anthara</td>
<td>antrà</td>
<td>vitora-ya</td>
</tr>
<tr>
<td>tritiya</td>
<td>thritya</td>
<td>τρίτη</td>
<td>tertia</td>
<td>thridyō</td>
<td>tréchià</td>
<td>treti-ya</td>
</tr>
<tr>
<td>chaturthā</td>
<td>tiūrya</td>
<td>τέταρτη</td>
<td>quarta</td>
<td>(sidvórdō)</td>
<td>ketwirtà</td>
<td>chetverts-ya</td>
</tr>
<tr>
<td>panchamā</td>
<td>pugdha</td>
<td>πεντάπτη</td>
<td>quinta</td>
<td>fimftō</td>
<td>penktà</td>
<td>pyata-ya</td>
</tr>
<tr>
<td>shashtha</td>
<td>oṣtvā</td>
<td>ἕκτα</td>
<td>sexta</td>
<td>saihstō</td>
<td>széssta</td>
<td>shesta-ya</td>
</tr>
<tr>
<td>saptamā</td>
<td>haptatha</td>
<td>ἕβδομα</td>
<td>septima</td>
<td>(sibundō)</td>
<td>sékma</td>
<td>sedma-ya</td>
</tr>
<tr>
<td>ushtamā</td>
<td>astēma</td>
<td>ὀχτώ</td>
<td>octava</td>
<td>ahtudo</td>
<td>ászma</td>
<td>osma-ya</td>
</tr>
<tr>
<td>navamā</td>
<td>nāuma</td>
<td>ἐννάτα</td>
<td>nona</td>
<td>niundō</td>
<td>devintā</td>
<td>devyata-ya</td>
</tr>
<tr>
<td>dasamā</td>
<td>dasēma</td>
<td>δεκάτα</td>
<td>decima</td>
<td>taihundō</td>
<td>deszintā</td>
<td>desyata-ya</td>
</tr>
<tr>
<td>ekādatā</td>
<td>ekaandaśa</td>
<td>é不肯</td>
<td>undeōima</td>
<td>(ainlīstō)</td>
<td>wienolikta</td>
<td>yedina-ya-na-desy</td>
</tr>
<tr>
<td>viṁsatī tamā</td>
<td>viṁsatīma</td>
<td>二十</td>
<td>vícesima</td>
<td>二十</td>
<td>dwideszintā</td>
<td>vtoraya-ya-na-desyaty</td>
</tr>
</tbody>
</table>

(Bopp 1885: 452)
Aggregate grammaticalisation score (jittered)

Datapoints, chronological

Summative grammaticalisation score

lowess

Aggregate grammaticalisation, with lowess regression line
(Correlation: Kendall tau = 0.126, p < 0.0001)

(Petré & Van de Velde 2014)
Historical Linguistics

Corpus research

Data

Mechanisms of language change

?
Historical Linguistics

- Iterative learning experiments
  - No concrete language changes
  - Limited in scope

- Agent-based modelling / multi-agent systems
Agent-based Modelling

(Guerreiro et al. 2013)  (Dhamdher & Dovrolis 2009)  (Bazghandi 2012)
Agent-based Modelling

(Steels & Spranger 2008)
Agent-based Modelling
Historical Linguistics

Corpus research

Data

Mechanisms of language change

Agent-based models
Agent-based Modelling

⇒ Can it happen like this?
Historical Linguistics

(van Trijp 2014: 3)
Historical Linguistics

- Explain the collapse of the Germanic, Romance,... case systems
  - Historical accident (Baerman 2009)
  - Universal case hierarchy (Hawkins 2004)
  - Language use (van Trijp 2012, 2013)
Historical Linguistics

(van Trijp 2014: 3)
Break
The Rise of the Weak Inflection in Germanic

Dirk Pijpops, QLVL research unit, University of Leuven
Katrien Beuls, Artificial Intelligence Lab, Vrije Universiteit Brussel
The Rise of the Weak Inflection in Germanic

- The Story
- The Model
- The Results
- The Conclusions

(Pijpops & Beuls 2015)
The Story
The Story
The Story
The Story

- Germanic past tense

  - Strong: \( ik \) loop → \( ik \) liep
    I run → I run-PAST

  - Weak: → \( ik \) loopte
    I run-PAST
The Story

- Competition between the strong and weak strategies

<table>
<thead>
<tr>
<th>Strong</th>
<th>Weak</th>
</tr>
</thead>
<tbody>
<tr>
<td>krijg</td>
<td>kreeg</td>
</tr>
<tr>
<td>lieg</td>
<td>loog</td>
</tr>
<tr>
<td>zuig</td>
<td>zooog</td>
</tr>
<tr>
<td>drink</td>
<td>dronk</td>
</tr>
<tr>
<td>zwem</td>
<td>zwom</td>
</tr>
<tr>
<td>sterf</td>
<td>stierf</td>
</tr>
<tr>
<td>spreek</td>
<td>sprak</td>
</tr>
<tr>
<td>zit</td>
<td>zat</td>
</tr>
<tr>
<td>vaar</td>
<td>voer</td>
</tr>
<tr>
<td>blaas</td>
<td>blies</td>
</tr>
<tr>
<td>vang</td>
<td>ving</td>
</tr>
</tbody>
</table>

lach → lachte
The Story

- Germanic past tense
  - Strong: \( ik\ loop \rightarrow ik \ liep \)
    - I run \rightarrow I run-PAST
  - Weak: \( ik\ lopen\ deed \rightarrow ik\ loopte \)
    - I to run did \rightarrow I run-PAST
The Story

