An agent-based model of linguistic diversity

Pieter de Bie
AI department
Rijksuniversiteit Groningen, the Netherlands
p.de.bie@ai.rug.nl

Bart de Boer
Institute of Phonetic Sciences
Universiteit van Amsterdam, the Netherlands
b.de.boer@ai.rug.nl

Abstract

Research in the field of language diversity mostly uses models where emerging language patterns visible in the real world can be hard to find. An example of this is the dialect continuum we see between the Netherlands and Germany, or the language border we find on the Dutch-French-line in Belgium.

The agent based model described here introduces language diversity as a consequence of language mutation. Agents adopt these mutations from other agents based on social impact theory, which states that less common varieties have a relatively bigger influence per individual speaking the variety. Using this model, it was possible to show different language patterns existing at the same time.

1 Introduction

Linguistic diversity is truly amazing. Not just in the infinite richness with which the approximately 6000 different languages of the world can express meaning, but also in the fact that it exists at all. Although there are indications that dialects and diversity also exist in animal communication systems – chimpanzees and lyrebirds (among other birds) appear to have dialects, and humpback whales appear to change songs regularly – nowhere is the diversity as pronounced as in human language. Although *why* language is so diverse is an interesting question, in the work presented here, we are concerned with the question how language can become diverse.

As with many aspects of language, language diversity can be studied from the perspective of individual behavior and from the perspective of the language as a whole. At the level of the individual, two opposing behaviors can be observed. When infants acquire a language, they learn a language that is very close to the language of their parents and their peers. Much closer, in fact, than is needed for successful communication. Although speakers of different closely related dialects have no trouble communicating with each other, they have equally little trouble in recognizing their interlocutor as the speaker of a different dialect. The opposite behavior to this tendency to conform, is a tendency to innovate. Speakers of language, especially young ones, are constantly renewing their language. This can be a conscious process (when inventing new, cool ways of saying things, for example) or an unconscious process (when adopting subtle variants of pronunciation to differentiate oneself from the old farts of the group, for example).

On the collective level, the level of the language of a population, this leads to slow but sure language change. This language change leads to two rather different kinds of linguistic diversity. One is called a *dialect continuum*. This occurs when languages gradually merge into each other over an extended geographical range. An example is the range of German and Dutch dialects. A very different situation occurs when neighboring languages are very different: this is called a *language border*. The border between French and Dutch in Belgium is a classical example and the more salient because there is no geographical or political barrier to support it, and because it has been stable for hundreds of years .

The existence of language borders without accompanying political or geographical boundaries indicates that such boundaries cannot always be part of the explanation of linguistic diversity. Migration can also not always be relied on as an explanation. The language border between Spanish and English in North America can obviously be explained as the result of the influx of populations that spoke a different language, but such straightforward explanations are not always possible. Speakers on both sides of the French-Dutch language border are closely related genetically, indicating extensive intermixing of populations, but the language border has remained stable for a long time nevertheless.

1.1 Previous work

It is clear that linguistic diversity is a phenomenon that needs to be explained as the result of an interaction between behavior on a collective level (the languages and corresponding cultural groups) and the behavior of individuals. The dynamics of such interactions can become extremely complex and this is why agent based computer models are a useful tool for investigating the emergence of complexity.

When one can assume geographical barriers, language borders can be explained straightforwardly, and can even be investigated mathematically (Patriarca and Leppanen, 2004). When such barriers cannot be assumed, the situation becomes more difficult, because contrary to for example biological evolution, small mutations do not tend to get carried on by offspring. (Nettle, 1999a) has identified two problems. One is the *averaging problem*, and the other is the *threshold problem*. The averaging problem occurs when the language learned by an individual is a mixture of the languages spoken around it. Any initial difference between languages (or differences caused by linguistic innovation) then tends to average out. Even when one assumes discretely inherited features of language, language change has difficulty spreading because of the threshold problem. Because of random drift in the population, variants of the language that are infrequent tend to disappear, thus hampering the spread of innovation.

That this is a serious problem is illustrated by early work to model linguistic (or cultural) diversity. Axelrod (1997), for example, has created a spatial model of cultural diversity that starts with maximum diversity. He has shown that agents that tend to interact with and copy from like agents converge to a situation where there are sharp borders between a small number of identical groups. At the same time, diversity tends to decrease. The same is true for a model by Barr (2004) that starts with high diversity and converges to few near-homogenous groups.

In order to overcome these problems, Nettle (1999a,b) has made use of social impact theory.

