A model of linguistic accommodation leading to language simplification

Stella Frank (stella.frank@ed.ac.uk) & Kenny Smith
Centre for Language Evolution, School of Philosophy, Psychology & Language Sciences,
University of Edinburgh, Dugald Stewart Building, 3 Charles Street, Edinburgh, EH8 9AD, UK

Abstract

Language complexity seems to be influenced by population characteristics such as the proportion of adult learners. One potential explanation for this link is that native speakers accommodate to non-native speakers, simplifying their language use during such interactions: learners may then acquire a less complex language. We model accommodation in interaction in a Bayesian framework, where in order to accommodate appropriately, an agent must first infer their interlocutor's linguistic abilities. We find that when the agent consistently accommodates, learners end up with a simplified language, due to a reinforcing effect between an initially underinformed learner and an accommodating native speaker.

Keywords: language evolution; language complexity; Bayesian models; interaction models

Introduction

Linguistic communication requires a shared language. However, no two speakers have had exactly the same linguistic experiences and thus no two speakers will have exactly the same language: they may differ at all linguistic levels, e.g. in their pronunciation, in the words they know, and in the grammatical constructions they use. Interaction between individuals therefore involves making inferences about the linguistic system of one's interlocutor, since knowledge of an interlocutor's linguistic system is required in order to accurately understand their intended meaning.

Listeners construct such a *speaker model* of their interlocutor, for example, to to correctly interpret UK/US ambiguous words (*flat*, *pants*) depending on the accent in which the words are spoken (Cai et al., 2017; Martin, Garcia, Potter, Melinger, & Costa, 2015); or to interpret entrained/dialogue-specific words in a speaker-specific way (Brown-Schmidt, 2009). More generally, interaction requires us to make potentially complex inferences based on the nested and interlocking linguistic and social groups our interlocutors belong to (e.g. cognitive scientist, Scot, English speaker, poor French speaker), because those affiliations determine how our interlocutors use and understand language (Clark, 1996, 1998).

We construct a Bayesian model of linguistic interaction between two individuals capable of making these sorts of inferences about their interlocutors' likely linguistic knowledge. We focus on a particularly asymmetric scenario, namely a conversation between two individuals with vastly different language experience, as would arise in an interaction between a native- and a non-native speaker. In this scenario, the native speaker is aware of having greater experience with the language than the non-native speaker, and is consequently aware that their interlocutor may have a different linguistic system; the native speaker builds a model of the likely linguistic system of their non-native interlocutor, which allows them to accommodate to the inferred language abilities of the non-native speaker, adjusting their linguistic output to match (their beliefs about) their partner's linguistic system (Giles, Coupland, & Coupland, 1991).

Asymmetric interactions of this type result in linguistic registers such as Child-Directed Speech and Foreigner Talk (Snow & Ferguson, 1977; Ferguson, 1975). Rather than modelling these registers for their own sake, our motivation for studying this scenario arises from our interest in the effects of interaction on the *complexity* of linguistic systems. Natural languages differ in the amount of complexity they encode within a domain (e.g. number of case markings, noun classes; size of phoneme inventories). Various authors have suggested that this kind of variation in complexity may reflect systematic differences in the size or composition of the populations in which those languages are spoken (e.g. Wray & Grace, 2007; Lupyan & Dale, 2010; Trudgill, 2011; Bentz & Winter, 2013). In the example of case, Bentz and Winter (2013) show across a sample of 66 languages that languages with more non-native speakers tend to have fewer distinct cases, and all languages with more than 50% non-native speakers have no case system at all; Lupyan and Dale (2010) suggest a similar link between the prevalence of non-native speakers and morphological simplicity more broadly.

Why would languages with more non-native speakers be simpler? Bentz and Winter (2013) speculate that native speaker accommodation to non-native interlocutors might act as a crucial linking mechanism: if non-native speakers tend to imperfectly acquire the language's case system (e.g. due to insufficient exposure), native speaker accommodation to this feature of non-native speech will result in such simplifications being more widely used, more frequent in the input available to language learners, and ultimately leading to a change in the language's grammar.

