A Bottom-up approach to the cultural evolution of bilingualism

Seán G. Roberts (sean.roberts@mpi.nl)

Language and Cognition Department, Max Planck Institute for Psycholinguistics

Wundtlaan 1, Nijmegen 6525 XD The Netherlands

Abstract

The relationship between individual cognition and cultural phenomena at the society level can be transformed by cultural transmission (Kirby, Dowman, & Griffiths, 2007). Top-down models of this process have typically assumed that individuals only adopt a single linguistic trait. Recent extensions include 'bilingual' agents, able to adopt multiple linguistic traits (Burkett & Griffiths, 2010). However, bilingualism is more than variation within an individual: it involves the conditional use of variation with different interlocutors. That is, bilingualism is a property of a population that emerges from use. A bottom-up simulation is presented where learners are sensitive to the identity of other speakers. The simulation reveals that dynamic social structures are a key factor for the evolution of bilingualism in a population, a feature that was abstracted away in the top-down models. Top-down and bottom-up approaches may lead to different answers, but can work together to reveal and explore important features of the cultural transmission process.

Keywords: Language; Cultural evolution; Bilingualism; Bottom-up.

Introduction

Bilingualism is prevalent in many communities, but this is, intuitively, an unstable situation. What drives the emergence of bilingualism? Previous top-down computational models have emphasises the role of the learning biases of individuals such as a language's prestige (Abrams & Strogatz, 2003), or expectations about variation in the input (Smith, 2009; Burkett & Griffiths, 2010; Smith & Thompson, 2012). However, these models assume that languages are monolithic, discrete and static. A bilingual is defined as an individual that adopts more than one linguistic variant from a set of discrete languages that has been defined *a priori*. This paper questions this assumption and explores how differences arise between the linguistic codes of communities in the first place and how they are subsequently adopted and maintained. In doing so, reveals that social factors, such as cultural identity and the dynamics of population structure, are crucial to the process of cultural evolution.

While low-level linguistic variants such as word order may plausibly have a discrete psychological reality (Diamond, 1991), 'languages' are more complex entities. Dividing continuous linguistic variation into categorical units or distinguishing 'languages' from 'dialects' is not straightforward and often involves complex notions such as politics, history, geography and identity (Sober, 1980; Haugen, 2009; Croft, in press; S. Roberts, 2012). The features of a language and the way it contrasts with others also changes over time. Therefore, this paper assumes that monolithic, static 'languages' (e.g. English, Welsh) have no psychological reality, but are emergent properties of populations and use. An abstract definition of bilingualism is presented which does not require

discrete categories: bilingualism is the amount of linguistic variation that is conditioned on social variables. That is, if I speak differently to Mary than to John, then I'm bilingual to some extent. This definition is compatible with the notion of audience design in sociolinguistics (Bell, 1984), and identifies bilingualism as a gradient property of interaction rather than a categorical feature that is identifiable in an isolated individual.

In order to explore the cultural evolution of bilingualism in this way, a bottom-up simulation is presented where learners are embedded in dynamic social structures and are sensitive to the identity of speakers. The top-down models assumed that social structures were static and focussed on cognitive explanations. In contrast, the bottom-up simulation will demonstrate that changes to social structures are an important factor. Rather than arguing that the bottom-up simulation is 'better' than the top-down model, this paper argues that different modelling tools tend to bias researchers towards making certain kinds of assumptions.

Simulation definition

Bilingual cultural transmission is simulated as iterated stepwise linear regression. The representation of language is highly abstract, but allows the simulation of the emergence of bilingualism in a complex social structure. The linguistic space is a continuous one-dimensional space. A linguistic utterance is a point on this space (a real number). A meaning is represented as a point in a multi-dimensional meaning space. Each dimension of the meaning space represents a different semantic variable, such as properties of the environment that a linguistic utterance might be referring to (e.g. colour, number, size, tense etc.). A point in the meaning space, then, represents a particular combination of semantic elements (e.g. one big yellow thing). Each semantic variable has a set of hidden parameters which describe the distribution from which values are sampled. This systematic variation ensures that the linguistic signal has some structure to emulate. An important exception s a semantic variable that represents the identity of the speaker who described the event (speaker ID).