(Carroll et al. 2012: 161)
The Story

- Competition between the strong and weak inflections
  - Weak inflection is becoming dominant
  - Weak inflection first takes over the low frequency verbs and then works its way up to the more frequent verbs

⇒ Why?

⇒ Influence of learners
The Story

- May work for the current situation (in English):
  - Strong vowel alternations are (mostly) irregular
  - Weak inflection is more frequent

- Doesn’t work for the situation in Germanic:
  - Strong vowel alternations are still regular
  - Weak inflection has only just been born

⇒ General applicability
The Story

- Adding a new inflection only further complicates matters

- krijg → kreeg
- lieg → loog
- zuig → zoog
- drink → dronk
- zwem → zwom
- sterf → stierf
- spreek → sprak
- zit → zat
- vaar → voer
- blaas → blies
- vang → ving
The Story

- Adding a new inflection only further complicates matters

- krijg → kreeg
- lieg → loog
- zuig → zoog
- drink → dronk
- zwem → zwom
- sterf → stierf
- spreek → sprak
- zit → zat
- vaar → voer
- blaas → blies
- vang → ving
- lach → lachte
The Model

- Agent-based
- 10 agents
- Past events: 257 verbs
- FCG grammar: 11 strong patterns + 1 weak one
- Memory of previously heard forms: Corpus of Spoken Dutch
The Model

- Competition formula:

  \( Ik \text{ schrijf } + PAST \)
  ‘I wrote’

  \( schreef \)

  \( schrijfdde \)
The Model

- Competition formula:

\[ Ik \text{ schrijf} + PAST \]
‘I wrote’

\[ p(“schreef”) \sim \text{heard(“schreef”) + heard(pattern 1)} \]

\[ p(“schrijfde”) \sim \text{heard(“schrijfde”) + heard(pattern weak)} \]
The Model

- Competition formula:

\[ Ik\ schrijf + PAST \]

‘I wrote’

\[ p(“schreef”) \sim \text{heard(“schreef”)} \]

\[ p(“schrijfde”) \sim \text{heard(“schrijfde”)} \]
The Model

- Replenishment
The Model

- Replenishment
The Model

- Learning

old agent  learner agent

VANGEN

???
Learning

"ik ving"

old agent → learner agent

a → i

"ving" + 1
Results

- Original corpus input, no replenishment
The Results

- Original corpus input, new agent every 2500 interactions
The Results

- Weak starting from vastly inferior position, new agent every 2500 interactions

![Graph showing token frequency and number of games played]
The Results

- Weak starting from vastly inferior position, new agent every 2500 interactions
The Results

- Weak starting from vastly inferior position, new agent every 2500 interactions
The Results

- Weak starting from vastly inferior position, new agent every 2500 interactions
The Results

- Given a high enough replenishment rate, ...

- The weak strategy can grow to become dominant, even starting from a vastly inferior position.

- The weak strategy first takes over the low frequency verbs, then the more frequent verbs.
The Results

- Rise of the weak inflection as a byproduct of language use
- Learners do not actively try to change the language, they just try to express something
The Conclusions

- Evolutionary advantages of both inflections
  - Strong inflection: shorter e.g. *I ran* ↔ *I runned*
  - Weak inflection: generally applicable
The Conclusions

- Conditions which favor the weak inflection
  - Sociohistorical conditions: many language learners
  - Linguistic conditions: low frequent verbs

  ⇒ Both inflections can co-exist for a long time

  ⇒ Expansion of weak inflection can be slowed down
The Story

- No short term effect

ij \rightarrow ee
ie \rightarrow oo
ee \rightarrow a
...  
* \rightarrow *-de

VANGEN

“vangde”
The Story

- Criticism

ij → ee
ie → oo
ee → a
a → i
...
* → *-de

VANGEN

“ving”
The Conclusions

- No short term effect
- Given the right conditions, huge long term effect
General conclusions

- In historical linguistics, sociohistorical changes, like increased language contact, influx of new learners etc.: often named as causes or catalysts of languages changes

- Agent-based models can be used to test their long-term effects
Techniques from Artificial Intelligence, such as agent-based modelling, can be useful in the most unexpected disciplines. Look outside of your field.
Thanks

- for further information: dirk.pijpops@kuleuven.be
References