There is some experimental evidence that accommodation during interaction between individuals with asymmetric linguistic knowledge can lead to this kind of simplification. For instance, Atkinson, Smith, and Kirby (submitted) show that when two experimental participants trained on artificial languages which differ in their complexity (e.g. where one participant is trained on the full language, and their partner is trained on a simpler language which lacks irregulars) then the pair preferentially align on the simpler language: the first participant is willing to move away from their own language to accommodate the other.

While this experimental evidence is consistent with the hypothesis (investigated in this paper) that accommodation can result in simplification, which might then spread through a population, the mechanisms at play are at present unclear:

inferences about a partner's linguistic knowledge, but also reciprocal priming between interlocutors and feedback-based reinforcement learning may all be involved. The model we present here allows us to explore the role of inferences about a partner's linguistic knowledge in a targeted way, in order to identify whether and when it results in linguistic simplification. Consistent with the accounts offered above, we demonstrate that accommodation by native speakers can lead to simplification during interaction, which in turn results in non-native speakers learning a language that is more regular than the original native speaker's language, and therefore can result in a net loss in linguistic complexity.

Model Framework

Our model and interactive setup is related to Bayesian models of pragmatics (Franke & Jäger, 2016), such as the Rational Speech Act model (Frank & Goodman, 2012), in which agents construct (probabilistic) representations of the beliefs of others about the world based on those agent's utterances. Our methodology also follows modelling work on language evolution, e.g. Reali and Griffiths (2009); Smith et al. (2017) where a simple language model is used to illustrate the process of *regularisation*, in which variation in a linguistic system is reduced, through learning, transmission or interaction, resulting in a simplification the system. The Bayesian agents in the above models behave like our basic learner agent, inferring a language in interaction with other agents. We add internal speaker-modelling of the other agents, i.e. agents possess a meta-level 'theory of language' (analogous to 'theory of mind') of their interlocutors' linguistic abilities.

In our model, two agents (named A and B) interact using a common language L. Prior to interaction, A and B have learned L from different exposures, leading to variation in their individual languages. As A and B see more data, they will, in the long run, converge to the common language. In the shorter term, however, divergence between the agents can be severe. In the scenario investigated here, A is the 'native' speaker, who has seen enough data to learn the language accurately. We designate B as the 'non-native' speaker who has had only limited exposure to L before their interaction with A; we show how this will lead to B having a language that is more regular, with a more peaked distribution towards a single variant. In conversation with B, as a rational interlocutor A should take B's linguistic abilities into account, despite not having direct access to B's internal language. Instead, A infers a *speaker model* over B's presumptive language.

The language that the agents are learning is represented as a probability distribution over variants, e.g. a set of ways to refer to an object's syntactic role, such as case marking or lexical strategies. Agents learn the language by inferring the parameters for the distribution (i.e. the probability of each variant) from the data they're exposed to, and speak by drawing a variant from their inferred distribution.

The regularity of a language $L = [p(x_0), p(x_1), \dots p(x_K)]$ is measured by the entropy of the language, $H(L)$:

$$H(L) = -\sum_{i=1}^{K} P(x_i) \log_2 P(x_i) \quad (1)$$

Languages in which all K variants have the same probability have the maximum entropy, $H = \log_2 K$. Languages with a highly likely variant have lower entropy, and the entropy of a single-variant language (where $P(x_i) = 1$ and $P(x_{j\neq i}) = 0$) is zero. The process of regularisation in a language can thus be captured as decreasing entropy.

Learning

Agents are modeled as rational learners who have internal representations of a language that they update based on what they hear, corresponding to learning.

Agents learn a distribution over possible languages using Bayes' rule to combine a prior, in the form of a distribution over possible languages indicating their prior beliefs $P(L)$, with the likelihood of the observed data $P(D|L)$: $P(L|D) \propto P(L)P(D|L)$. The specific prior used in this model corresponds to the agent's prior beliefs about the regularity of the language, i.e. the extent to which a single variant will have nearly all the probability mass, or whether multiple variants will have high probability. This takes the form of a symmetric Dirichlet distribution (hyper-)parameterised by a_0; if $a_0 > 1$, learners expect to hear all variants often (a flat distribution), whereas with a prior with $a_0 < 1$, they expect a more regular language with a dominant variant (a peaked distribution). In all our experiments, we set $a_0 = 0.01$. Note that the prior does not reveal *which* variant will be dominant, only that there is no expectation of multiple frequent variants.

