Language Evolution and Computation
Proceedings of the Workshop/Course at ESSLLI, Vienna 2003

Simon Kirby (ed.)
University of Edinburgh

http://www.ling.ed.ac.uk/~ simon/esslli.html

Modelling of sound systems

Bart de Boer
Artificial Intelligence Lab – Vrije Universiteit Brussel
Pleinlaan 2, B-1050 Brussels, Belgium
bartb@arti.vub.ac.be

Abstract

This paper is an introduction to the computer modelling of systems of speech sounds. It focuses mainly on modelling that studies (the prerequisites of) evolution, so pure speech recognition is not treated. However, both cultural and biological evolution are considered as well as aspects of learning and social interactions that might be relevant to the evolution of speech. It will be argued that speech sounds are a relatively well-studied and simple-to-model aspect of language that nevertheless has interesting aspects in common with syntax. Compositionality is probably among the most exciting of these.

1 Introduction

Speech is an aspect of language that has been well investigated in linguistics. Universals, historical change, perception, production, processing and acquisition have all been researched in great detail. However, this detailed knowledge is not reflected in the amount of computer modelling effort that has gone into understanding the evolution of linguistic sound systems. Although one of the first computer models investigating factors in language origins was about sound systems (Liljencrants & Lindblom, 1972), the more recent surge of interest in computational models of language origins (see Kirby, 2002 for an overview) has resulted in relatively few papers on modelling of sound systems. de Boer (2002) presents an overview of recent work on computer modelling of evolution of speech sounds, but this body of work is relatively small compared to work on syntax and semantics.

A lot of information is available about sound systems of human language. This is partly because sounds are concrete signals that can be recorded and measured objectively. In contrast, more abstract aspects of language such as syntax and semantics, can only be investigated in a much more indirect way. Results of simulations involving sound systems can therefore be compared much more readily with observations of real human language. It is also easier to build models of perception, production and processing that are close to how humans handle speech sounds, although, of course, there is always the issue of computational complexity. These properties make speech sounds an attractive area of computer modelling research.

It could be suggested that speech sounds are less interesting than syntax and semantics, and effectively, systems of speech sounds show less complexity than syntactic and semantic systems. One can also argue that syntax is the fundamental property of human language, and that examples of call systems similar to human speech can be found in songs of songbirds and whales. Nevertheless, speech sounds do have a complexity of their own. Just as individual words can be combined into sentences according to the rules of syntax, speech sounds can be combined into words according to definite and sometimes reasonably complex rules. These rules have to be learned by infants in a way that is very similar to the way infants have to learn syntactical rules. On the basis of these and similar observations Carstairs-MacCarthy (1999) has argued that phonology is the evolutionary precur-

sor of syntax. Jackendoff (2002, section 8.5) also observes that phonological systems must have been the first combinatorial systems in the evolution of human language.

Different aspects of evolution of sound systems have been modelled. These include the way in which linguistic diversity can emerge (e.g. Livingstone & Fyfe, 1999), the way in which sound systems can be learned and the role of mother-child interactions in this process (e. g. de Boer & Kuhl, 2001), the emergence of sound systems in populations (*e.g.* de Boer 2000), and the emergence of phonemic coding (e. g. Lindblom, MacNeilage, & Studdert-Kennedy, 1984; Steels & Oudeyer, 2000). Different approaches have also been applied to the problem of understanding origins and evolution of speech sounds. Models have been based on straightforward optimisation, genetic algorithms or agent-based models. Speech signals have been implemented as abstract symbols or as more realistic signals. Learning of speech sounds has been implemented with connectionist as well as more classical learning mechanisms, and some approaches (especially those using genetic algorithms) do not implement learning of speech at all.

This paper is meant as an introduction to speech sound modelling for an interdisciplinary audience. Therefore it contains material that might not be of interest to all readers. Section 2 contains a crash course in phonetics and phonology and can safely be skipped by linguists. Only the sections on universals of speech sounds (2.1) might be of interest to them. Section 3 contains an overview of work on modelling of speech sounds, and is hopefully of interest to all audiences. Section 4 in turn contains some considerations of modelling and can safely be skipped by dyed-in-the-wool cognitive modellers. Casual computer scientists, on the other hand are encouraged to read this section, though. The discussion at the end of the paper contains some suggestions of future work and is, hopefully, again of interest to all audiences.

