Historical Linguistics and AI

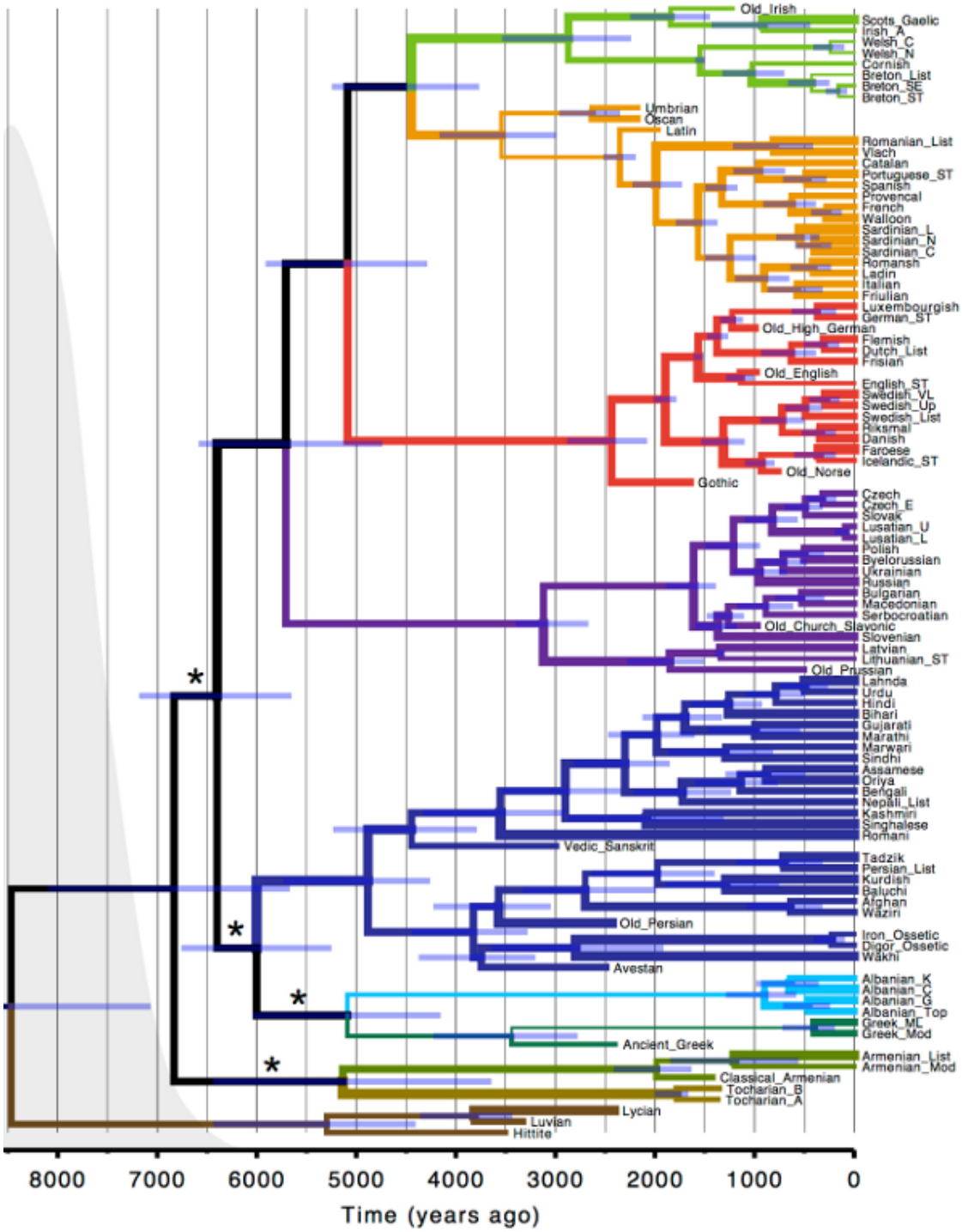
Dirk Pijpops, QLVL research unit, University of Leuven

Overview

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

Historical Linguistics

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


Old_Irish
Scots_Gaelic
Irish_A
Welsh_C
Welsh_N
Cornish
Breton_List
Breton_SE
Breton_ST
Umbrian
Oscan
Latin
Romanian_List
Vlach
Catalan
Portuguese_ST
Spanish
Provencal
French
Walloon
Sardinian_L
Sardinian_N
Sardinian_C
Romansh
Ladin
Italian
Friulian
Luxembourgish
German_ST
Old_High_German
Flemish
Dutch_List
Frisian
Old_English
English_ST
Swedish_VL
Swedish_Up
Swedish_List
Riksmal
Danish
Faroese
Icelandic_ST
Old_Norse
Gothic
Czech
Czech_E
Slovak
Lusatian_U
Lusatian_L
Polish
Byelorussian
Ukrainian
Russian
Bulgarian
Macedonian
Serbocroatian
Old_Church_Slavonic
Slovenian
Latvian
Lithuanian_ST
Old_Prussian
Lahnda
Urdu
Hindi
Bihari
Gujarati
Marathi
Marwari
Sindhi
Assamese
Oriya
Bengali
Nepali_List
Kashmiri
Singhalese
Romani
Vedic_Sanskrit
Tadzik
Persian_List
Kurdish
Baluchi
Afghan
Waziri
Old_Persian
Iron_Ossetic
Digor_Ossetic
Wakhi
Avestan
Albanian_K
Albanian_C
Albanian_G
Albanian_Top
Greek_ML
Greek_Mod
Ancient_Greek
Armenian_List
Armenian_Mod
Classical_Armenian
Tocharian_B
Tocharian_A
Lycian
Luvian
Hittite
8000
7000
6000
5000
4000
3000
2000
1000
0
Time (years ago)

Historical Linguistics

- Cycles
- Drifts

Historical Linguistics

- Cycles: grammaticalization of the French future tense

Latin-Old French		Middle French		Modern French
ego cantabo I sing-FUT				
je chanter ai I to sing have	→	*je chanterai* I sing-FUT	→	*je chanterai* I sing-FUT
		je vais chanter I go to sing	→	*je vais chanter* I sing-FUT

Historical Linguistics

- Drifts: deflexion, e.g. loss of case in the Germanic and Romance languages

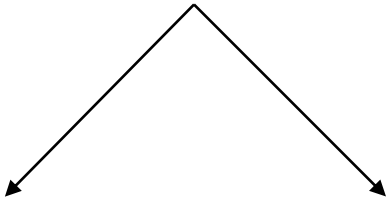
$$\exists\, x, y\colon girls(y)\ \&\ women(x)\ \&\ caressing(x, y)$$

puell-as *femin-ae* *permulcent*

De	*vrouwen*	*aaien*	*de meisjes.*
The	women	caress	the girls

Historical Linguistics

Dat zijn de meisjes die de vrouwen aaien.



...de meisjes die zij aaien

...de meisjes die hun aaien

subject	object
ik	mij
jij	jou
hij/zij	hem/haar
wij	ons
jullie	jullie
zij	hun

Historical Linguistics

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

- Dan zegt hem weer…
  Then says him again

subject	object
ik	mij
je	
hem/ze	
we	ons
jullie	
hun	

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

- *iets leuks* → *iets leuk*
  something fun → something fun

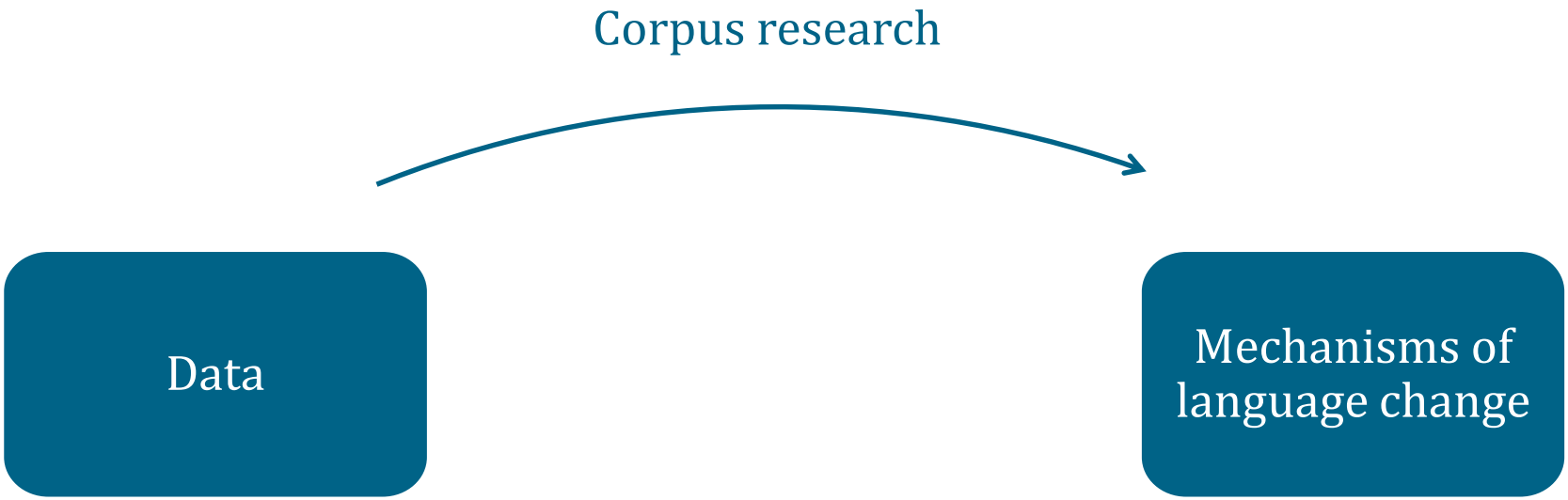
Historical Linguistics

- 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)
    - …

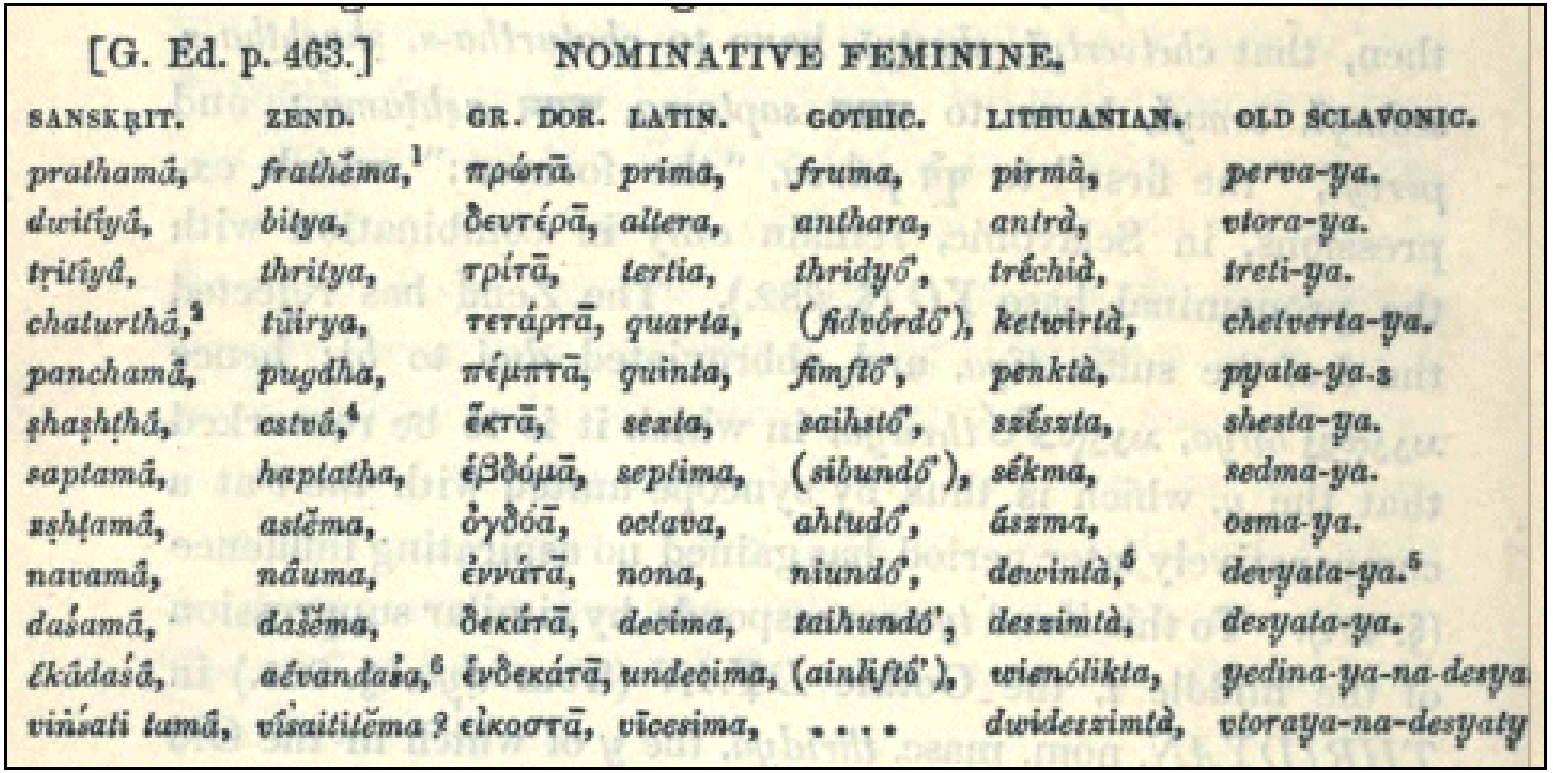
Historical Linguistics

- Why is Artificial Intelligence of interest to historical linguists?