The likelihood function is Categorical (i.e. the probability distribution over seen variants), for which the Dirichlet prior is conjugate, leading to a posterior distribution that is also a Dirichlet distribution, with hyperparameters updated by the seen counts. At the first update, the counts break the symmetry of the Dirichlet prior: the parameterisation of the posterior Dirichlet is now a vector $\boldsymbol{a}$, where $a_i = a_0 + c_i$, i.e. the previous prior value for the category i (in this case, a_0) plus the number of i-category items seen. In the interaction setting, learning is iterative, with agents updating after every utterance they hear. The predictive posterior probability of a word of category i is the normalised value of the updated hyperparameter, $p(x = i) = \frac{a_i}{\sum_j a_j}$.

In this paper, we denote an agent's language as L_A (indexed to speaker-agent A); this corresponds to their current posterior distribution over variants in the language, encoded by their updated hyperparameter $\boldsymbol{a}$ vector and based on their data exposure D_A. Agents speak by drawing samples from their language distribution $x \sim P(L_A|D_A, a_0)$.

All speakers have the same prior; consequently, differences in speakers' (distributions over) languages arise solely from exposure to different data. An agent who has not seen much data will be influenced more by the prior. Given a prior that prefers regular languages ($a_0 < 1$), an agent that has only seen

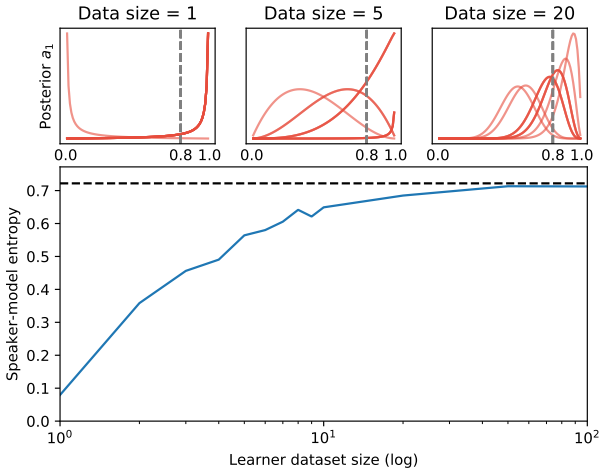

Figure 1: Speaker model inferences based on assumed learner dataset sizes. Top: (Hypothetical) learners who have seen little data have an extreme posterior, while with more data, the posteriors converge to the true value ($L^* = [0.8, 0.2]$, posterior probability of the first variant is plotted, equivalent to Beta-binomial). Below: Speaker-model entropy (calculated as the average entropy of the set of hypothetical learners) at different values of the size of the hypothetical learners' datasets. Speaker-modelling agents expect learners with little data exposure to have highly regular, low-entropy languages.

a few datapoints will infer a language that is highly skewed towards the seen variants. Additionally, the few datapoints are likely to over-represent the high-probability variants and under-represent the low-probability variants, a phenomenon known as *minority undersampling* (Hertwig, Barron, Weber, & Erev, 2004; Hahn, 2014). Together, this means that an agent with little experience (corresponding to the non-native speaker in our scenario) will initially have an overly-regular language, compared to the language data they are learning from. However, as they are exposed to more data, they will converge to the true (data-generating) language, in essence becoming a native speaker.

An Agent's Speaker Model

An agent constructs an internal *speaker model* of their interlocutor representing their inferences about their interlocutor's internal language (note that this is separate from their own language model). In our scenario of native- and non-native speaker interaction, the native speaker (A) constructs a speaker model of the non-native speaker (B), in order to accommodate to their language use.