A 'learner' observes 'teachers' describing meanings with utterances and learns a mapping between the linguistic signal and the meaning space. The learning mechanism is a linear regression which results in an *linguistic model* that maps points in the meaning space to points in the linguistic space. The learner can then use this model to produce linguistic utterances itself.

Real languages exhibit flexibility with regards to which aspects of meaning condition linguistic variation. For example, in French, the form of a demonstrative (*ce,cette,ces*) depends
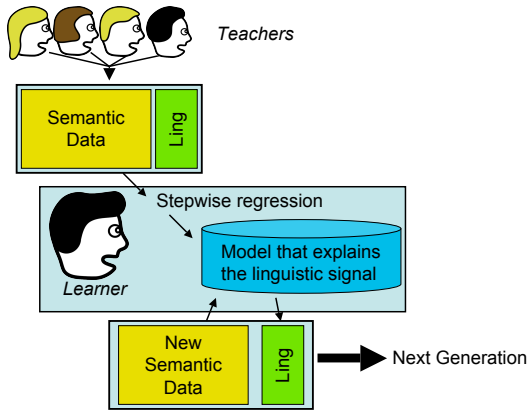
Figure 1: A diagram of how the bottom-up simulation works. Teachers produce linguistic and semantic data for a learner to observe. The learner uses stepwise regression to build a linguistic model, which is then used to describe some new semantic variables for the next generation.

on the grammatical gender of the referent and whether it is singular or plural, while the distance of the referent from the speaker is not important. In contrast, in Panjabi, the form of a demonstrative (*ih,uh*) is conditioned on distance from the speaker, but not gender or number. In order to capture this flexibility, the linguistic model is selected by stepwise linear regression. A stepwise regression selects the minimum number of salient (semantic) variables that maximises the statistical fit, according to an information criterion. This allows the learner's linguistic model to prioritise or ignore different semantic variables in its linguistic utterances, including the identity of the speaker.

This learning process is iterated (Smith, Kirby, & Brighton, 2003) in the following way. Learners observe a number of 'teachers' describing points in the meaning space with linguistic utterances, as above. The observations are affected by a small amount of noise. The probability of observing an utterance from a particular speaker depends on the structure of the community (see below). After the learners induce a linguistic model, the teachers are removed from the population. The learners become teachers for a new generation that is added to the population. This process repeats for many generations.

Learners have no explicit biases for particular 'languages', and the speaker ID variable is not privileged over other aspects of meaning. However, the learner is biased in a general way by the information criterion for the stepwise regression *IC*, which affects the number of variables an individual is willing to include in their linguistic model. The results presented below use the Akaike information criterion ($IC = 2$). In general, iterated stepwise linear regression has a bias towards shallow slopes and small intercepts. These do not affect the general results regarding bilingualism.

Population parameters Generations of individuals are separated by discrete timesteps $t_1, t_2 ... t_n$. A population of P_t learners in the current generation observe data produced by P_{t-1} teachers in the previous generation. There are a number of communities in each generation and a set $C(t)$ of P_t discrete labels represents which community each individual belongs to. A community interaction matrix $I(t)$ defines how much contact there is between each community. The probability $W(t)_{i,j}$ of learner i receiving data from teacher j is calculated as:

$$W(t)_{i,j} = \frac{I(t)_{C(t)_i, C(t-1)_j}}{Sum_W} \tag{1}$$

Where Sum_W is the sum of all weights between individuals. The community structure can therefore reflect situations from simple ones such as 'there are two communities' to a weighted, directed graph between individuals.

The community interaction matrix $I(t)$ can be simplified to a vector of single numbers by assuming that the probability of receiving data from any community that a learner does not belong to is equal.

$$W(t)_{i,j} = \begin{cases} \frac{I(t)_i}{Sum_W} & \text{if } C(t)_i = C(t-1)_j \\ \frac{1 - I(t)_i}{Sum_W} & \text{otherwise} \end{cases} \tag{2}$$

This assumption will be adequate for the examples in this paper, and allows manipulation of the social structure through a single parameter for each community.