Historical Linguistics



Historical Linguistics

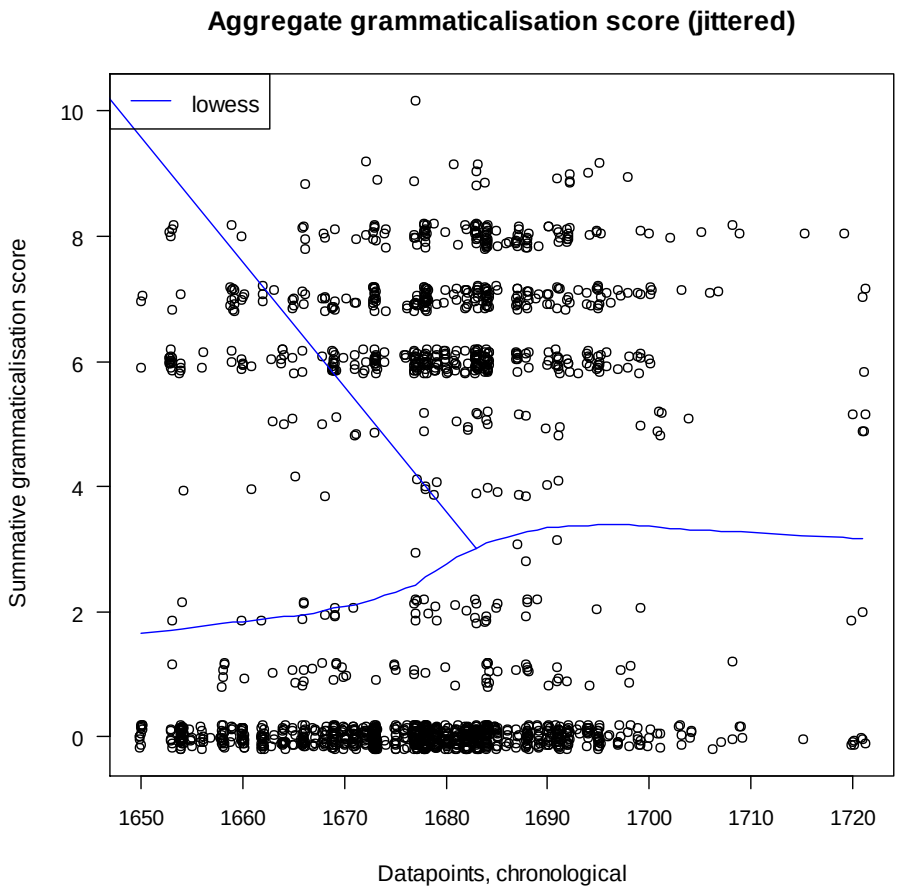


[G. Ed. p. 463.] NOMINATIVE FEMININE.

SANSKṚIT.	ZEND.	GR. DOR.	LATIN.	GOTHIC.	LITHUANIAN.	OLD SCLAVONIC.
prathamâ,	*fratĕma,*[1]	πρώτᾱ.	*prima,*	*fruma,*	*pirmà,*	*perva-ya.*
dwitîyâ,	*bitya,*	δευτέρᾱ,	*altera,*	*anthara,*	*antrà,*	*vtora-ya.*
tṛitîyâ,	*thritya,*	τρίτᾱ,	*tertia,*	*thridyô',*	*trĕchià,*	*treti-ya.*
chaturthâ,[2]	*tûirya,*	τεταρτᾱ,	*quarta,*	(*fidvôrdô'*),	*ketwirtà,*	*chetverta-ya.*
panchamâ,	*pugdha,*	πέμπτᾱ,	*quinta,*	*fimftô',*	*penktà,*	*pyata-ya.*[3]
ṣhaṣhṭhâ,	*estvâ,*[4]	ἕκτᾱ,	*sexta,*	*saihstô',*	*szĕszta,*	*shesta-ya.*
saptamâ,	*haptatha,*	ἑβδόμᾱ,	*septima,*	(*sibundô'*),	*sĕkma,*	*sedma-ya.*
aṣhṭamâ,	*astĕma,*	ὀγδόᾱ,	*octava,*	*ahtudô',*	*ãszma,*	*osma-ya.*
navamâ,	*nâuma,*	ἐννάτᾱ,	*nona,*	*niundô',*	*dewintà,*[5]	*devyata-ya.*[5]
daśamâ,	*daśĕma,*	δεκάτᾱ,	*decima,*	*taihundô',*	*deszimtà,*	*desyata-ya.*
ékâdaśâ,	*aêvandaśa,*[6]	ἑνδεκάτᾱ,	*undecima,*	(*ainlifta'*),	*wienólikta,*	*yedina-ya-na-desya.*
viṁśati tamâ,	*vîsaitiĭema ?*	εἰκοστᾱ,	*vīcesima,*	. . . .	*dwideszimtà,*	*vtoraya-na-desyaty*

(Bopp 1885: 452)

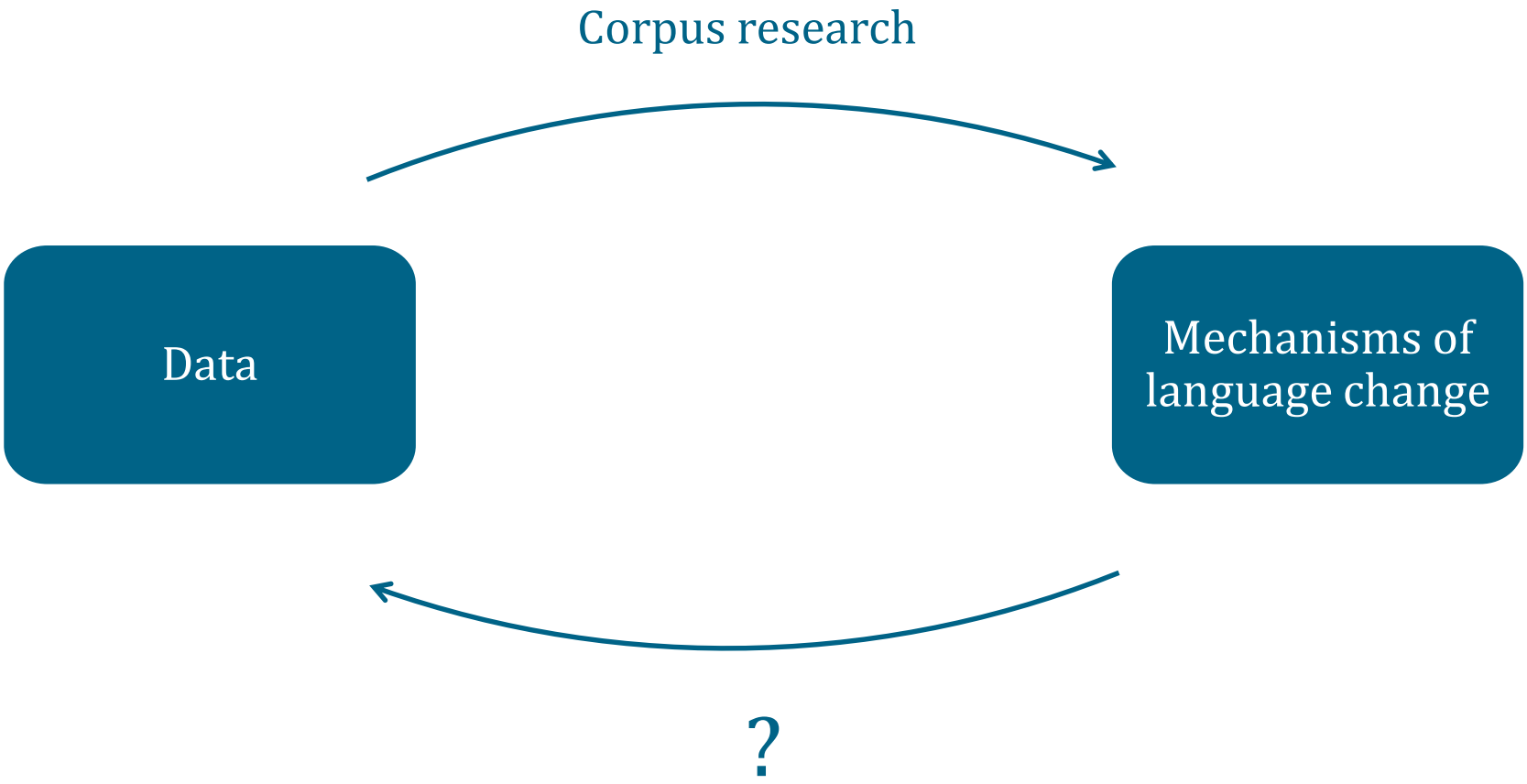
Historical Linguistics



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

(Petré & Van de Velde 2014)

Historical Linguistics



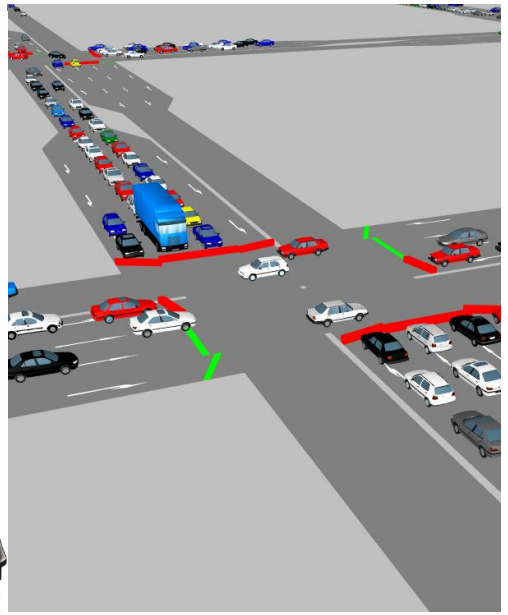
Historical Linguistics

- Iterative learning experiments
  - No concrete language changes
  - Limited in scope
- Agent-based modelling / multi-agent systems