To construct this speaker model, A assumes B is a learner of the same Bayesian form and with the same prior parameterisation a_0. A's aim is to discover B's posterior language L_B, which (given a known shared prior a_0) is dependent only on the data that B has seen, D_B. However, A does not know B's past history or exposure D_B, and thus cannot calculate L_B exactly. Instead, in our model, A guesses at B's possible past exposure; to limit the risk of guessing wrongly, A computes a mixture over several guesses Z and their corresponding posteriors:

$$P_A(L_B) = \sum_z^Z \omega_z P(L_z | D_z, a_0) \tag{2}$$

Each component of this mixture corresponds to a hypothetical learner who has been exposed to hypothesised dataset D_z. Figure 1 (top) shows the languages these hypothetical learners might have: as with the real learners, hypothetical learners with little exposure (small D_z) are likely to have a skewed estimate of the language. The size of each guessed dataset is drawn from a Poisson distribution that has a $Gamma(\gamma, 1)$ prior; the ensuing datasets will have a mean size of γ and variance of 2γ. We set γ to different values in our experiments, representing A's different prior beliefs over B's likely data exposure. The hypothetical datasets are then generated by sampling from A's language ($P_A(L)$) until they are the desired size. Note that A must trust their own language to be representative: they do not have access to the true language. The weights ω_z are initialised uniformly.

Updating the Speaker Model during interaction The agent must update their speaker model of their interlocutor as they interact with them, since the utterances in the conversation will provide two different kinds of information relevant to the speaker model:

1. Utterances heard by B (produced by A) will result in updates to B's language, which in turn need to be reflected in A's model of B.

2. Utterances produced by B provide evidence for B's language, which in turn is evidence about B's prior language exposure D_B.

The first kind of information (A's utterances) are added to the data heard by B and incorporated into B's posterior. Within A's internal speaker model, each hypothetical learner is also updated with the new datapoints.

The second kind of information (B's utterances) result in Bayesian updates of the speaker model, which amount to updating the weights over possible languages (posteriors given hypothesised historical language data). The new weight for a specific hypothetical learner, ω'_z, after hearing B produce the variant x is the old weight updated by the probability of the word under the language inferred from the hypothesised dataset D_z.

$$\omega'_z = \omega_z + p(x = i | L_z) \tag{3}$$

$$= \omega_z + \frac{a_{zi}}{\sum_j a_{zj}} \tag{4}$$

Figure 2 shows how A's speaker model weights are updated as A hears utterances produced by B.

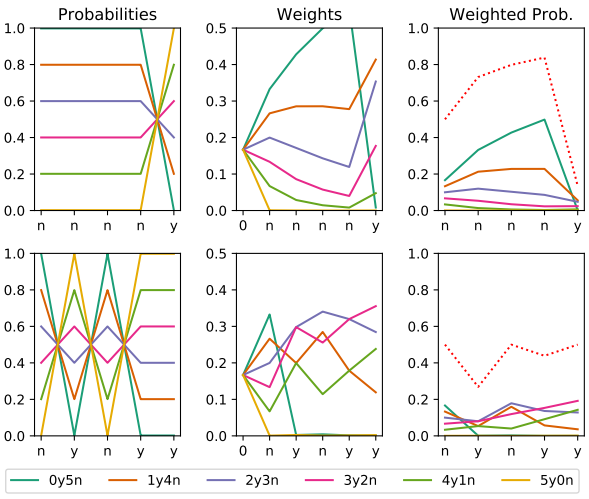

Figure 2: Two examples (top and bottom) of how successive utterances from B lead to updates in the speaker model. B speaks by uttering either 'y' or 'n' (shown on the x axes); values of the speaker model are shown after updating with this utterance. The speaker model consists of six hypothesised learners (coloured lines), each having seen a dataset of size 5, where '4y1n' corresponds to the learner having seen four 'y's and one 'n'. The hypothetical learners do not alter their posterior probabilities (the first column shows each learner has the same probability of 'y'/'n' throughout) but their weights get updated (second column), leading to changes in weighted probabilities (third column) and a re-estimated mixture posterior (third column, red dotted line). The top graph shows that as B produces more 'n', the components with a history of more 'n' ('0y5n', '1y4n') get weighted increasingly heavily, until B produces a 'y', at which point the all-'n' history gets completely down-weighted. The bottom graph shows the same process with a different sequence of B utterances, leading to different posteriors.