This framework allows different types of social structures. Given a situation where there are two teachers and two learners $P_{t-1} = P_t = 2$ and two communities at each generation $C(t) = C(t-1) = \{A, B\}$, different settings of I can then result in many social dynamics. Below I give some examples of matrices, with the learners (rows) labelled as L_1 and L_2 and the teachers (columns) labelled as T_1 and T_2. For example, a society with two communities that are completely integrated and balanced (effectively a single community):

$$I(t) = \{0.5, 0.5\} \longrightarrow \begin{array}{c c c} & T_1 & T_2 \\ L_1 & 0.5 & 0.5 \\ L_2 & 0.5 & 0.5 \end{array} \tag{3}$$

In the matrix above, for example, learner 1 (L_1) has an equal probability of receiving data from either teacher. Alternatively, L_1 only receives data from T_1 and L_2 only receives data from T_2. This simulates two communities that are completely isolated:

$$I(t) = \{1, 1\} \longrightarrow \begin{array}{c c c} & T_1 & T_2 \\ L_1 & 1 & 0 \\ L_2 & 0 & 1 \end{array} \tag{4}$$

The prestige of a community can also be manipulated. Below is a matrix for a situation where one community only receives input from its own members (superstrate), while the other community receives input from both communities (substrate). It is predicted that this will lead to an analogue of a

minority language situation where everyone speaks one language (the majority language), and some speak a second language (the minority language).:

$$I(t) = \{0.5, 1\} \longrightarrow \begin{array}{c|cc} & T_1 & T_2 \\ \hline L_1 & 0.5 & 0.5 \\ L_2 & 0 & 1 \end{array} \qquad (5)$$

Measuring bilingualism Since 'languages' are not encoded in the simulation, the amount of bilingualism must be calculated from the bottom-up. This is defined as the amount of linguistic variation that is conditioned on social variables. In the simulation this is based on two measures of intelligibility, assuming that utterances are intelligible to speakers if their linguistic model would produce the same utterance given the same meaning (obverter assumption, similar to Oliphant & Batali, 1997). The first is a measure of **comprehensive intelligibility**: the proportion of utterances that one speaker typically produces that another understands. For example, a monolingual speaker of English understands half of the utterances spoken by a balanced bilingual speaker of English and Welsh. In the simulation, this is a measure of the proportion of the variance in one learner's productions that is explained by another learner's linguistic model. If we're comparing individual A and B, this is implemented in the following way (see figure 2):

1) Take identical samples of meanings M_A and M_B
2) Sample speaker ID evenly in M_A and M_B
3) Given M_A, produce utterances U_A with A's linguistic model and given M_B, produce utterances U_B with B's linguistic model
4) Calculate the correlation between U_A and U_B

If two learners have the same linguistic model, then this correlation should be high. Individual A with a very different model from individual B will produce linguistic signals with a variation that is poorly explained by learner B's model, and so the correlation will be low.

We can also define a **functional intelligibility** score which is the proportion of utterances that interlocutors understand when they design their utterances for each other (figure 3). That is, a bilingual speaker of English and Welsh and a monolingual speaker of English could always make themselves understood by using English. In the simulation, this is calculated in a similar way to the comprehensive intelligibility score, except step (2) above is changed to

2) Set the speaker ID in M_A to B and in M_B to A

In this case, an individual with a linguistic model that used speaker identity as a conditioning factor would adjust its variation to better suit its receiver (i.e. in the Welsh-English example, by speaking only English).

The two intelligibility measures can be combined to get a measure of bilingualism by subtracting the comprehensive intelligibility from the functional intelligibility. This can be calculated for a population by taking the mean bilingualism score for all pairs of speakers.

$$B(t) = \frac{2}{n^2 - n} \sum_{i=0}^{P_t - 1} \sum_{j=i+1}^{P_t} Func(i,j) - Comp(i,j) \qquad (6)$$

Where $Func(i,j)$ and $Comp(i,j)$ calculate the functional and comprehensive intelligibility between two speakers.

If $B(t) \approx 0$, the comprehensive and functional intelligibility are, on average, similar. This means that everyone shares the same mapping between linguistic utterances and meanings - what might be called the same linguistic code or 'medium' (Gafaranga, 2008). While this code might exhibit variation, this could be interpreted as a single 'language' (monolingualism).

