Observing and simulating changes in the Germanic past tense system
Dirk Pijpops & Isabeau De Smet
Quantitative Lexicology & Variational Linguistics, University of Leuven
Research Foundation Flanders (FWO)

How to combine empirical research with computer simulation and why?

**THEORIZE**

*Assume*
- Openness
- Single mechanism: exemplar-based
- Fledgling weak inflection

*Do not assume*
- Irregularity
- Memory constraints
- Segmentability

*Why*
- Use Occam’s Razor
- Shift the burden of proof
- It’s ridiculously easy: Babel2

---

**OBSERVE**

What does reality look like?

1. Gradual Rise

   Weakened verbs

   Time [Anno Domini]

2. Conserving Effect

   Extremely low frequency verbs
   Low frequency verbs
   Middle frequency verbs
   High frequency verbs

   Time [Anno Domini]

3. Class Resilience

   Weakened verbs in 2000 AD

   Frequency < 0.01%

   Weak instances

---

**SIMULATE**

What should reality look like?

1. Gradual Rise

   Weakened verbs

   Time (millions of interactions)

2. Conserving Effect

   Extremely low frequency verbs
   Low frequency verbs
   Middle frequency verbs
   High frequency verbs

   Time (millions of interactions)

3. Class Resilience

   Weak instances

---


Empirical Data

To allow for easy comparison with English (Lieberman et al. 2007) and German (Carroll et al. 2012), the data selection procedure replicated that of these earlier studies as closely as possible. 164 verbs were selected which were marked as strong in several dictionaries and reference grammars of Old Dutch (860-1200) and which could be tracked in dictionaries or reference grammars of Middle Dutch (1200-1500), Modern Dutch (1500-1900) and Contempory Dutch (1900 onwards; see references for used dictionaries and grammars). These verbs were coded as strong (1), varying (0.5) or weak (0). Only base forms without suffixes were taken up, unless exclusively complex forms were attested. While coding, only preterite forms were considered, not participles. Not selected were the preterite-presents, irregular weak verbs, and verbs whose choice of preterite was dependent upon its meaning. The frequency of each verb was counted in the Corpus of Spoken Dutch, and divided by the total frequency of all verbs in the corpus. The 4 frequency bins shown in the graph above contain verbs with frequency > 1%, 1%-0.1%, 0.1%-0.01%, and < 0.01%.

Simulation Design

Before each interaction, a verb is selected from a set of 40 nonsense verbs. Each verb’s chance of being selected corresponds to its frequency. These frequencies follow a Zipfian distribution, with the verb v of rank n having the frequency freq(v) = n⁻¹/log(n). Next, a speaker and a hearer agent are randomly selected from a population of 100 agents and interact according to the flow chart above. All starting agents are initiated with a memory of 39 strong forms for the 39 most frequent verbs and a single weak form for the least frequent verb. The initial memory count of verb v of rank n is count(v) = n⁻¹/log(n). The 39 initially strong verbs are distributed across 7 about classes to create classes with equal token frequency, but different type frequency and vice versa. Every 10,000 interactions, 1 agent is replaced by a new agent with an empty memory. In the current settings, verbs are never replaced. The graphs show the running averages and standard deviations of 20 series of each 20 million interactions. The 4 frequency bins shown in the graph above contain verbs with frequency > 4%, 4%-1.5%, 1.5%-0.7%, and < 0.7%.

Acknowledgments

We cordially thank Kathien Beuls and Freerk Van de Velde for their indispensable contributions to both the empirical study and the agent-based simulation. In addition, we would like to thank Remi van Trijp for interesting discussions and useful advice about the simulation, as well as the participants of the SLE-48 workshop Shifting classes: Germanic strong and weak preterites and participles.

References