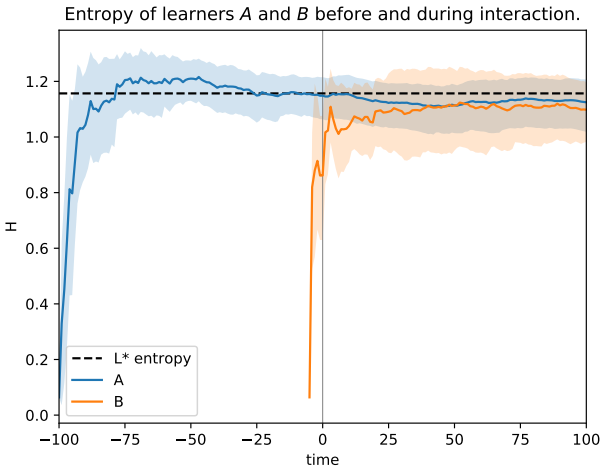

Figure 3: Two learners converge on the original language $L* = [0.7, 0.2, 0.1]$. Learner A has seen 100 items from $L*$ prior to interaction while learner B has seen only 5. During interaction (from timestep 0 onwards), both learners update from each other's utterances. Despite the distance between the languages at the start of interaction, B converges to A's language, instead of both finding a middle ground, because of A's more informed posterior. Shaded regions depict the 95% confidence interval over 10 samples.

Sampling from the speaker model involves first sampling a hypothetical learner's language $P(L_z)$ from a distribution parameterised by ω_z, and then sampling a word from that language as in the language learner model.

Similarly, the entropy of the speaker model is measured as the (weighted) average over the entropies of L_z. Figure 1 (bottom) shows how the speaker model can capture the expectation that with smaller datasets, the (hypothesised) speaker will produce lower-entropy languages.

Interaction

A native- and non-native speaker interact in a dialogue, where each agent speaks in turn (drawing a single sample from their language) and the other listening and updating their posterior and speaker model as appropriate. As described above, the difference between the two types of speakers is in the amount of data they have been exposed to: the native speaker has seen 2–3 orders of magnitude more data (1–20 instances for the non-native speaker vs. 1000 for the native speaker).

The native speaker also has a speaker model over the non-native speaker. We set the hyperparameter γ controlling the size of the speaker model's hypothesised datasets to be the size of the non-native speaker's pre-interaction exposure; this implies that the native speaker can accurately gauge the non-native speakers' prior experience.

It would be theoretically possible to also have the non-native speaker modelling the native speaker. However, the non-native speaker does not have a sufficiently accurate language model to generate plausible datasets for the native-speaker; in effect, the non-native's speaker model of the native speaker would recapitulate (and possibly exaggerate) the errors/bias of the non-native speaker's own language. We assume the non-native speaker is aware of this and thus chooses to use the native-speaker only as a trusted source for learning.

During interaction, both individuals update their models based on what they hear from the other. For a non-native speaker, this constitutes a large percentage of their total exposure to the language. On the other hand, the native speaker will already have a very sharp posterior as their language distribution, and a few low-likelihood utterances from the non-native speaker will not strongly alter this posterior. Figure 3 shows how this interaction, in the absence of accommodation by the native speaker (modelled below) leads the interacting pair to converge on the true language, L^*.