If $B(t) < 0$, the functional intelligibility between speakers is, on average, lower than the comprehensive intelligibility. For example, in the functional measure, speaker A would adapt their linguistic signal for speaker B and B would adapt their linguistic signal for speaker A. This yields a low functional similarity. However, their comprehensive similarity is high (their overall linguistic system is similar), so $B(t)$ is negative. This would be interpreted as bilingualism in the sense that each community is associated with a different mapping or 'code', and individuals can use each others' codes to some extent. A lower $B(t)$ means 'more' bilingualism in the lay sense. This is meant to represent the amount of linguistic variation that is conditioned on social variables, and so is analogous to an entropy-like measure where lower values indicate more order (the linguistic system is more conditioned on social factors = bilingualism) and higher values indicate more disorder (the linguistic system is less conditioned by social factors = monolingualism).

If $B(t) > 0$, the comprehensive intelligibility score is lower than the functional intelligibility score. For example, A adapts their linguistic signal for B, but B does not adapt their linguistic signal for A. This leads to a high functional intelligibility, but a low comprehensive intelligibility. This means that both communities share one code, but one community has at least one other code. This might be interpreted as a minority situation in which one community speaks a minority language as well as the majority language. As we well see below, it's useful to be able to distinguish between 'balanced' communities ($B(t) < 0$) and 'minority' situations ($B(t) > 0$).

Results

Since simulation is complex, basic findings are presented for simulations with 2 communities with 2 individuals each and 2 semantic variables, but the results scale up many semantic variables and hundreds of individuals. To summarise: unconditioned variation is unstable and bilingualism tracks social change. Figure 4 shows how B changes in different social structures. When the two communities are completely integrated (integration parameter $I = 0.5$), then they quickly converge to using the same linguistic code ($B = 0$). When the
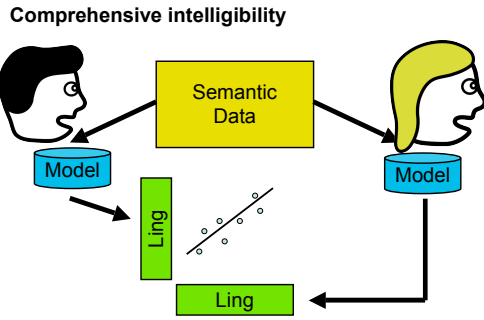
Comprehensive intelligibility

Figure 2: The comprehensive intelligibility measure. Two individuals are given the same meanings and produce linguistic utterances with their linguistic models. The correlation between these utterances in the linguistic space is measured.
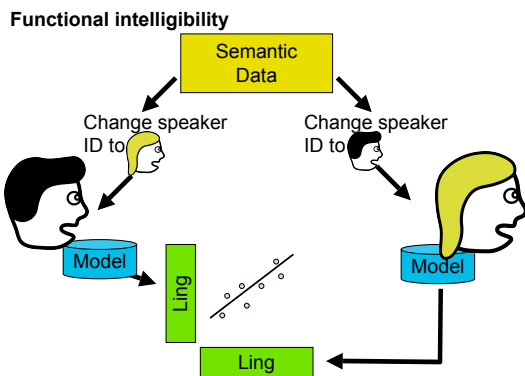

Functional intelligibility

Figure 3: The functional intelligibility measure. Two individuals are given the same meanings, but the speaker ID is set to the other individual in the pair. They produce linguistic utterances with their linguistic models, and the correlation between the utterances in the linguistic space is measured.

two communities are partially isolated ($I = 0.8$), their varieties will take longer to converge and 'bilingualism' ($B < 0$) persists for some generations. The results are slightly different in a substrate/superstrate situation where learners from one community receive input equally from both communities (the minority, $I = 0.5$), but the other community mainly receives input from speakers from its own community (the majority, $I = 0.9999$). In this case, B remains positive for many generations (a 'minority' language situation).