2 Speech sound crash course

All human languages, except those used by deaf communities, use sound as their means to transmit signals. As human language must be able to convey a large and potentially infinite number of different messages efficiently, it needs a large and extensible repertoire of signals. The solution human language uses for this is to recombine a small set of basic sounds into larger assemblies. I will discuss properties of human speech in a slightly simplified fashion below. It should be kept in mind, however, that every aspect of human language is complex and idiosyncratic, and that for each of these aspects, including speech, many competing theories abound. The material I present reflects my understanding of the subject and also reflects what I think is necessary to successfully make computer models for studying (evolution of) human speech. Nonetheless, the reader is encouraged to study the linguistic and phonological theory independently. Any good textbook on the subject of phonology of phonetics that is not too much centred on one particular theory will do, although Ladefoged and Maddieson's book (Ladefoged & Maddieson, 1996) is especially recommended for those who like to marvel at the diversity and complexity of human languages.

Most linguists would say that phonemes are the basic unit of speech. Phonemes are defined as the smallest articulations that can change the meaning of a word. An example for English would be /l/ and /r/, as changing /l/ into /r/ can change the meaning of a word: "law" and "raw" have different meanings in English. Two words that differ only in one phoneme are called minimal pairs. By carefully studying minimal pairs in a language, linguists can determine the set of phonemes a language uses. However, there are complications. Phonemes are not always pronounced identically in all contexts. In standard English, /l/ at the beginning of a word is pronounced quite differently from /l/ at the end of a word. When /l/ occurs at the end of a word, it is velarised, at the beginning of a word it is not. However, not velarising an /l/ at the end of the word does not change its meaning, it only makes the speaker sound funny, or possibly as if he or she has an Irish accent. There are languages that do use this distinction. In Russian, for example, changing a palatalised /l/ (which sounds ap-

proximately like English /l/ at the beginning of a word) into a velarised /l/ changes the meaning of a word. For instance: мель “shoal”, ending in palatalised /l/ contrasts with мел “chalk” ending in velarised /l/. Sounds that are phonemes in one language do not have to be phonemes in another, and they are therefore language-dependent.

There are other levels at which speech can be considered combinatorial. Phonemes are problematic in certain ways. First of all, their independent existence is not so clear as it might seem at first glance. Independently pronounceable phonemes, such as vowels, nasals (/m/, /n/) and fricatives (/v/, /s/) clearly do have an independent existence. But what about sounds such as /p/ and /t/? These can only be pronounced clearly in the context of a vowel. And indeed, it turns out that many speech sounds are detected mostly by the way they influence neighbouring sounds. Even independently pronounceable sounds are heavily influenced by the surrounding sounds. This effect is called co-articulation. This causes the interesting phenomenon that a speech signal cannot usually be cut up into its constituent phonemes and be recombined into another understandable utterance. Although phonemes definitely do have some status as units of recombination, it might be necessary to look at the syllable level as well.

One encounters many subtle and complex problems when studying syllables, but a simplified account would consider syllables to consist of an onset, a nucleus and a coda. Typically, the nucleus consists of the vowel part of a syllable[1]. The onset consists of the consonants preceding the nucleus and the coda consists of the consonants following the nucleus. Thus in English “sprint”, /i/ is the nuclues, /spr/ is the onset and /nt/ is the coda. Syllables can also contain what is called suprasegmental material (modifications of the signal spread over the whole syllable), such as tone. It is usually possible to cut up a speech signal into syllables and recombine these, such that an intelligible result ensues. Therefore, syllables are uncontroversial as a unit of recombination. However, the number of syllables per language can become quite large, and these syllables can often easily be analysed in terms of smaller units – either phonemes or onsets, nuclei and codas. It is therefore likely that both the syllabic level and the phonemic level have cognitive significance.

A third level of analysis of speech is also generally assumed. This is the level below that of phonemes. It turns out that the phoneme inventories of languages are systematic. For example, languages that use sounds like /b/, /d/ and /g/ tend to use /p/, /t/ and /k/ as well. These sounds can be paired such that the elements of each pair only differ in voicing. The sound /b/, for example is articulated with the lips and is voiced (the vocal chords need to vibrate) while the sound /p/ is articulated in the same manner, but it is voiceless. These sounds are said to differ only in the *feature* ‘voicing’. Speech sounds can thus be analysed as consisting of several distinctive features. Apart from voicing, distinctive features can represent place of articulation, manner of articulation or nasalisation, for example. Almost always distinctive features are considered to be binary. Different researchers of speech sounds often use different sets of features. Many processes that have to do with how sounds are combined into words and how words are simplified in rapid, casual speech can be explained in terms of distinctive features. Although the exact nature and cognitive reality of distinctive features is hotly debated, there is no doubt that they are a useful tool for linguistic description, and speech can definitely be analysed as combinatorial on the level of distinctive features.