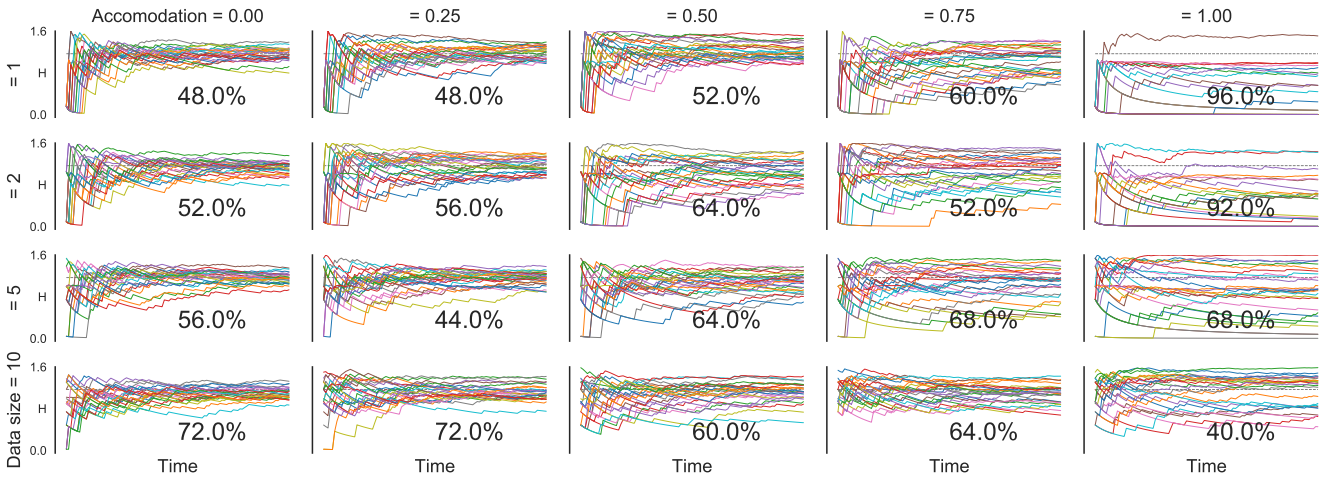

Figure 4: Learner language entropy (y-axis) during interaction (x-axis) with a native speaker accommodating at different levels (left to right). The learner has seen more (bottom) or less (top) data prior to interaction. The lines show different 25 interaction runs with the same parameters (100 timesteps, Z=25, a_0=0.01). The percentage shows the proportion of runs that resulted in a non-native speaker language with lower entropy than the initial language ($L^* = [0.7, 0.2, 0.1]$, $H(L^*) = 1.16$). Higher accommodation levels lead to the learner learning more variable languages and frequently regularising.

Accommodation during interaction

In order to be understood, the native speaker may speak in a way that conforms less to their own language and more to the language they believe their (non-native) interlocutor to have: they may *accommodate* their partner. Accommodation happens in the model when the native speaker uses their speaker model of the non-native speaker's language to speak, rather than their own language. The non-native speaker updates their language based on what they hear from the native speaker (accommodated or not), as before; likewise the native speaker updates both their own language and their speaker model from the non-native speaker's utterances.

The degree of accommodation is set by a fixed parameter that indicates the probability of the native speaker generating utterances from their speaker model of their interlocutor's language versus their own language. A speaker with an accommodation level of 0.75 uses their own native language 25% of the time and their speaker model 75% of the time.

Figures 4 and 5 show how the non-native speaker's language develops in interactions with varying degrees of accommodation. Without accommodation (left-most column of Fig. 4), the learner ends up with a language close to the original language, L^*. With more accommodation, the learner's final language is further from the original language and likely to be regularised, with lower entropy compared to the original language. This is due to a reinforcing positive feedback loop between accommodation by the native speaker and regularisation by the non-native speaker: the native speaker's expectation that the non-native speaker will use a regularised form of the language leads them to produce more regular data; this regular data leads the non-native speaker to infer a more regularsed language, and produce relatively regular utterances which in turn confirm the native speaker's belief than the non-native speaks a regularised version of the language.

Figure 4 also shows the relationship between prior experience of the learner (size of their initial dataset D) and accommodation. In general, more experienced non-native speakers require higher accommodation levels, which is unsurprising: the accommodating native speaker has to outweigh the learner's initial exposure to the language. In a population-level model, learners who are only ever exposed to accommodated language (similar to our small-D learners) may require lower levels of accommodation in order to produce a regularised version of the language.

Finally, characteristics of the original language affect learner regularisation, as shown in Figure 5. Languages with initial high entropy undergo more consistent regularisation. The pattern is less clear with less variable, low-entropy languages, indicating that this model will not capture the extinction of variants (a general issue with parametric probabilistic models), or at least not in a single bout of interaction. Languages with lower complexity, in the form of fewer variants, are somewhat less susceptible to regularisation. More complex, higher entropy languages are more likely to lead to extreme (regularised) variants early in learning, in which one or more variants are unseen; when the native speaker accommodates to this initial regularisation, the non-native speaker's language therefore remains highly regularised.