Agent-based Modelling

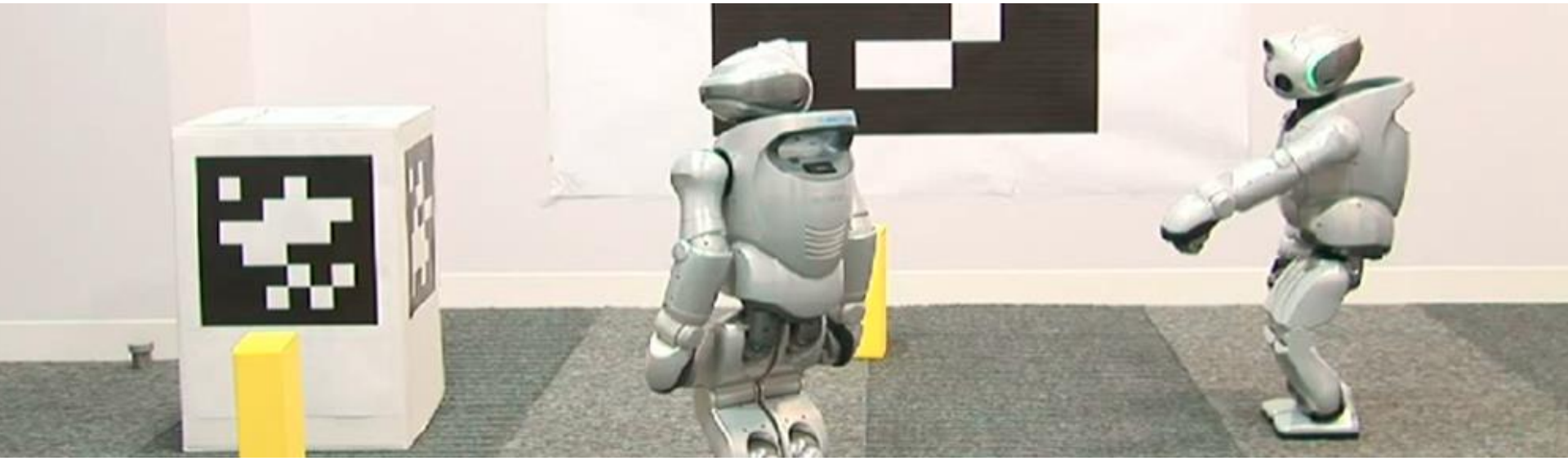
(Guerreiro et al. 2013)

(Dhamdher & Dovrolis 2009)



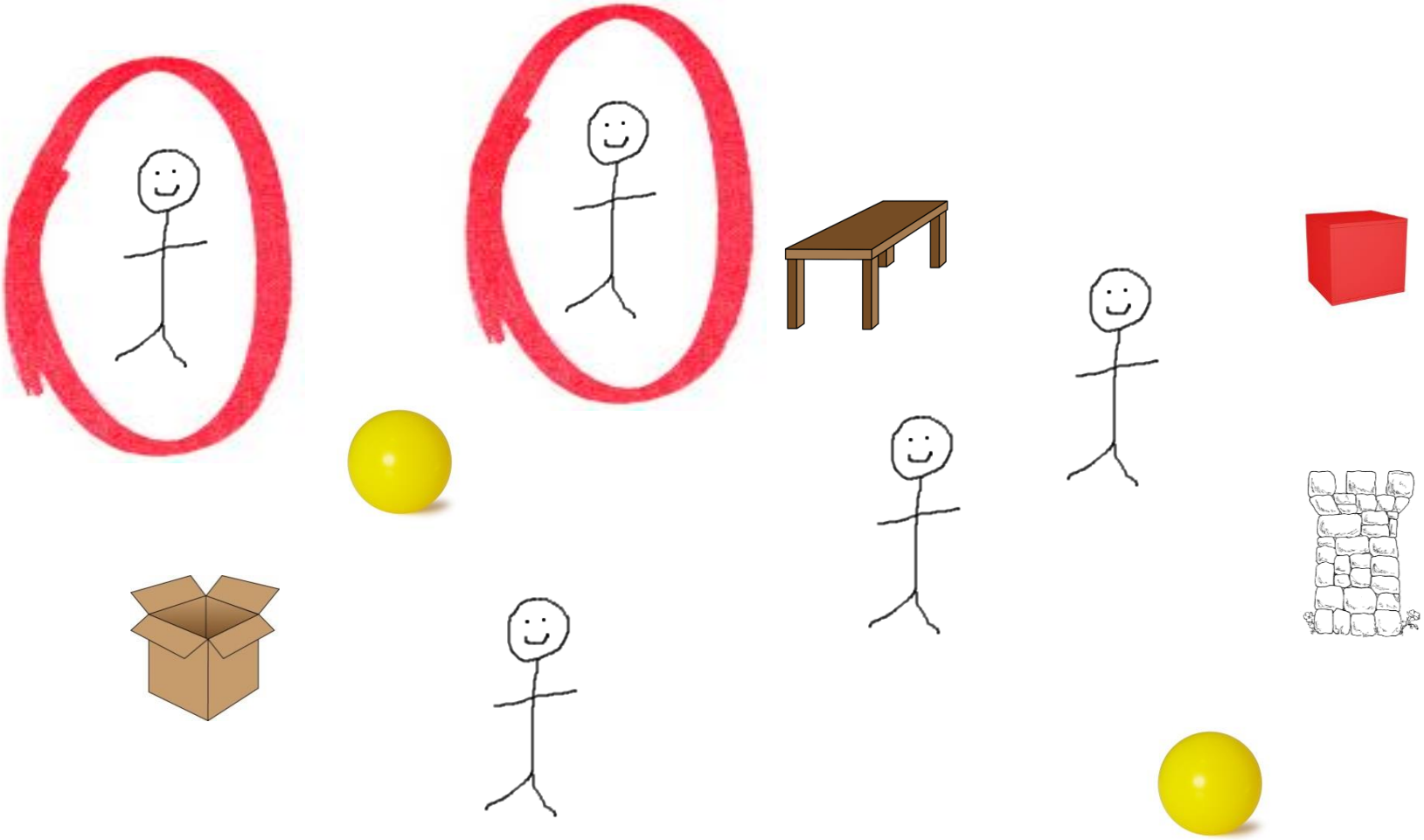
(Bazghandi 2012)

Agent-based Modelling

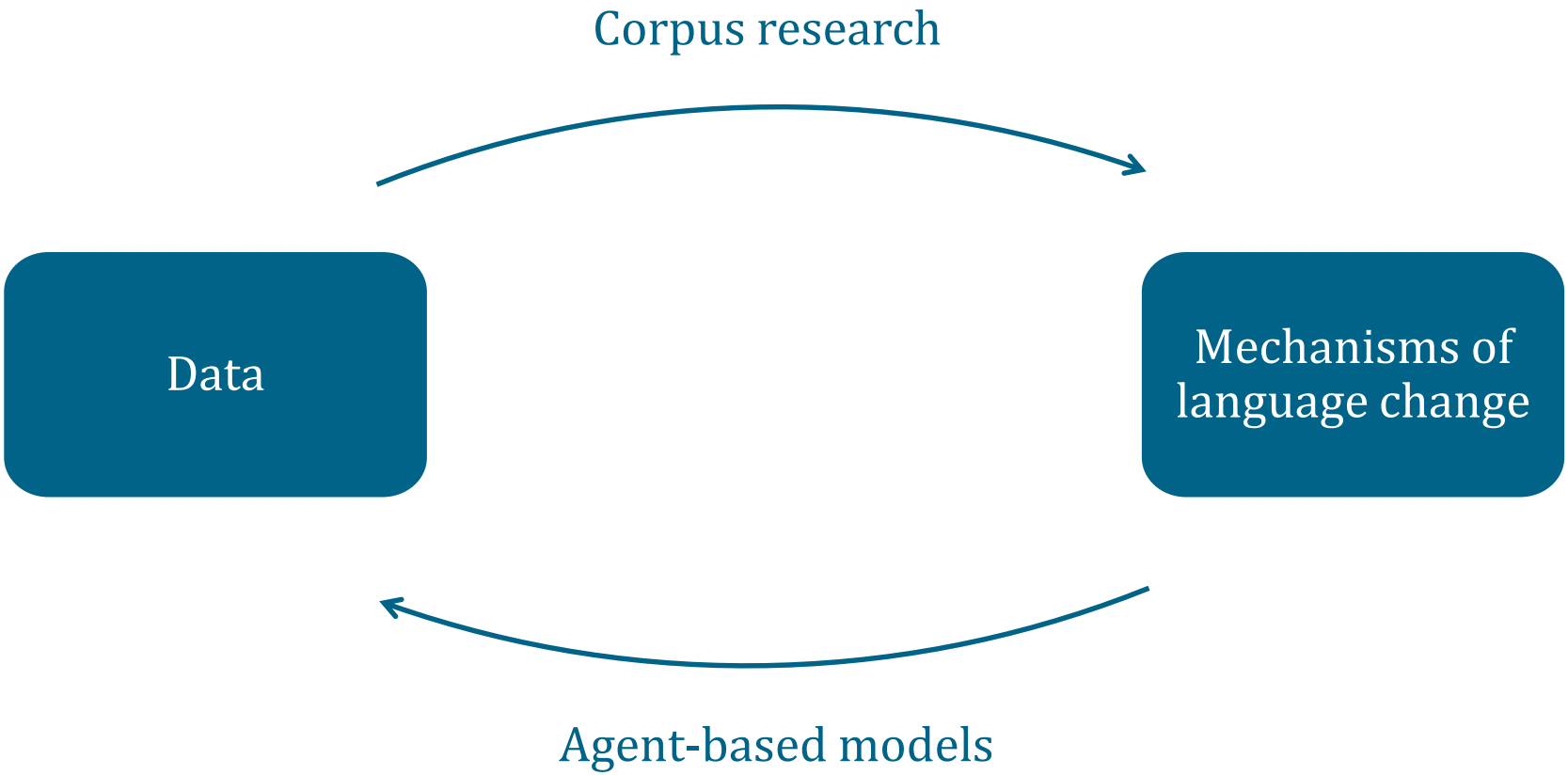


(Steels & Spranger 2008)

Agent-based Modelling



Historical Linguistics



Agent-based Modelling

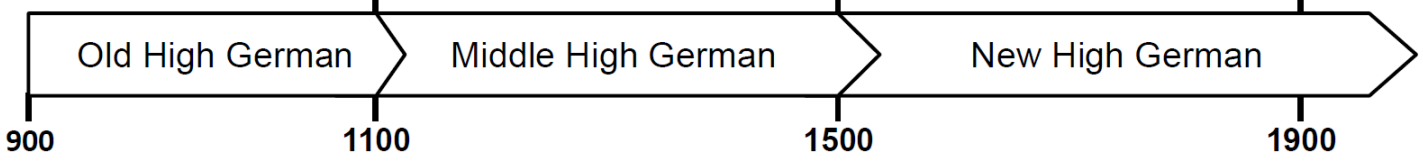
$\Rightarrow$ Can it happen like this?

Historical Linguistics

	Masc	Neut	Fem
	Singular		
NOM	dër	daz	diu
ACC	dën	daz	die
DAT	dëmu	dëmu	dëru
GEN	dës	dës	dëru
	Plural		
NOM	die	diu	deo
ACC	die	diu	deo
DAT	dēn	dēn	dēn
GEN	dëro	dëro	dëro

Masc	Neut	Fem
Singular		
dër	daz	diu
dën	daz	die
dëm	dëm	dër
dës	dës	dër
Plural		
die	diu	die
die	diu	die
dën	dën	dën
dër	dër	dër

Masc	Neut	Fem
Singular		
der	das	die
den	das	die
dem	dem	der
des	des	der
Plural		
die	die	die
die	die	die
den	den	den
der	der	der



(van Trijp 2014: 3)

Historical Linguistics

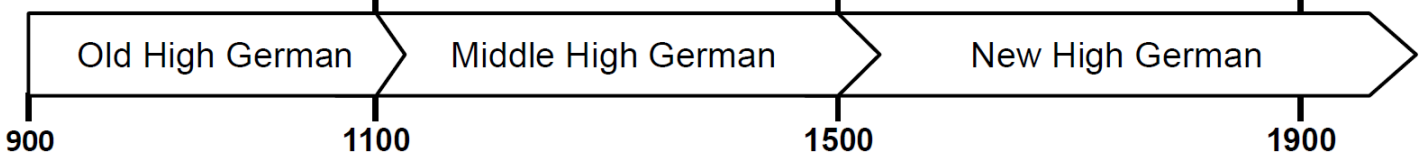
- Explain the collapse of the Germanic, Romance,… case systems
  - Historical accident (Baerman 2009)
  - Universal case hierarchy (Hawkins 2004)
  - Language use (van Trijp 2012, 2013)

Historical Linguistics

	Masc	Neut	Fem
	Singular		
NOM	dër	daz	diu
ACC	dën	daz	die
DAT	dëmu	dëmu	dëru
GEN	dës	dës	dëru
	Plural		
NOM	die	diu	deo
ACC	die	diu	deo
DAT	dēn	dēn	dēn
GEN	dëro	dëro	dëro

Masc	Neut	Fem
Singular		
dër	daz	diu
dën	daz	die
dëm	dëm	dër
dës	dës	dër
Plural		
die	diu	die
die	diu	die
dën	dën	dën
dër	dër	dër