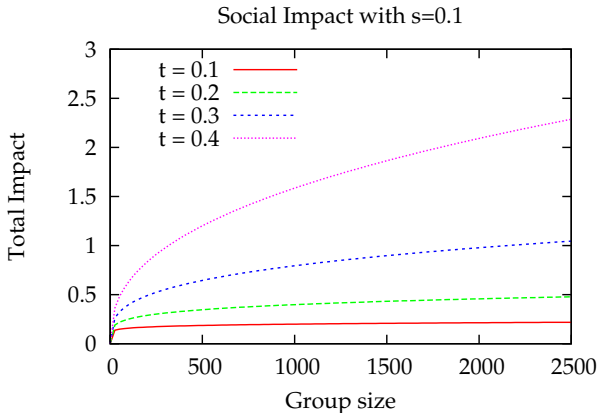


Figure 1: Sociale impact as a function of total group size, with different values of t

This social impact theory, originally devised by Latané (1981), models the influence of a social event given the number of people involved in the event. Latané's findings were that the total impact of the group is a power function in the form:

$$I = sN^t \tag{1}$$

where I is the total social impact, N is the group size and s and t are used for fitting the curve. As can be seen in figure 1, the curve tends to flatten out, thus the individual pressure of participants in the group tends to *decrease* with group size increase.

Nettle (1999b) uses this theory to explicitly boost the rare variant in a language modeled as either variant p or q. Using this, or the use of super-influential individuals, he shows that language diversity can be obtained. Livingstone (2002), finally, has shown that dialect continua can emerge in groups of agents that are situated along a one-dimensional continuum. The bottom line, however, is that none of these models can generate both dialect continua and language borders.

2 Model

We propose a simple spatial model in which both dialect continua and language borders can emerge. This model is based on a population of agents that prefer to interact with agents that speak like them, and sometimes change their language to be more similar to their interlocutors. The agents prefer language variation that are considered a minority, making use of social impact theory.

Languages are represented as strings of binary features (16 in all simulations presented here). The distance between two languages is defined as the number of all features that are different between the two languages, divided by the number of features per language. The meaning of the features is not defined; they may be compared to Nettle (1999a)'s *linguistic items*, and can include words, but also sentence structure or pronounciation hits.

Algorithm 1 Communication between two agents

```
for ∀a ∈ N do {for all agents}
  for ∀b ∈ Neighbourhood(a) do
    if random ≤ p_i(a, b) then
      i ← random vector index
      if a_i ≠ b_i then
        if random ≤ p_a(i, b_i) then
          a_i ← b_i
        end if
      end if
    end if
  end for
end for
```

2.1 Communication and language adoption

In each time step all agents can interact with all their direct neighbors (using 4-connectivity). The whole process of communication can be seen in algorithm 1. The individual steps will be explained further.

The probability of interaction between agents a and b depends on the distance $d(a, b)$ between their languages as follows:

$$p_i = \alpha + (1 - \alpha)10^{-\beta \cdot d(a,b)}$$

where α is a minimal interaction probability and β determines how fast p_i drops off with increasing distance.

When an agent interacts with an other agent, it selects a random feature from its language, and when it is different, adopts it from the other agent with a probability:

$$p_a = \max\left(1, \gamma\left(\frac{N_{i,v}}{N}\right)^{\kappa-1}\right)$$

where γ is a minimal adoption probability, $N_{i,v}$ is the total number of agents that use variant v of feature i, and N is the total number of agents. The parameter $0 \leq \kappa \leq 1$ determines how fast p_a drops off with the number of speakers of variant i. This probability causes rare variants to be more desirable.

This formula is in fact based on the social impact theory as in equation 1. However, since the communication proceeds on an individual level, the formula calculates individual social pressure and not total group pressure.

2.2 Language introduction and stability

Each time step, an agent mutates its language with a probability p_m set to a fixed value in the range $[10^{-5} : 10^{-3}]$. A mutation means the flipping of a single feature in the agents linguistic vector.

Agents can learn for the first ten time steps of their lives. After this period, they can only function as language providers, thus providing some stability in the system. Agents are susceptible to die from their 15th time step, after which they have a probability of dying of 50% each time step.

After agent death, a new agent replaces the old one. This agent will speak the language most common among his direct neighbors. The alternative to this choice is to pick the most common variant of each individual language feature. Since this could result in a mix of languages, where the child is not able to speak with any of its parents, this seems implausible, and thus the idea was discarded.

3 Results

The described model was implemented in the Repast framework(Railsback et al., 2006). For the random functions, the Mersenne Twisteer MT19937 (Matsumoto and Nishimura, 1998) was used.