Conclusion

The goal of the model was to test a mechanism whereby language simplification can happen as a result of interaction between agents with different levels of linguistic ability. Our account involves agents that not only infer their own language,

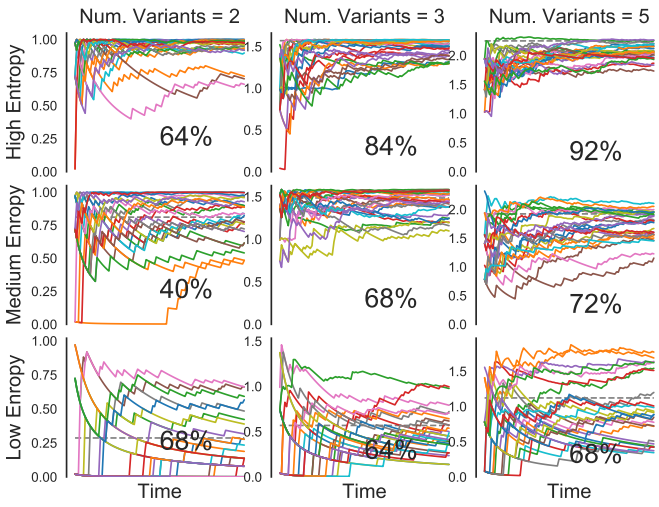

Figure 5: Learner language entropy (y-axis) during interaction (x-axis), varying the initial language L^* (shown in subplot title). Accommodation levels are set throughout at 0.75; the non-native's initial dataset size is 5, other parameters set as in Fig 4. From left to right, the language includes more variants (is higher-dimensional), while from top to bottom the language decreases in initial entropy. Languages with more variants are more likely to regularise, along with languages with high initial entropy. The languages are: first column, $K = 2$: [.55, .45] (H=0.99), [.75, .25] (H=0.81), [.95, .05] (H=0.29); second column, $K = 3$: [.4, .3, .3] (H=1.57), [.5, .3, .2] (H=1.49), [.9, .05, .05] (H=0.57), third column, $K = 5$: [.3, .15, .15, .15, .15] (H=2.25), [.5, .2, .15, .1, .05] (H=1.92), [.8, .05, .05, .05, .05] (H=1.12).

but also make inferences about the language of others. Native speaker agents who act cooperatively by accommodating their interlocutor's language skills lead their non-native interlocutors to learn simplified languages, due to not having exposure to the true language. The more complex the initial language is, the stronger the drive towards regularisation.

We plan to extend the current dialogue model to the population level. Within a population, native speakers could also end up with a simplified language, either due to sufficient exposure to regularised non-native speaker languages, or as a result of generational turnover and initial learning from a mixture of native and non-native languages. This setting will also allow us to explore the potential differences between accommodation in two types of asymmetric interactions, i.e., the native–non-native interactions explored here as well as adult–child learner interactions.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 681942).

References

Atkinson, M., Smith, K., & Kirby, S. (submitted). Adult learning and language simplification.

Bentz, C., & Winter, B. (2013). Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change*, *3*, 1–27.

Brown-Schmidt, S. (2009). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language*, *61*(2), 171–190.

Cai, Z. G., Gilbert, R. A., Davis, M. H., Gaskell, M. G., Farrar, L., Adler, S., & Rodd, J. M. (2017). Accent modulates access to word meaning: Evidence for a speaker-model account of spoken word recognition. *Cognitive Psychology*, *98*, 73–101.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Clark, H. H. (1998). Communal lexicons. In K. Malmkjaer & J. Williams (Eds.), *Context in language learning and language understanding* (pp. 63–87). Cambridge University Press.

Ferguson, C. A. (1975). Toward a characterization of English Foreigner Talk. *Anthropological Linguistics*, *17*, 1–14.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998.

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, *35*(1).

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In H. Giles, J. Coupland, & C. N. (Eds.), *Contexts of accommodation* (p. 1-68). Cambridge University Press.

Hahn, U. (2014). Experiential limitation in judgment and decision. *Topics in Cognitive Science*, *6*(2), 229–244.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*(8), 534–539.

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

Martin, C. D., Garcia, X., Potter, D., Melinger, A., & Costa, A. (2015). Holiday or vacation? The processing of variation in vocabulary across dialects. *Language, Cognition and Neuroscience*, *31*(3), 375–390.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*, 317–328.

Smith, K., Perfors, A., Fehér, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017). Language learning, language use, and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society B*, *372*, 20160051.

Snow, C., & Ferguson, C. A. (1977). *Talking to children: Language input and acquisition*. Cambridge: Cambridge University Press.

Trudgill, P. (2011). *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford University Press.

Wray, A., & Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, *117*, 543–578.