The evolutionary dynamics of language as a function of demography

Dirk Pijpops, Katrien Beuls & Freek Van de Velde

Research Foundation Flanders (FWO)
Artificial Intelligence Lab, Vrije Universiteit Brussel
QLVL, University of Leuven

Constant Rate Hypothesis

- Weak vs. strong verbal inflection: *throwed* vs. *threw*
- *The half-life of irregular verbs is proportional to the square root of their frequency* (Lieberman et al. 2007: 714)
- Language changes independently of its sociocultural context

Contra Constant Rate Hypothesis

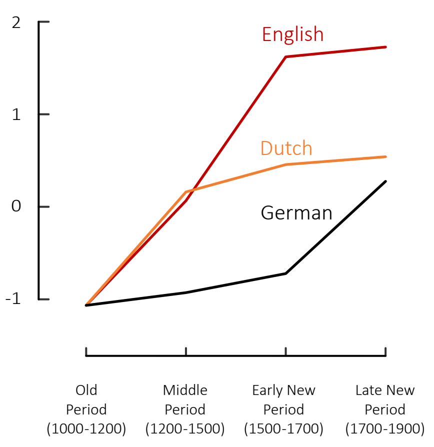
- Constant Rate does not hold for German (Carrol et al. 2012)
- Constant Rate does not hold for Dutch (De Smet 2016)
- Constant Rate does not hold for English, given extra measurement point (De Smet et al. 2017, cf. also Cuskley et al. 2014)

Alternative: Linguistic Niche Hypothesis

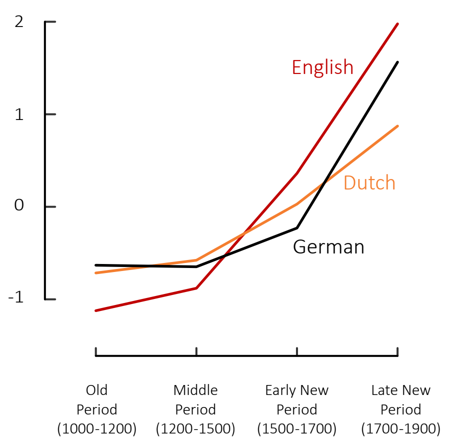
- Dale & Lupyan (2012)
- Different degrees of urbanization in English, Dutch and German areas; urbanization means immigration in pre-industrial Europe; immigration means interdialectal contact
- Weak inflection profits from competition between dialectal forms thanks to its general applicability (Pijpops et al. 2015)

Real World

Cumulative percentage of weak verbs, z-scored

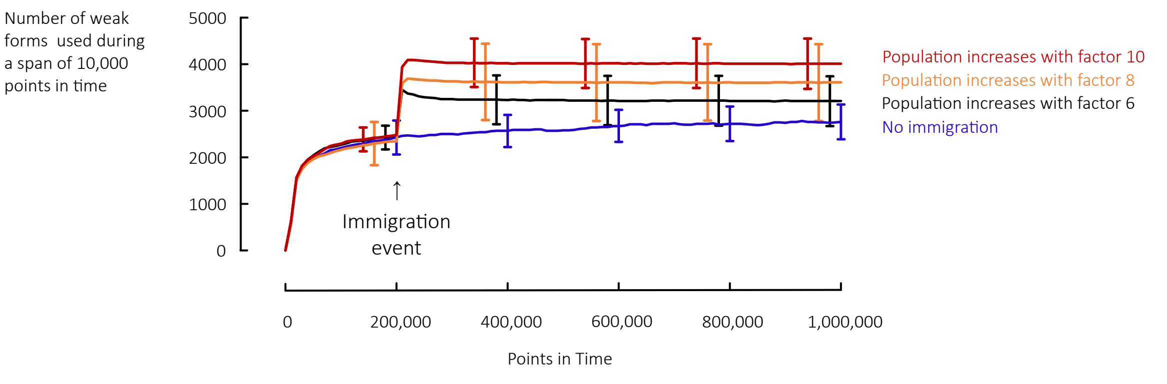


Logarithm of the population size of the largest cities, z-scored



	Demography		
	English	Dutch	German
English language	**0.87** (p=0.13)	0.86 (p=0.31)	0.69 (p=0.31)
Dutch language	**0.72** (p=0.28)	**0.72** (p=0.28)	0.56 (p=0.44)
German language	0.96 (p=0.03)	0.97 (p=0.03)	**0.99** (p=0.01)

Simulation



Pijpops, Dirk, Katrien Beuls & Freek Van de Velde. 2018. **The evolutionary dynamics of language as a function of demography**. *Evolang 12*. April 17, Torun.