These results are for communities with static social structures. We can manipulate the social structure to demonstrate that linguistic diversity also tracks the *change* in social structures. Figure 5 shows the results of simulations with dynamic social structures. The communities go through cycles of being integrated, isolated, integrated and isolated again, with a few transition generations between each phase where the integration parameter is interpolated. As shown above, if two communities are integrated, they will come to speak effectively a single code ($B \approx 0$, see figure 4). However, as the
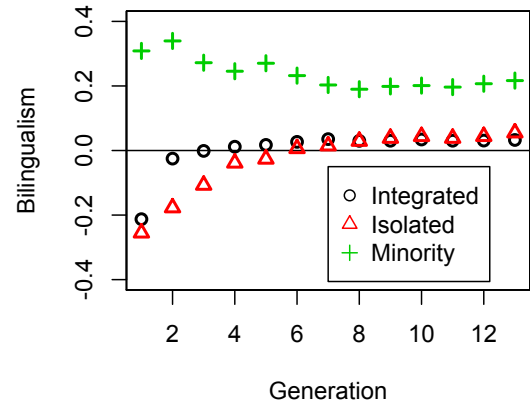

Figure 4: The bilingualism score (means for 100 simulations) over generations for 2 communities of 2 individuals in three social structures: Integrated (circles), Isolated (triangles) and Minority (crosses).

communities become more isolated, B increases. This is also in line with the results above. However, as the communities increase their interactions after this, B decreases (everyone speaks a single code). Then we can split the communities apart and two codes will emerge again with some amount of bilingualism. That is, the distribution of linguistic variation tracks the changes in social structure.

More complex factors that affect B were determined by analysing many runs of the simulation (analysis done using linear regression and stepwise linear regression). $B < 0$ is inherently unstable in this simulation. As soon as individuals start mutually accommodating the linguistic signal of other communities, this neutralises the distinction over speaker ID. This is in line with the expectation that unconditioned linguistic variation is unstable (e.g. Smith & Wonnacott, 2010).

$B < 0$ is much more likely to emerge if speaker identity is the most important conditioning factor, while positive bilingualism scores can emerge if speaker identity is less important. Negative bilingualism is also more likely if individuals rank speaker identity in their models similarly. There are some more complex interactions. For example, $B < 0$ tends to emerge when: the speaker ID is more important in the previous generation, except when communities are diverging, when it can be higher; when the community with the most complex linguistic model also considers speaker ID to be less important; when the mean and standard deviations of the speaker id rank are correlated; and when there is a stronger correlation between the difference in linguistic signal means and model fit ratio between communities.

Figure 5 shows that, after the first contact situation, only $B > 0$ tends to emerge. This is partly due to the linguistic signal of two communities adapting to the same semantic distributions, and so becoming more alike. Situations where $B < 0$ requires that there are large differences in the utterances of each community so that speaker ID conditions a large amount
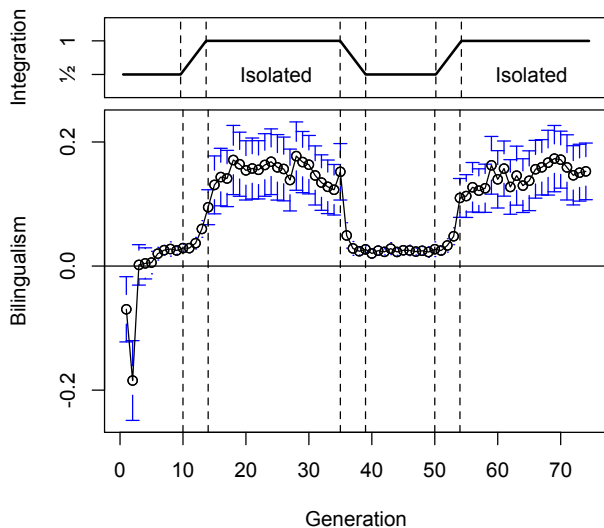
Figure 5: How the integration parameter $I(t)$ (top, identical for both communities) affects the levels of bilingualism (bottom) in 2 communities with 2 learners each, which are alternatively integrated and semi-isolated (means over 100 runs with 95% confidence intervals).

of variation. When $B > 0$ there is an imbalance in the extent to which different communities adapt to each other's linguistic signal.