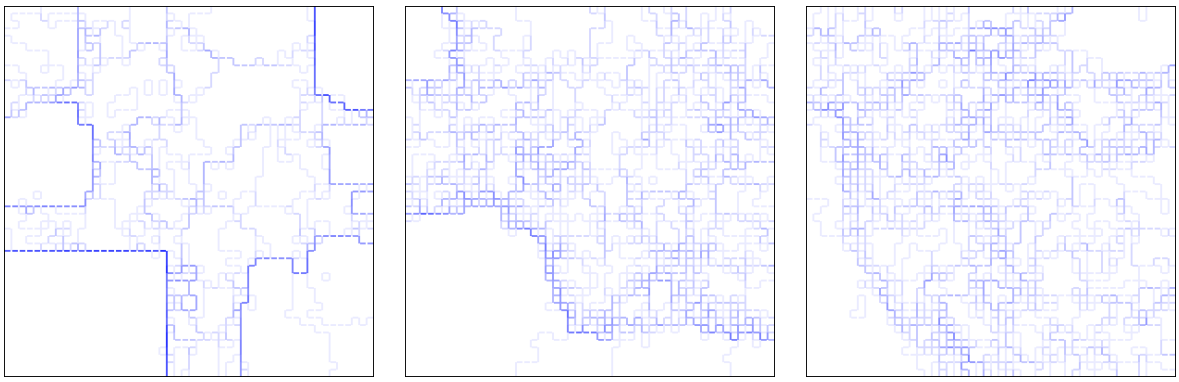
Population size was set to 2500 individuals, implying a space size of $50 \cdot 50$ points. The feature space size was key constant to 16 binary possibilities.

Increasing the minimal interaction probability α resulted in less contained groups and a more freely flowing language continuum, as can be seen in figures 2(a) and 2(b). Increasing the drop-off speed actually decreases the amount of language fading, and increases the amount of language borders, as can be seen in figures 2(c) and 2(d).

A higher minimal adoption probability unsurprisingly increases the amount of variety in the system. The slope of the adoptation probability is reflected in the 'sharpness' of the borders; a steeper slope will result in clearer borders, as can be seen in figures 2(e) and 2(f).

To test the usefulness of the adoption probability (and thus, the social impact theory), a null-model was devised. In this case, the adoption probability function was discarded in favor of a constant value. This model was tested under a wide variety of parameters, but never was the model able to maintain variety. This can be explained by the same experience Axelrod (1997) had. He calls the effect 'random walk with absorbing barriers', which means that in a situation where there is enough room to mutate, the less used mutation will probably be absorbed by the more frequent one.

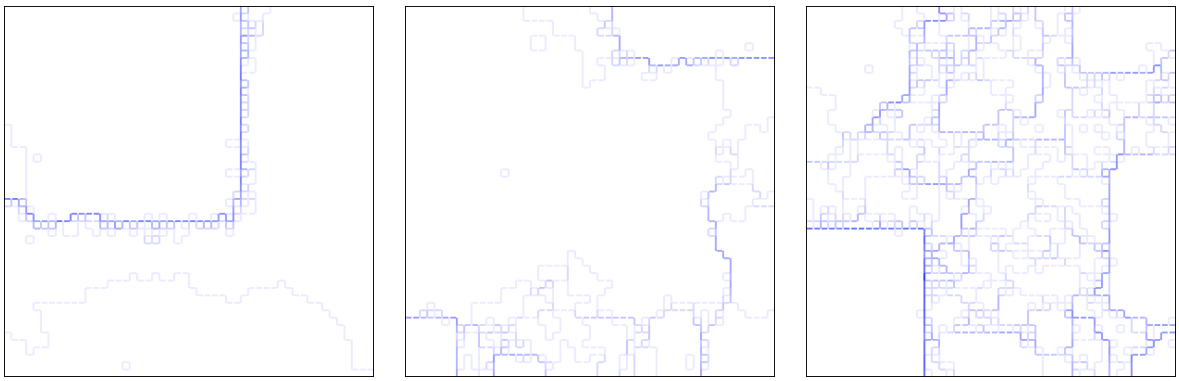
Surprisingly, the agent death had a big influence on the model. Disabling agent death and birth, and always allowing agents to learn from each other, the results changed dramatically. The model was not able to generate any kind of language border. Neither is there any form of group creation. This can be explained by the fact that agent death has a stabilizing function in the system, by temporarily introducing non-changing agents, as well as allowing the model to converge, by introducing new agents that speak the same language as its immediate neighbors. Figure 3 shows cases of the experiment without agent death, for both a high and a low mutation introduction chance. As can be seen, under both conditions there is no grouping.