Pijpops, Dirk, Katrien Beuls & Freek Van de Velde. 2015. **The rise of the verbal weak inflection in Germanic. An agent-based model**. *CLIN-Journa*l, 5, 81–102.

Observational data

For the demographic data, we made use of the combined databases of Bairoch et al. (1988), De Vries (1984), Mitchell (1998), and, for the year 1900, supplemented by Chandler (1987), giving estimates for the population size of European cities through time. We sampled the urban population size of the U.K., Northern Belgium & The Netherlands, and Germany & Austria, concentrating on those cities which have a population of at least 100,000 inhabitants in 1800. The growth of these cities is too large to have come about through natural births, and strongly suggests immigration (Howell 2006: 208). For the linguistic data, we made use of the datasets of Lieberman et al. (2007) for English, De Smet (2016) for Dutch, and Carroll et al. (2012) for German. The second graph depicts the weighted mean of the log-transformed number of inhabitants of the 3 largest cities in 1000-1200, the 5 largest cities in 1200-1500, the 7 largest cities in 1500-1700, and the 9 largest cities in 1700-1900 (z-scored). This increment was motivated by the fact that there are fewer cities whose population can be reliably estimated for the older time periods.

Simulation Design

The simulation is made up of a population of computational agents that communicate about past events. At each point in time, the population forms up into pairs of interacting agents and one of 40 different events is selected for each pair. The probability p_n that the *n*th most frequent event e_n is selected, follows a distribution that instantiates Zipf's Law, as $p_n = [100/n] / \sum_{i=1}^{40}[100/i]$. In each pair of agents, the randomly designated speaker agent will have to express an event e_n using a past tense form of the corresponding verb v_n. The speaker chooses this past tense form according to the probabilities in its individual language system. This language system is exemplar-based, i.e. agents simply retain all past tense forms they hear in memory, which may thus contain several past tense forms for the same verb v_n. This would correspond to, for instance, an English speaker having heard *drank, drunk* and *drinked* as past tense forms of the event of *drinking*. For each past tense form $f_{n,m}$ of verb v_n that instantiates grammar rule g_m, a count $c_{n,m}$ is retained, which tallies the number of times the agent has heard the form, representing its entrenchment in memory. The probability $q_{n,m}$ with which the speaker selects form $f_{n,m}$ from the competitors $\{f_{n,t}, \ldots f_{n,u}\}$ is directly derived from these counts, according to $q_{n,m} = c_{n,m} / \sum_{i=t}^{u} c_i$. In the case of no past tense forms of v_n being present in memory, the agent reverts to checking which known grammar rules may be applicable to v_n. The probability r_m with which the agent applies a grammar rule g_m with count d_m from the applicable competitors $\{g_t, \ldots g_u\}$ is calculated analogous to the choice of forms, as $r_m = d_m / \sum_{i=t}^{u} d_m$. The selected rule is then used to build a new past tense form. Should no past tense form or grammar rule be available in the speaker's memory, the speaker remains silent.

Two types of grammar rules are available to the agents. First, 7 vowel-alternating rules $\{g_1, \ldots g_7\}$, such as English *i ⟶ a,* as in *sing ⟶ sang* and *drink ⟶ drank,* correspond to original 7 classes of the Germanic strong inflection and represent the initially dominant system. Second, a single suffixation rule g_8, such as English *+ed*, as in *kick ⟶ kicked*, corresponds to the Germanic weak inflection and represents a recent innovation. The only qualitative difference between both types in the simulation is that the application of a vowel-alternating rule is dependent on the stem vowel of the verb, such that a rule *i ⟶ a* cannot be applied to verbs with a stem containing an *a*, like *draw*. Another vowel-alternating rule, such as *a ⟶ e,* may be applicable, though. Meanwhile, the suffixation rule can in principle be applied to any verb, indiscriminately of stem vowel.