Masc	Neut	Fem
Singular		
der	das	die
den	das	die
dem	dem	der
des	des	der
Plural		
die	die	die
die	die	die
den	den	den
der	der	der



(van Trijp 2014: 3)

Break

The Rise of the Weak Inflection in Germanic

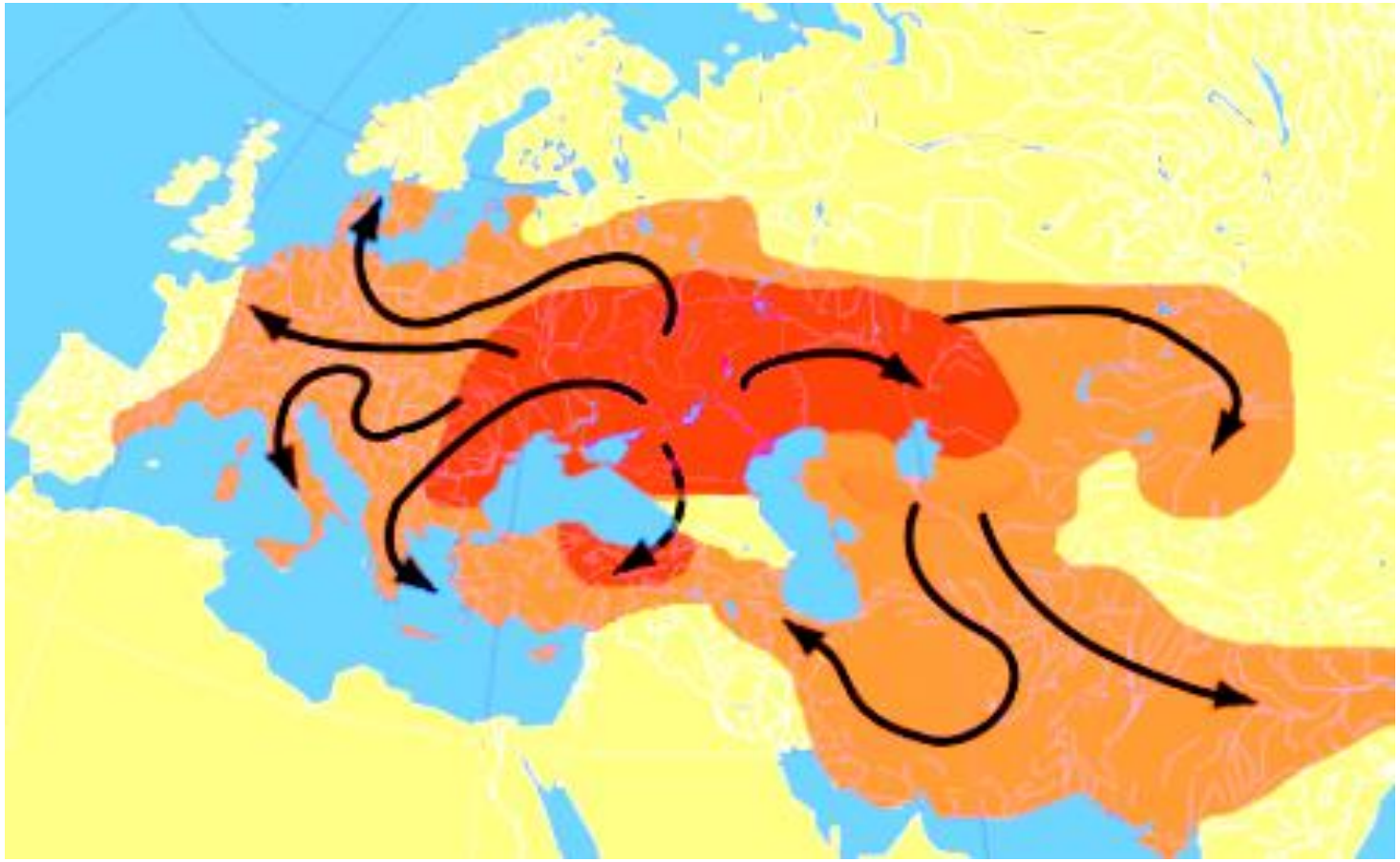
Dirk Pijpops, QLVL research unit, University of Leuven
Katrien Beuls, Artificial Intelligence Lab, Vrije Universiteit Brussel

The Rise of the Weak Inflection in Germanic

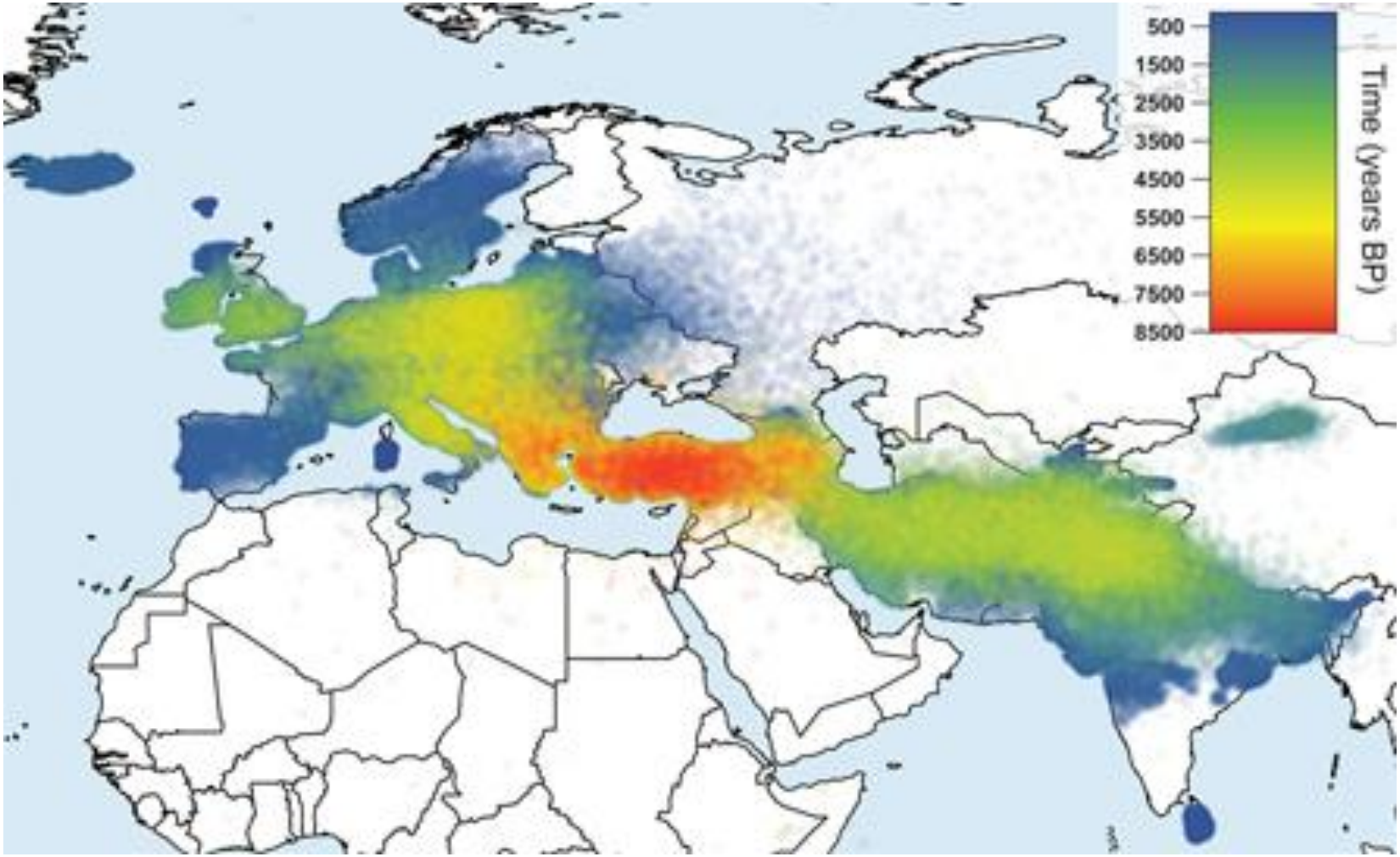
- The Story
- The Model
- The Results
- The Conclusions

(Pijpops & Beuls 2015)

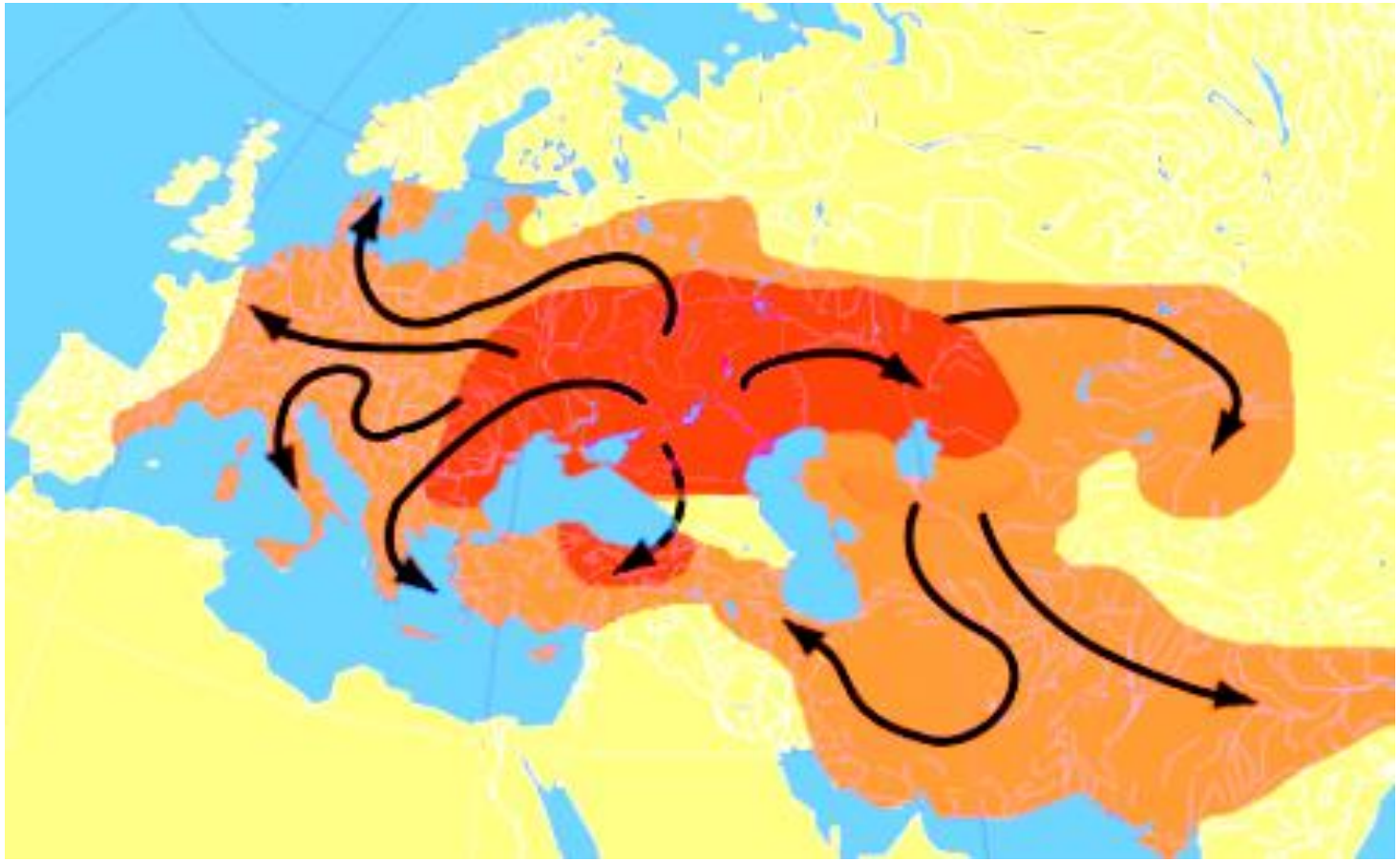
The Story



The Story



The Story



The Story

- Germanic past tense
  - Strong: *ik loop* → *ik liep*
    I run → I run-PAST
  - Weak: → *ik loop**te***
    I run-PAST