It is possible to identify a 'superstrate' community as the one whose linguistic signal changes least between the generations of contact (as measured by the difference in a community's comprehensive intelligibility between generations). The difference in the linguistic utterance means between generations is the main determiner of the superstrate community. If community X's mean is absolutely greater than community Y's mean in the previous generation, then community X's linguistic models will change more than community Y. This affect arises due to the bias in the learning mechanism for small intercepts. However, this trend is only strong in the first generation of contact. During diverging generations, there is a 41% chance of a switch in superstrate community in the first two generations of divergence (from 100 simulations, significantly different from no switch: t = 16.7, df = 399, p-value < 0.001, but also random switching: t = -3.55, df = 399, p < 0.001). In contact situations, there is a 49% chance change of a switch in superstrate community in the first two generations (from 100 simulations, significantly different from no switch: t = 13.8, df = 199, p < 0.001; but not significantly different from random switching: t = -0.28, df = 199, p = 0.78). In one generation a community might adapt to another, but this can cause the models in that community to better fit the data, leading to a pressure for the other community to adapt in the subsequent generation. Although a preliminary result, this may be compatible with phenomena such as 'mixed languages' where the emerging language in a contact situation uses the lexicon of one source language, but the grammar and morphology of the other (e.g. Muysken, 1997). If lexical items and morphology take different amounts of time to learn (as suggested by Clahsen, Felser, Neubauer, Sato, & Silva, 2010), then the 'mixing' might be partially due to this alternation in the community that adapts: the lexicon is taken first from one language, and later the morphology from another.

Discussion

Dynamic social structures are a key aspect for explaining the emergence of bilingualism in this simulation. In the top-down models, social structures were static and so they could not form a part of the explanation. The bottom-up simulation can be more flexible because it doesn't require learners to be fully rational or optimal, as opposed to some Bayesian models.

The linguistic contrast between communities will diminish if there is no contrast in the social variables. However, it does not mean than bilingualism in the lay sense is unstable. Firstly, B is not necessarily an index of an intuitive idea of bilingualism. Communities like those in Catalonia might actually have $B \approx 0$, because many people speak both Catalan and Spanish. Secondly, in the real world, linguistic variation might be dictated by social factors not simulated here, such as location, formality or stage of the conversation (e.g. Labov, 1963; Meyerhoff, 2008). Finally, this simulation includes no pressures to maintain a linguistic identity such as prestige, politics or resistance to freeriders (G. Roberts, 2010). Rather, it shows that bilingualism that can emerge just from the process of cultural transmission - a kind of baseline behaviour on top of which more complex factors may be applied.

The top-down models specified a prior bias over the amount of variation to expect in an agent's input, fitting the learning mechanism to the problem being addressed. In contrast, bilingualism emerges in the bottom-up simulation without individuals having a specific mechanism for dealing with bilingualism. All that is required is a general learning mechanism which conditions a linguistic signal on semantic variables. There are no expectations over the amount of variation to expect within or between speakers. Indeed, if social variables do not explain any of the variance, they do not play any role in an agent's linguistic internal representation.

Furthermore, the simulation maintains a division between population level phenomena and individual learning mechanisms: 'bilingualism' can emerge at the population level without discrete, static languages being encoded in the linguistic model of individuals. This suggests that that 'bilingualism' is a property of populations which is not necessarily related to specific individual learning biases. That is, whether humans have an expectation about the number of languages that will be in their input, or whether learning two languages is more difficult than learning one are not necessarily the most relevant questions. Rather, one should ask how contrasts in social variables support the maintenance of linguistic variation.

Conclusion

The simulation works as a proof-of-concept for the abstract definition of bilingualism. Bilingualism is measurable in this simulation without encoding a discrete, monolithic, static concept of a 'language'. The measure behaves as we would expect in integrated, isolated and substrate/superstrate population structures. The results suggest that dynamic social structures are an important part of explaining the cultural evolution of bilingualism. This differs from the conclusions of top-down models, demonstrating that different kinds of models may be biased towards certain conclusions.