(a) Low minimal interaction probability ($\alpha = 0.1, \beta = 3$)

(b) High minimal interaction probability ($\alpha = 0.5, \beta = 3$)

(c) Flat interaction probability ($\alpha = 0.4, \beta = 1$)



(d) Steep interaction probability ($\alpha = 0.4, \beta = 11$)

(e) Flat adoption probability ($\gamma = 0.5, \kappa = 0.3$)

(f) Steep adoption probability ($\gamma = 0.5, \kappa = 0.5$)

Figure 2: Comparison of conditions under which realistic diversity emerges. Shown are the agent spaces (50x50 agents). Dark lines indicate large distances between agents languages, light lines small distances. The middle frame shows a simulation with both preference for rare linguistic features and agent birth and death. Note the emergence of both language borders and dialect continua. The left frame shows a simulation without a preference for rare features, and the right frame a simulation without agent birth and death. Only uninteresting variation emerges.

4 Conclusion

Given the right parameter settings, this model results in linguistic (or cultural, depending on how one would interpret the feature strings) diversity that shows both dialect continua and language borders. Dialect continua appear almost for all parameter settings, but language borders only appear when the probability of accepting changes drops sufficiently quickly with the frequency of these changes. In other words agents must have a preference for rare utterances.

Interestingly, it also appears necessary that there is a flux of agents in the population. If agents do not die and are replaced with younger languageless agents, interesting linguistic diversity does not develop. This is an interesting result, which cannot be compared to earlier models: Nettle (1999b,a); Livingstone (2002) all used agent death in their models, but failed to include data on the case without death.

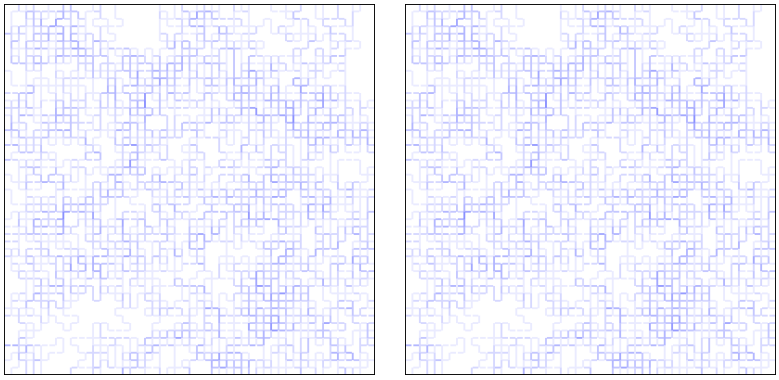


Figure 3: Results of experiment runs without agent death.

Although the analysis of these results is still in progress, and more experiments are necessary to clarify the exact circumstances under which diversity emerges in the model, they nevertheless show that under realistic conditions diversity can emerge without barriers and without unrealistic assumptions about social influence of single individuals.

References

Robert Axelrod. The dissemination of culture: A model with local convergence and global polarization. *The Journal of Conflict Resolution*, 41(2):203–226, apr 1997. ISSN 0022-0027.

Dale J. Barr. Establishing conventional communication systems: Is common knowledge necessary? *Cognitive Science*, 28(6):937–962, November-December 2004. doi: 10.1016/j.cogsci.2004.07.002.

B. Latané. The psychology of social impact. *American Psychologist*, 36(4):343–356, 1981.

Daniel Livingstone. The evolution of dialect diversity. In Angelo Cangelosi and Domenico Parisi, editors, *Simulating the Evolution of Language*, chapter 5, pages 99–118. Springer Verlag, London, 2002.

Makoto Matsumoto and Takuji Nishimura. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 8(1):3–30, 1998.

D. Nettle. *Linguistic Diversity*. Oxford University Press, 1999a.

Daniel Nettle. Using social impact theory to simulate language change. *Lingua*, 108(2-3): 95–117, June 1999b. doi: 10.1016/S0024-3841(98)00046-1.

Marco Patriarca and Teemu Leppanen. Modeling language competition. *Physica A: Statistical Mechanics and its Applications*, 338(1–2):296–299, July 2004. doi: 10.1016/j.physa.2004.02.056.

S.F. Railsback, S.L. Lytinen, S.K. Jackson, and J.S. Computing. Agent-based Simulation Platforms: Review and Development Recommendations. *Simulation*, 82(9):609, 2006.