The Story

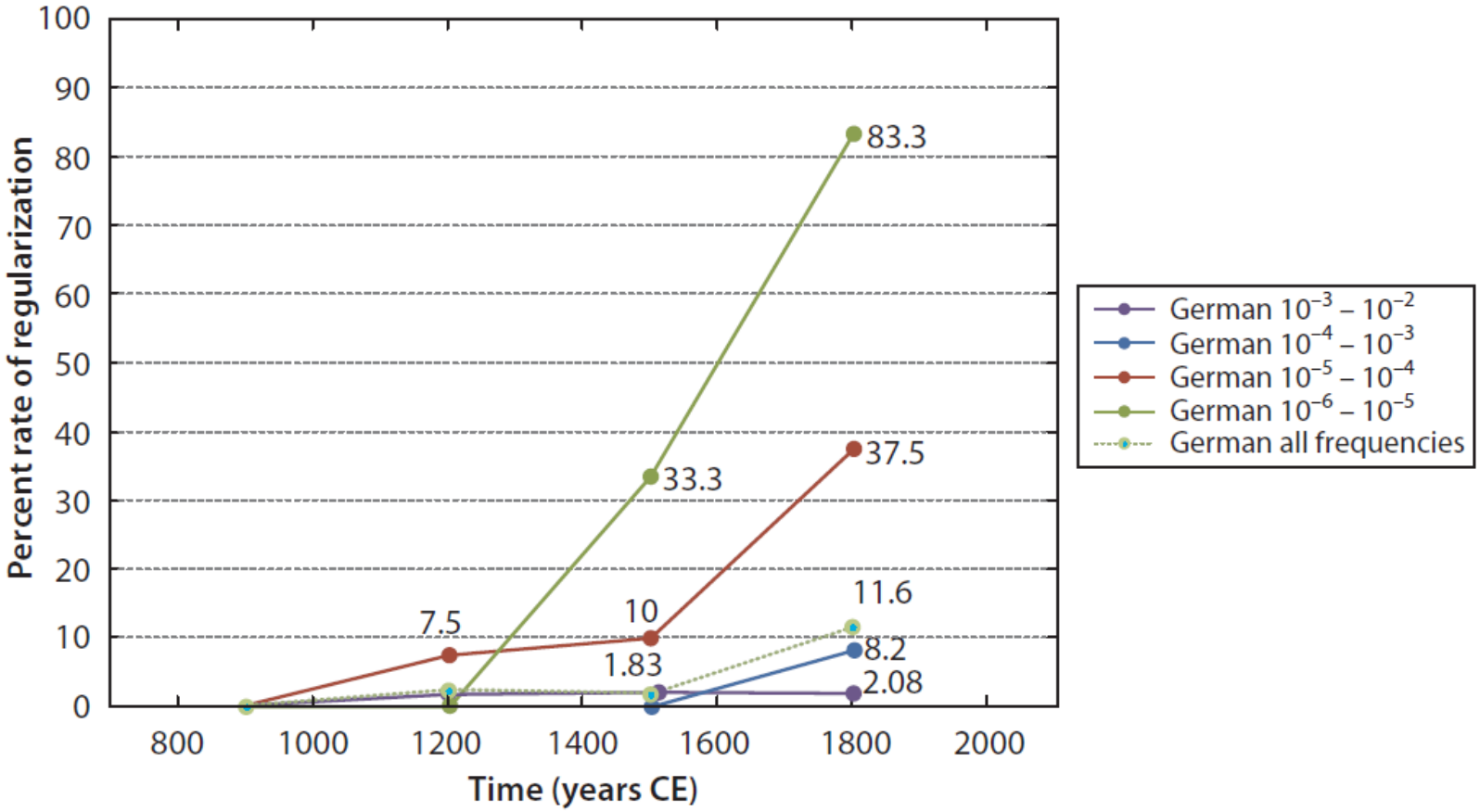
- Competition between the strong and weak strategies

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

The Story

- Germanic past tense
  - Strong: *ik loop* → *ik liep*
    I run → I run-PAST
  - Weak: *ik lopen deed* → *ik loop**te***
    I to run did → 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 schrijf + PAST
'I wrote'

schreef

schrijfde

The Model

- Competition formula:

Ik schrijf + PAST
'I wrote'

p("*schreef*") ~ heard("*schreef*") + heard(pattern 1)

p("*schrijfde*") ~ heard("*schrijfde*") + heard(pattern weak)

The Model

- Competition formula:

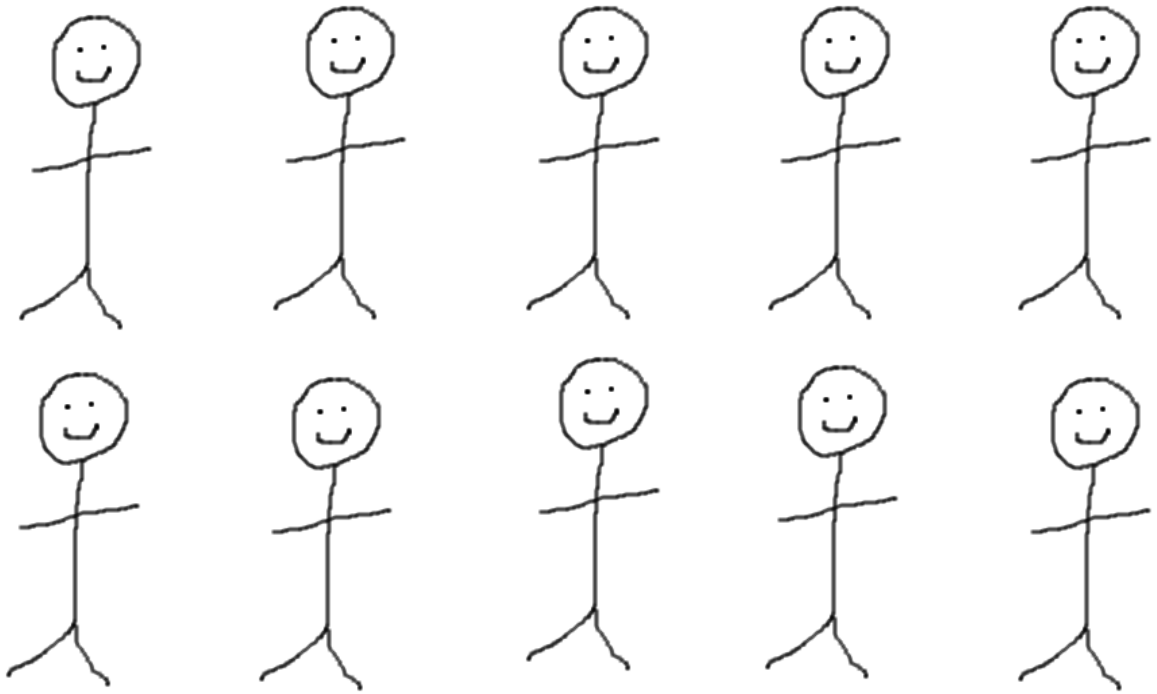
Ik schrijf + PAST
'I wrote'

p("*schreef*") ~ heard("*schreef*") + heard(pattern 1)

p("*schrijfde*") ~ heard("*schrijfde*") + heard(pattern weak)