Top-down models considered the stability of bilingualism given assumptions about individual learning (Burkett & Griffiths, 2010) and the most likely expectation for individuals to have about the number of languages in their input (Smith & Thompson, 2012). These might suggest research directions such as estimating the expectations human learners have about the number of languages to expect in their input, the amount of noise in transmission or whether the social structure was one that caused bottlenecks on learning. However, in the bottom-up simulation, because bilingualism tracks social change, asking whether individuals should expect many languages in their input does not make sense without also thinking about *dynamic* social structures. This suggests that the questions asked by the top-down model are misleading. The bottom-up simulation suggests researching dynamic social structures, and how linguistic variation, social structures and learning biases coevolve.

Both the top-down and bottom-up models are very abstract, and it would be a difficult to determine which was more 'realistic' or fitted real data better. Instead, both approaches can be seen as converging on a common solution to the problem from different angles. The top-down model is better at yielding good analytic results, but the bottom-up model allows more flexibility in terms of social dynamics. The bottom-up simulation presented here has suggested that some of the assumptions of the top-down models require more scrutiny. In response, a top-down model could be built which addressed the most relevant points raised by the bottom-up simulation perhaps using techniques such as variational Bayesian analysis (e.g. Kurihara & Sato, 2006). This process of exploring results and uncovering important assumptions using mutually supporting approaches can lead to more robust theories.

Acknowledgments

References

Abrams, D. M., & Strogatz, S. H. (2003). Modelling the dynamics of language death. *Nature*, *424*, 900.

Bell, A. (1984). Language style as audience design. *Language in society*, *13*(2), 145–204.

Burkett, D., & Griffiths, T. (2010). Iterated learning of multiple languages from multiple teachers. In A. Smith, M. Schouwstra, B. de Boer, & K. Smith (Eds.), *The evolution of language: Proceedings of evolang 2010* (p. 58-65).

Clahsen, H., Felser, C., Neubauer, K., Sato, M., & Silva, R. (2010). Morphological structure in native and nonnative language processing. *Language Learning*, *60*(1), 21–43.

Croft, W. (in press). Language use and the evolution of languages. In K. Smith & P. Binder (Eds.), *The language phenomenon.* Berlin: Springer.

Diamond, J. (1991). *The rise and fall of the third chimpanzee.* Radius, London.

Gafaranga, J. (2008). *Talk in two languages.* Palgrave Macmillan.

Haugen, E. (2009). Dialect, language, nation. *American anthropologist*, *68*(4), 922–935.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*(12), 5241-5245.

Kurihara, K., & Sato, T. (2006). Variational bayesian grammar induction for natural language. *Grammatical Inference: Algorithms and Applications*, 84–96.

Labov, W. (1963). The social motivation of a sound change. *Word*, *19*, 273-309.

Meyerhoff, M. (2008). Bequia sweet/bequia is sweet: syntactic variation in a lesser-known variety of caribbean english. *English Today*, *24*(01), 33–40.

Muysken, P. (1997). Media lengua. *Contact languages: a wider perspective*, *17*, 365.

Oliphant, M., & Batali, J. (1997). Learning and the emergence of coordinated communication. *Center for research on language newsletter*, *11*(1), 1–46.

Roberts, G. (2010). An experimental study of social selection and frequency of interaction in linguistic diversity. *Interaction Studies*, *11*(1), 138–59.

Roberts, S. (2012). *Evolutionary approaches to bilingualism.* Unpublished doctoral dissertation, University of Edinburgh.

Smith, K. (2009). Iterated learning in populations of Bayesian agents. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society.* Cognitive Science Society.

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: a framework for the emergence of language. *Artificial Life*, *9*(4), 371-386.

Smith, K., & Thompson, B. (2012). Iterated learning in populations: Learning and evolving expectations about linguistic homogeneity. In T. C. Scott-Phillips, M. Tamariz, E. A. Cartmill, & J. R. Hurford (Eds.), *The Evolution of Language: Proceedings of the 9th International Conference (EVOLANG9)* (p. 227-233). World Scientific.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*(3), 444–449.

Sober, E. (1980). Language and psychological reality: some reflections on chomsky's rules and representations. *Linguistics and Philosophy*, *3*(3), 395–405.