Every *h* points in time, a factor *j* of the population is replaced by new agents. To represent the result of a preceding period of language acquisition, these newly introduced agents inherit *k* of the counts in the language system of a parent agent, which is randomly assigned from the surviving population. At the start of the simulation, all agents have perfect knowledge of all initial verbs. Because the weak inflection represents the youngest strategy, it is at first not well-established in the verbal inventory; it will rather need to fight its way to the top from a vastly inferior starting position. Only the least frequent verb is conjugated weakly, with even the least frequent vowel-alternating rule being more than 5 times as frequent as the nascent weak rule.

At some point in time, the population may increase with factor *w* due to immigration. To model immigration from a different dialect area, these new immigrant agents may have access to 0-7 different vowel-alternating rules $\{g_9, \ldots g_{15}\}$. An agent may thus have access to maximally three different past tense forms for any verb: one formed according to the applicable grammar rule from the 'native' vowel-alternating rules $\{g_1, \ldots g_7\}$, one formed according to the weak inflection g_8, and one formed according to the applicable grammar rule from $\{g_9, \ldots g_{15}\}$. Our hypothesis states that these immigrants should have a positive effect on the success of the weak inflection. Equipping these agents with any initial knowledge of the weak inflection would equate to building in this effect. As such, these immigrant agents never have any knowledge of the weak suffixation rule g_8 upon entering the simulation, but will need to acquire this rule through contact with the original agents. The third graph shows running averages and standard deviations over 100 series with 10 starting agents. The parameter settings are h = 100, j = 0.1, k = 0.3.

Acknowledgments

We are thankful to Ryan Carroll, Ragnar Svare, Joseph Salmons and Isabeau De Smet for generously sharing their data.

References

Bairoch, P., Batou, J., & Chèvre, P. (1988). *La population des villes européennes: banque de données et analyse sommaire des résultats, 800-1850.* Genève: Droz.

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

Chandler, T. (1987). *Four thousand years of urban growth.* Lampeter: Edwin Mellen Press.

Colaiori, F., Castellano, C., Cuskley, C., Loreto, V., Pugliese, M., & Tria, F. (2015). General three-state model with biased population replacement: Analytical solution and application to language dynamics. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, *91*(1–1), 12808.

Cuskley, C., Pugliese, M., Castellano, C., Colaiori, F., Loreto, V., & Tria, F. (2014). Internal and External Dynamics in Language: Evidence from Verb Regularity in a Historical Corpus of English. *Plos One*, *9*(8), e102882.

Dale, R., & Lupyan, G. (2012). Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in Complex Systems*, *15*(3), 1150017-1-1150017–16.

De Smet, I. (2016). *De verzwakking van het preteritum in het Nederlands*. Master's thesis, University of Leuven.

De Smet, I., Beuls, K., Pijpops, D., & Velde, F. Van de. (2017). Language-specific differences in regularization rates of the Germanic preterite. In *International Conference on Historical Linguistics (ICHL)*. 2 August, San Antonio.

De Vries, J. (1984). *European urbanization 1500 - 1800*. London: Methuen.

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., & Nowak, M. (2007). Quantifying the evolutionary dynamics of language. *Nature*, *449*(7163), 713–716.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS One*, *5*(1), e8559.

Mitchell, B. (1998). *International historical statistics : Europe 1750-1993* (4th ed.). London: Macmillan.

Pijpops, D., Beuls, K., & Van de Velde, F. (2015). The rise of the verbal weak inflection in Germanic. An agent-based model. *Computational Linguistics in the Netherlands Journal*, *5*, 81–102.

Ten Kate, L. (1723). *Aenleiding tot de kennisse van het verhevene deel der Nederduitsche sprake. Eerste deel [Introduction to the understanding of the lofty part of the Dutch Language. First part].* Amsterdam: R. & G. Wetstein.