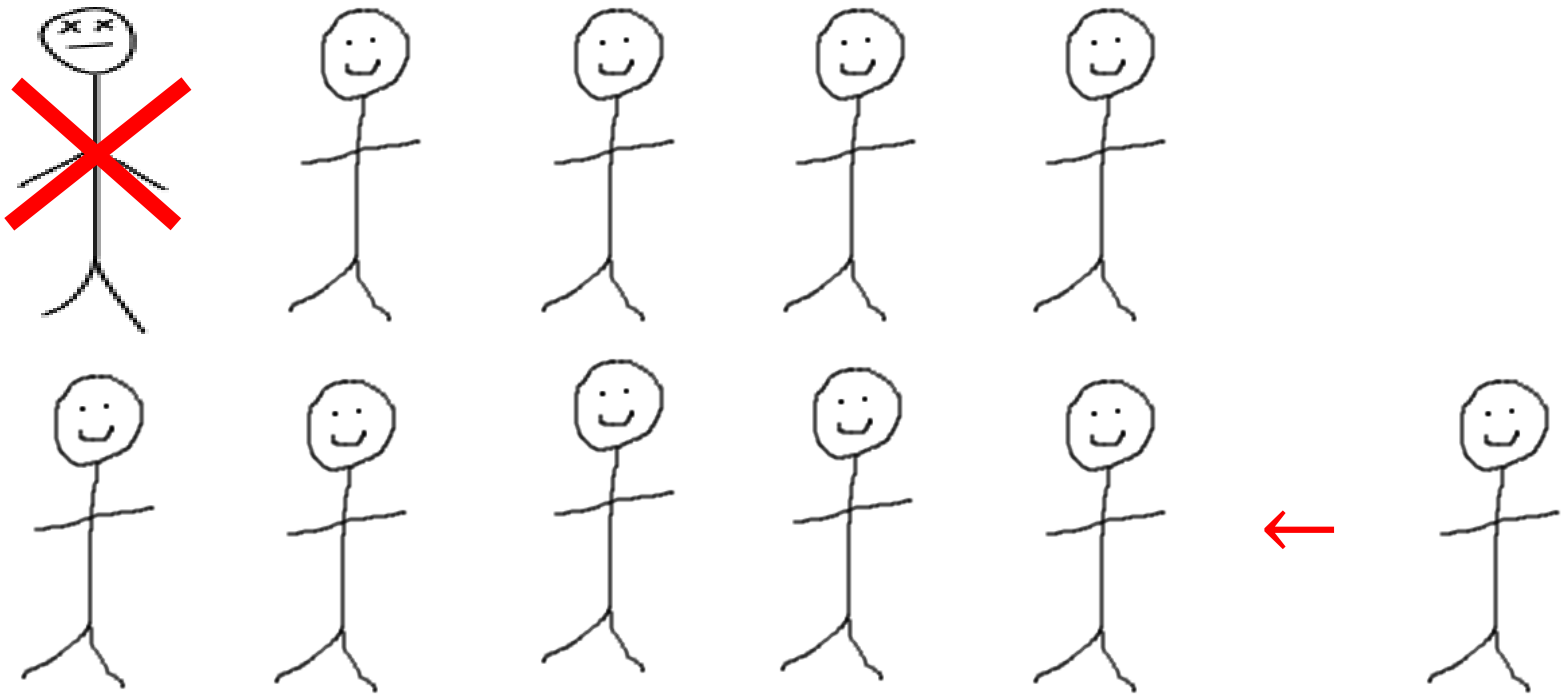
The Model

- Replenishment



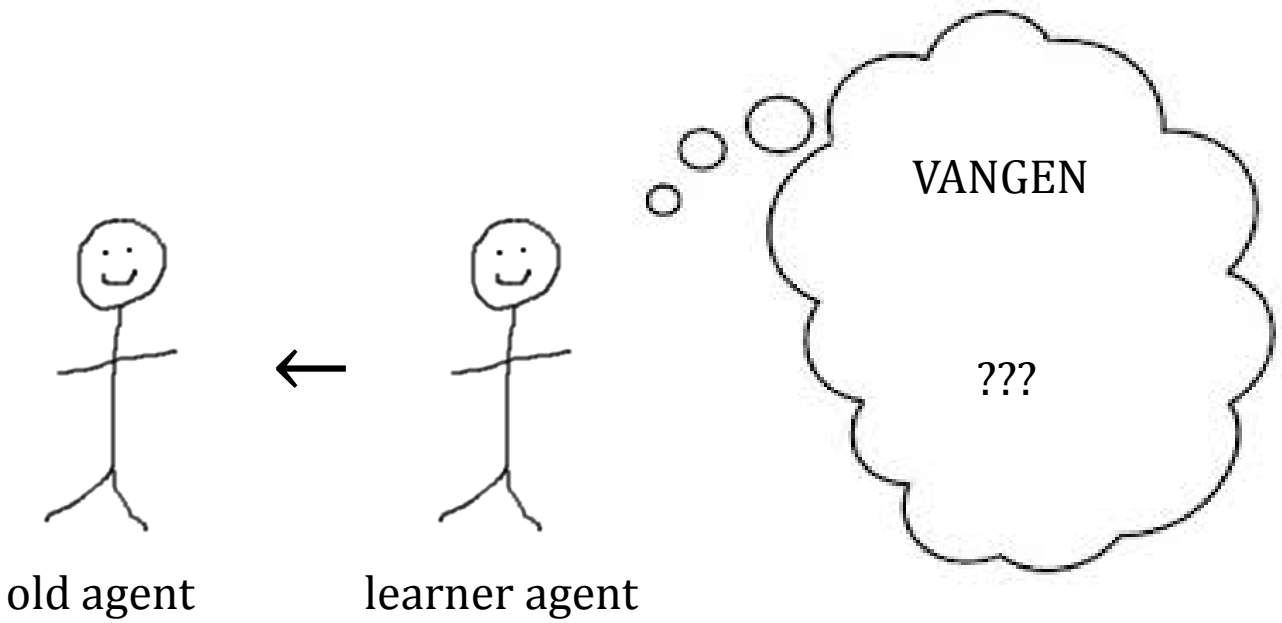
The Model

- Replenishment



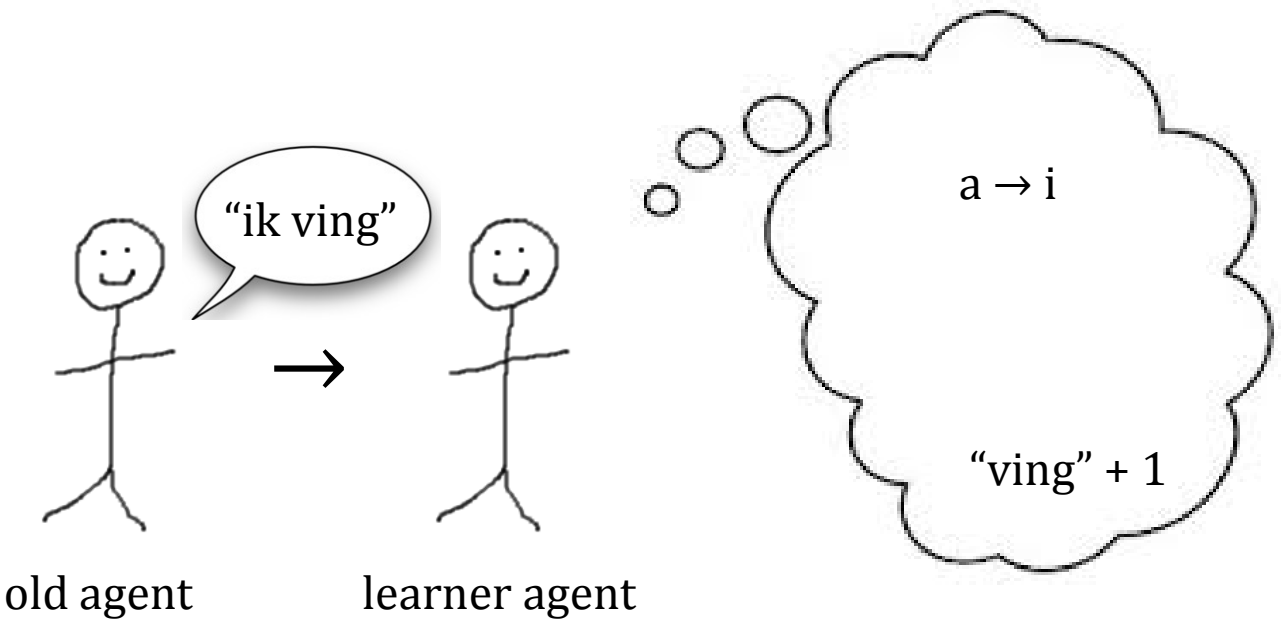
The Model

- Learning



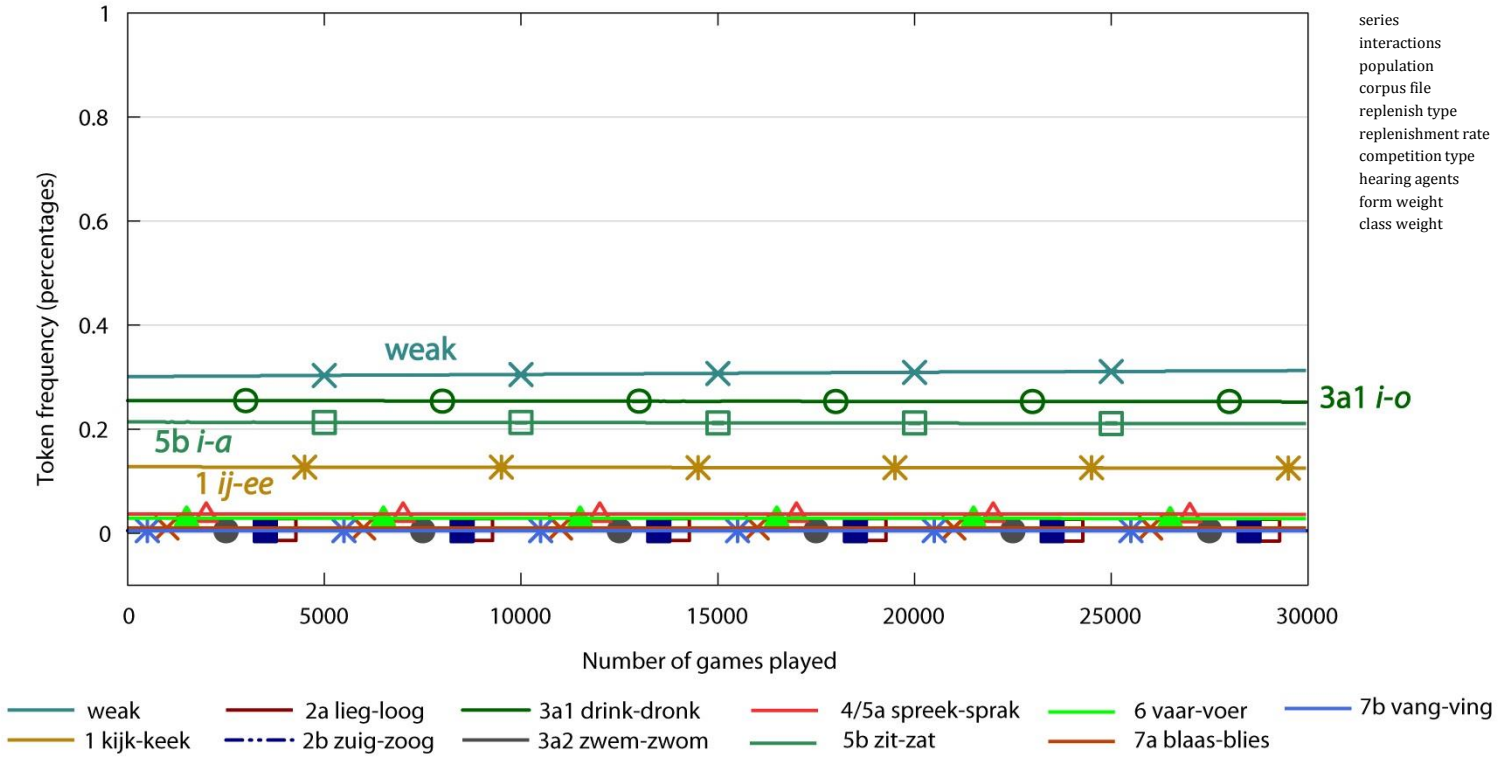
The Model

- Learning



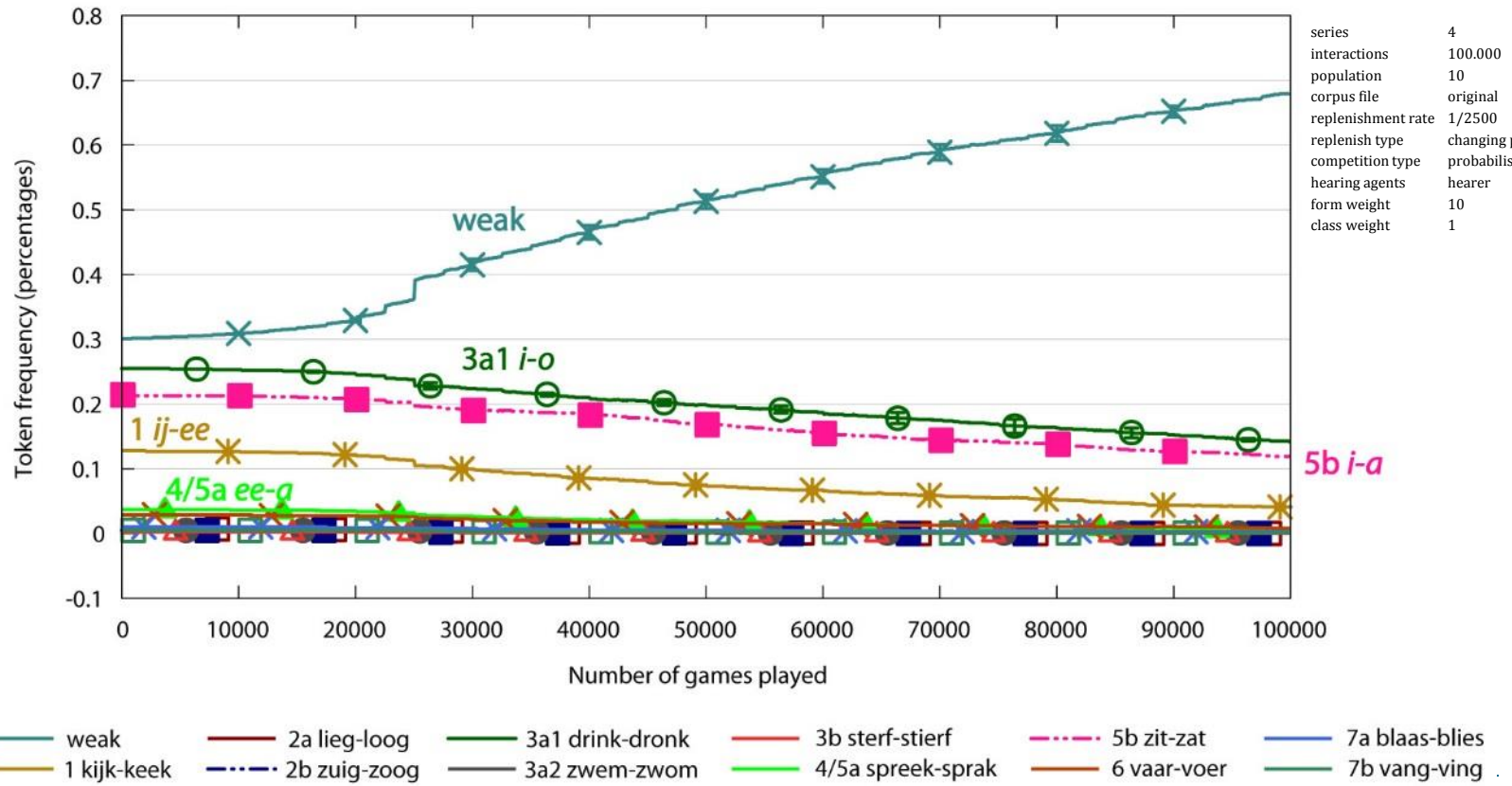
Results

- Original corpus input, no replenishment



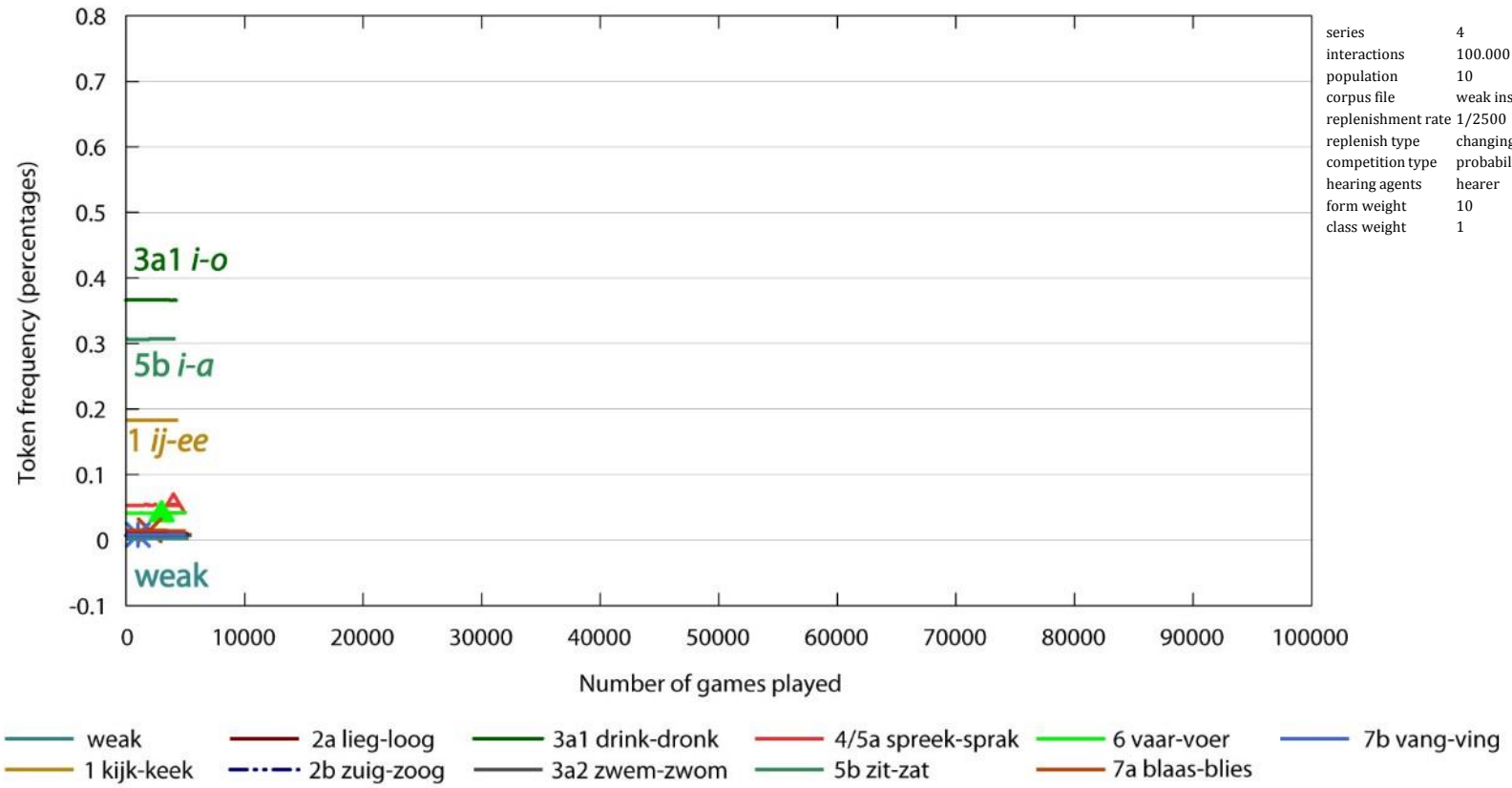
The Results

- Original corpus input, new agent every 2500 interactions



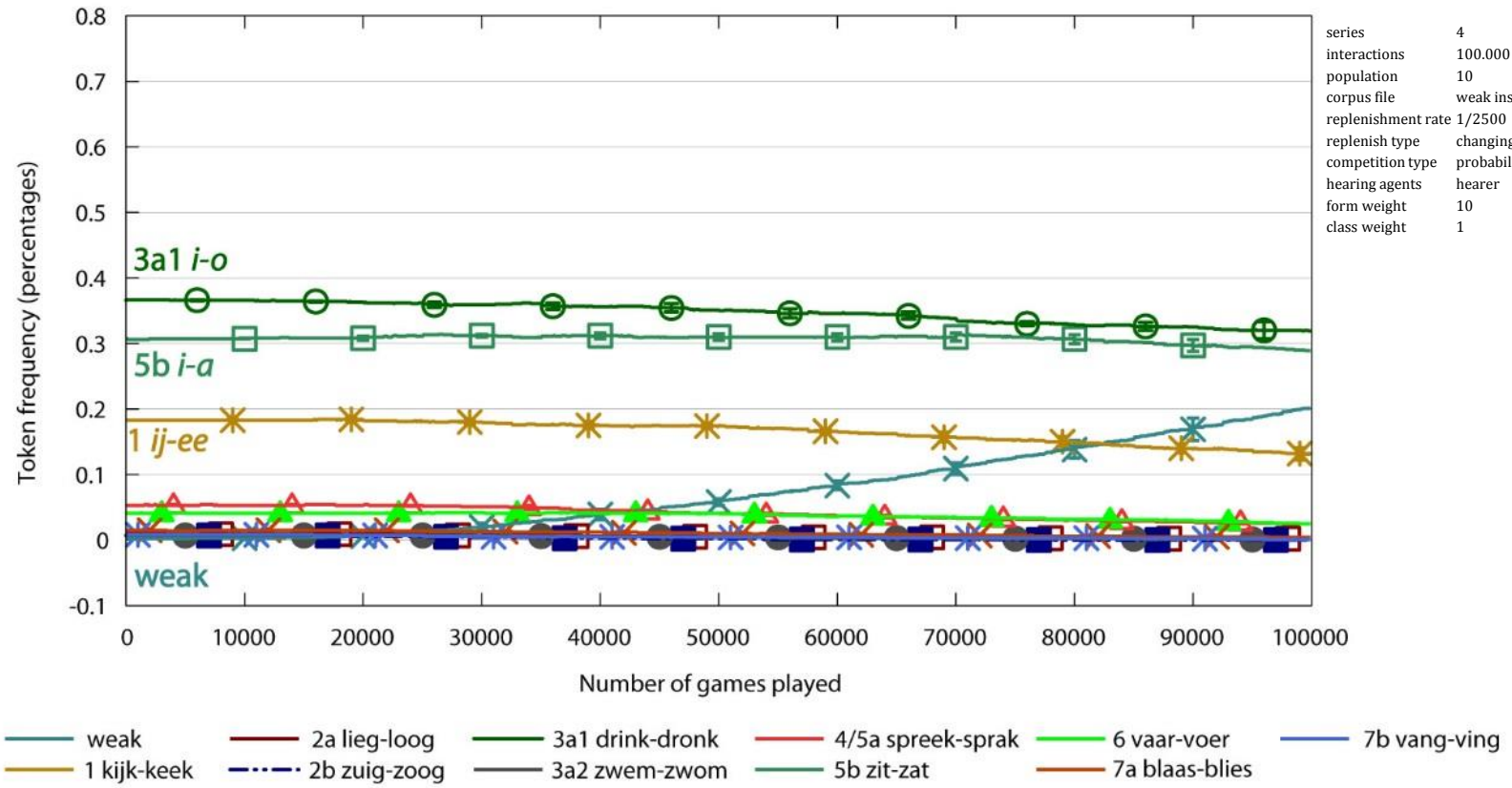
The Results

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



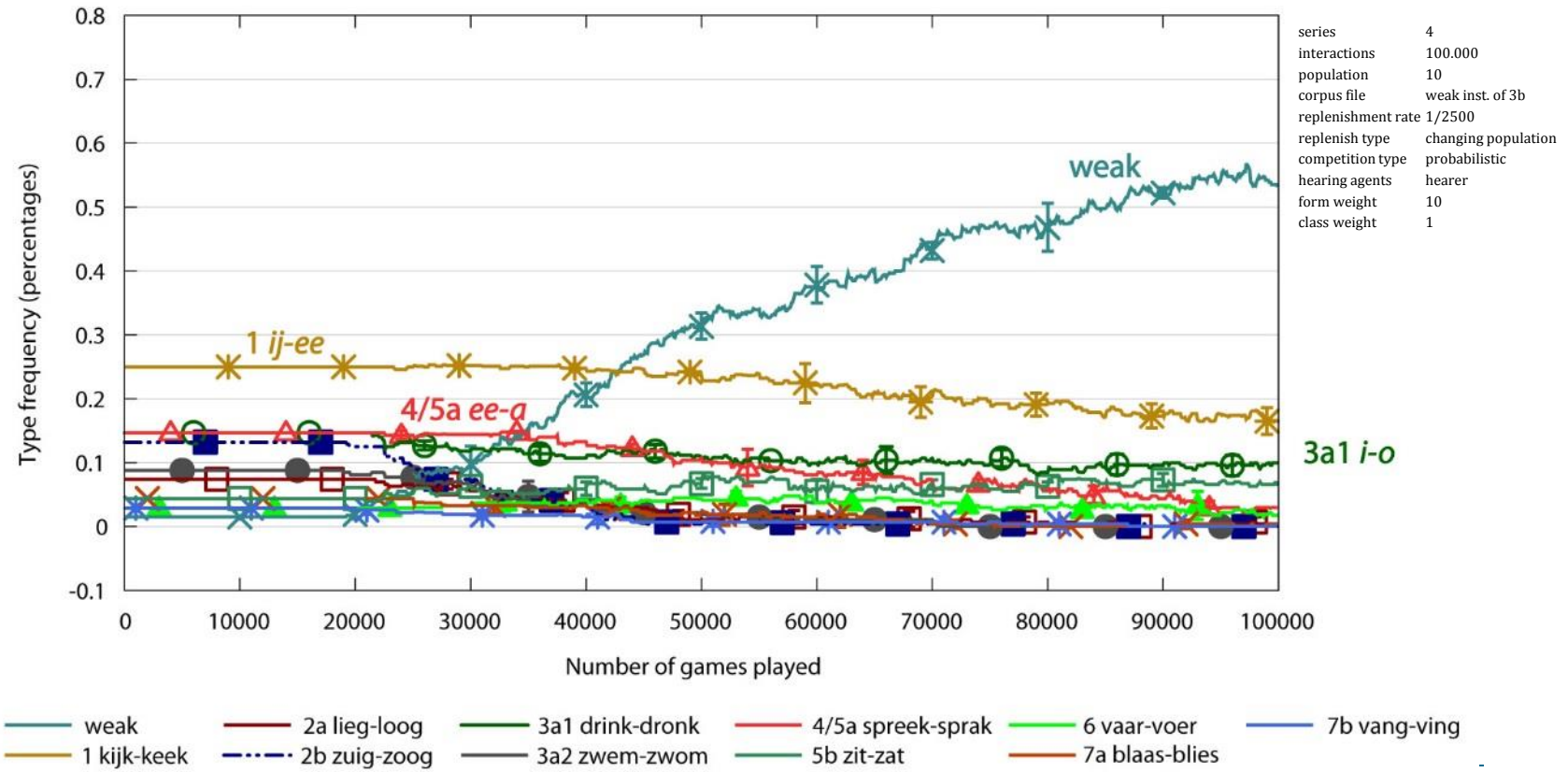
The Results

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



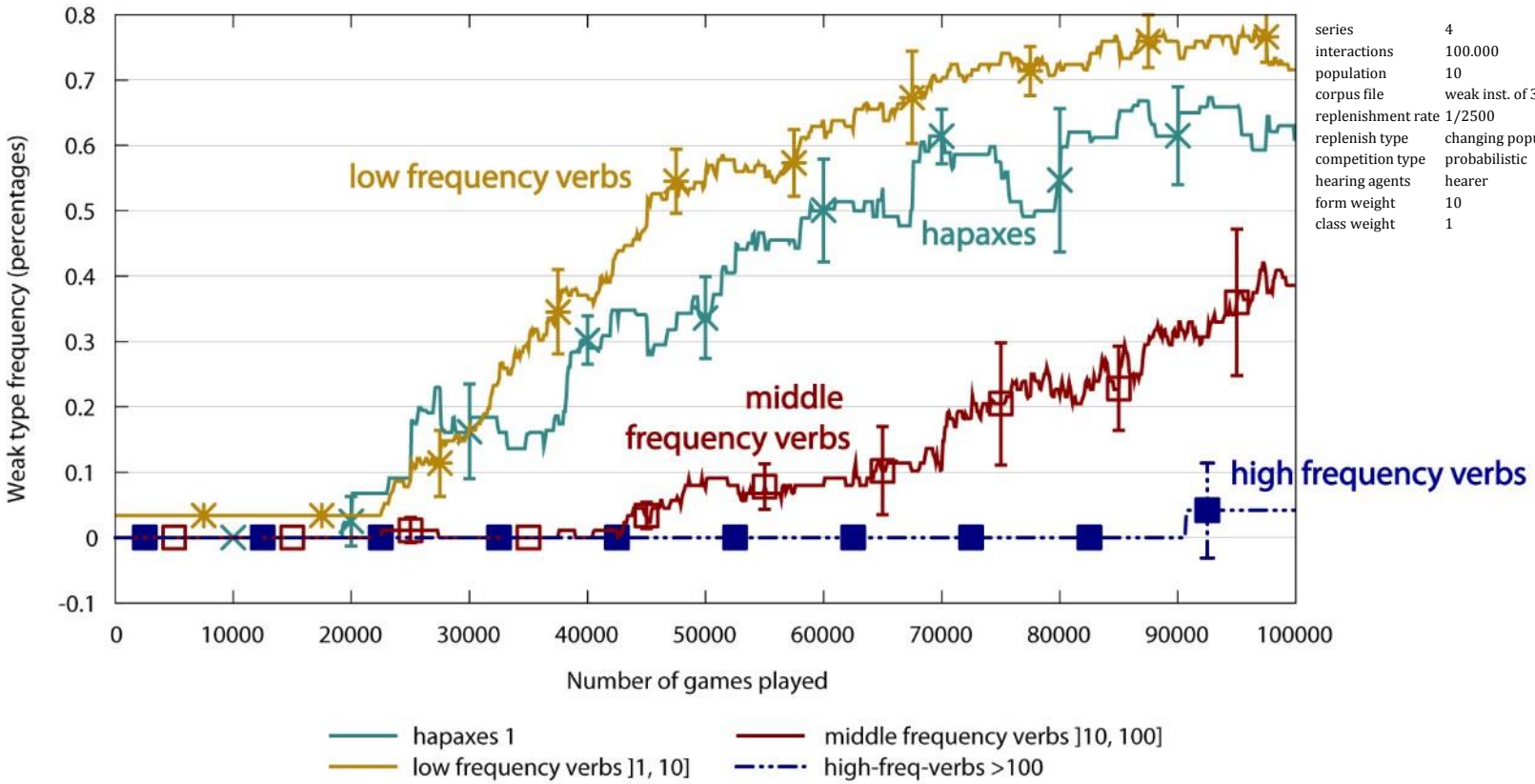
The Results

- 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

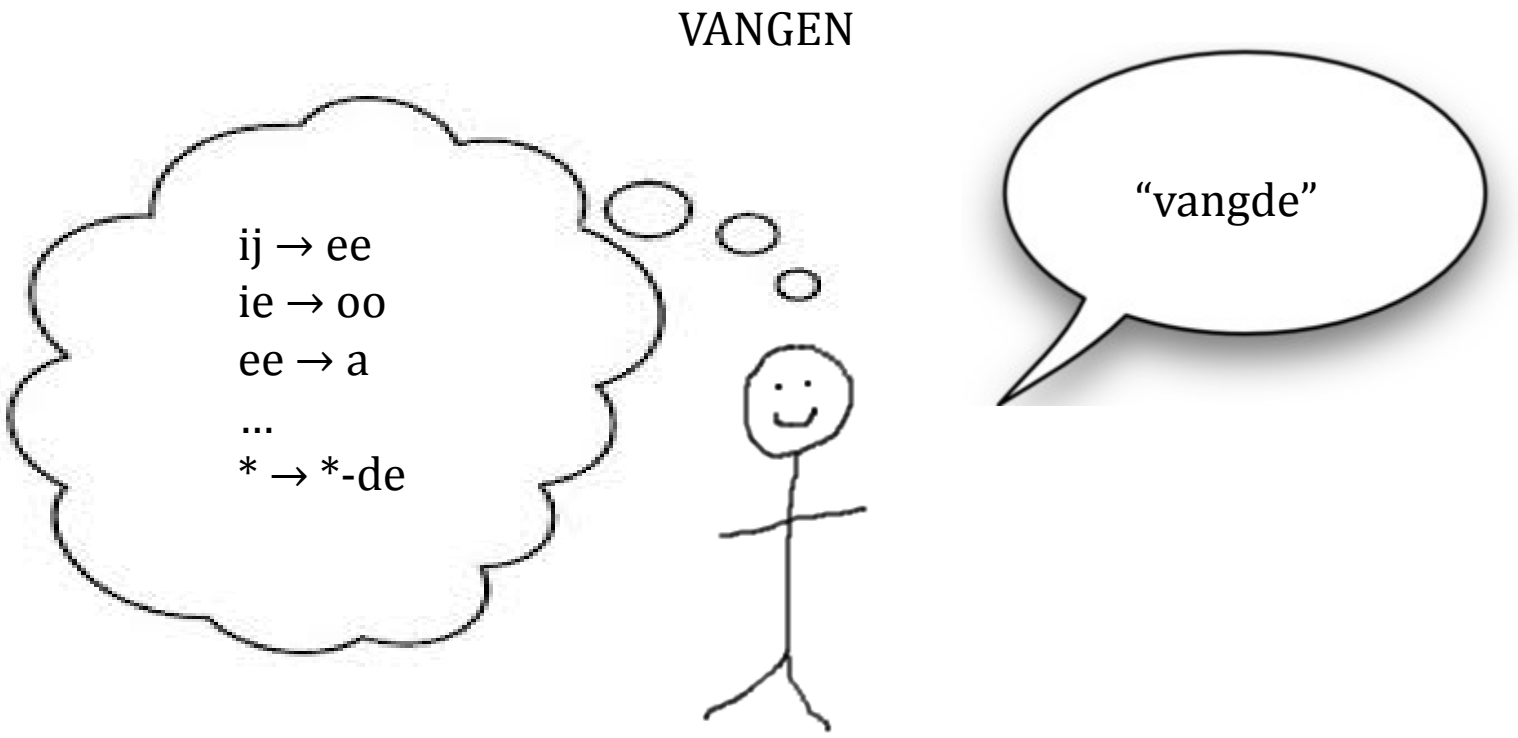
- Conditions which favor the weak inflection
  - Sociohistorical conditions: many language learners
  - Linguistic conditions: low frequent verbs

  ⇒ Both inflections can co-exist for a long time

  ⇒ Expansion of weak inflection can be slowed down

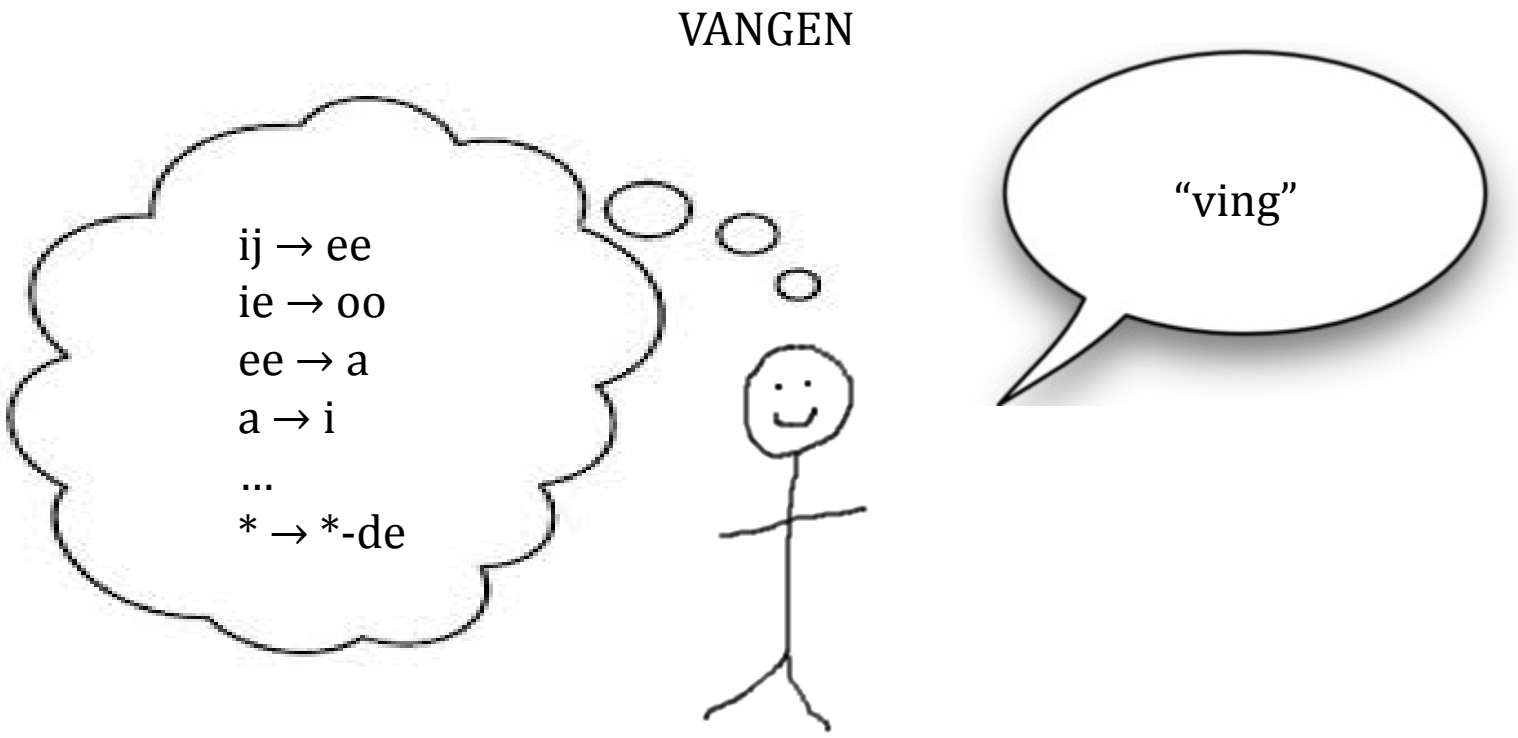
The Story

- No short term effect



The Story

- Criticism



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

General conclusions

- 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

Baerman, Matthew. 2009. Case Syncretism. *The Oxford handbook of case*, 219–230. Oxford: Oxford University Press.

Bazghandi, Ali. 2012. Techniques, Advantages and Problems of Agent Based Modeling for Traffic Simulation. *International Journal of Computer Science Issues* 9(1). 115–119.

Bopp, Franz. 1885. *A comparative grammar of the Sanskrit, Zend, Greek, Latin, Lithuanian, Gothic, German, and Slavonic languages.* 4th ed. London: Williams.

Carroll, Ryan, Ragnar Svare & Joseph Salmons. 2012. Quantifying the evolutionary dynamics of German verbs. (Ed.) Ryan Carroll. *Journal of Historical Linguistics* 2(2). 153–172

Dhamdher, Amogh & Constantine Dovrolis. 2009. An agent-based model for the evolution of the internet ecosystem. *COMSNETS. 1st International Conference on Communication Systems and Networks*, 1–10.

Guerreiro, Orlando & Miguel Ferreira. 2013. Towards an Agent Based Modeling : The Prediction and Prevention of the Spread of the Drywood Termite Cryptotermes brevis. *Progress in Artificial Intelligence. 16th Portuguese Conference on Artificial Intelligence, EPIA 2013*, 480–491.

Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford university press.

Petré, Peter & Freek Van de Velde. 2014. Tracing real-life agents' individual progress in ongoing grammaticalization. In Luc Steels, Freek Van de Velde & Remi Van Trijp (eds.), *How grammaticalization processes create grammar: from historical corpus data to agent-based models*, 48–51.

Pijpops, Dirk & Katrien Beuls. 2015. Strong "islands of resilience" in the weak flood. Dutch strategies for past tense formation implemented in an agent-based model. *Computational Linguistics in the Netherlands (CLIN)*. February 6, Antwerp.

Steels, Luc & Michael Spranger. 2008. The robot in the mirror. *Connection Science* 20(4). 337–358.

Trijp, Remi van. 2012. Self-Assessing Agents for Explaining Language Change: A Case Study in German. *Proceedings of the 20th European Conference on Artificial Intelligence.* Montpellier: France.

Trijp, Remi van. 2013. Linguistic Selection Criteria for Explaining Language Change: A Case Study on Syncretism in German Definite Articles. *Language Dynamics and Change,* 3(1):105-132.

Trijp, Remi Van. 2014. Fitness landscapes in cultural language evolution: a case study on german definite articles.