2.1 Universals

Systems of speech sounds of human languages are not random collections of sounds. If this were the case, there would be very little to study or model. Although humans can produce and distinguish an amazing number of different sounds (Ladefoged & Maddieson, 1996), the phoneme inventories of human languages show remarkable regularities. Most languages from a reasonably representative

[1] There are many languages that allow consonants as the nucleus, for example, Czech “krk” (neck, throat). Deciding what exactly the syllable nuclei are in such languages can be tricky.

sample of 451 languages (Maddieson, 1984; Maddieson & Precoda, 1990) have between 20 and 37 phonemes, while the most frequently occurring number of phonemes is 25. The minimum number of phonemes appears to be 11 for the East-Papuan language Rotokas (Firchow & Firchow, 1969) and the South-American language (Murà-) Pirahã (Everett, 1982; Sheldon, 1974) spoken by men and 10 for Piraha spoken by women (Daniel Everett, personal communication). The maximum number of phonemes for a language in UPSID is 141 for the Khoisan language !Xũ (Snyman, 1970) while over 160 phonemes have been reported for the Khoisan language !Xóõ (Traill, 1985). Because of the way clicks are analysed in these languages, these numbers might be slightly inflated, and the real maximal number of phonemes per language might be closer to 90, for example for certain North-East Caucasian languages (e. g. Catford, 1977).

Not only are there regularities in the repertoire sizes, but there are also regularities in the kinds of sounds that occur. Certain sounds occur much more often than others, and certain speech sounds also co-occur more frequently than predicted from the frequencies of the individual sounds themselves. For regularities of phoneme inventories see e.g. (Schwartz, Boë, Vallée, & Abry, 1997b) while for more general regularities see (Maddieson, 1984). Also, languages with small inventories tend to use a limited number of basic articulations, while languages with larger inventories tend to use more elaborated articulations that can often be analysed as combinations of more basic articulations (Lindblom & Maddieson, 1988). It is not always possible to distinguish between sequences of phonemes on the one hand and complex phonemes on the other hand. This is especially true for the complex clicks found in some of the Khoisan languages (as mentioned above) that are analysed as single phonemes, but that could in principle also be analysed as sequences of basic clicks and secondary articulations (Traill, 1985) (incidentally reducing the number of phonemes in these languages substantially).

The number of syllables in languages varies much more than the number of phonemes. Maddieson (1984, section 1.9) presents the numbers of possible syllables for nine languages. The numbers of theoretically possible syllables ranges from 162 for Hawaiian to 23 638 for Thai (taking into account tones as well). However, the way in which speech sounds can be combined into larger wholes (syllables, words) does show regularities. Certain speech sounds tend to occur close to the nucleus of a syllable, while other sounds tend to occur near the periphery. Speech sounds can be ordered hierarchically, such that sounds higher on the hierarchy tend to occur closer to the nucleus of a syllable when co-occurring with sounds occurring lower on the hierarchy. This is often called the sonority hierarchy. For a description, see (Vennemann, 1988). It is approximately as follows a > w > l > n > s = t (where each phoneme represents the sounds from its category). Many languages allow only very simple syllables, such as syllables consisting of a vowel only, or syllable consisting of a single consonant followed by a single vowel. Such constraints imply that phonemes cannot be combined freely. These constraints are probably due to articulatory and acoustic factors, such that speech that follows them is easier to produce and perceive.

Tone systems, the ways in which languages use pitch and pitch contours to make distinctions between words, also has its universals. These have to do with the size of the inventories, with the frequencies with which tones occur (falling tones appear to be preferred over rising tones) and the way in which tones combine into tone inventories. Unfortunately, I am quite ignorant on the subject of universals of tone, but the interested reader is referred to (Maddieson, 1978).

These regularities observed in human languages and which are usually called *phonological universals* are the factual basis that successful computational models should explain. If a model gives results that are significantly different from what is observed in real human languages, it will not be interesting for linguists, even though the model might have other merits. Also, it is important for researchers coming from a different domain to have a good knowledge of the linguistic facts, in order to gain the necessary scientific street credibility to have ones ideas accepted in the linguistic community.

2.2 Other sources

Linguistic universals are probably the most interesting source of data for researchers interested in modelling language evolution. Results of experiments can be "grounded" in this kind of data. However, when one wants to build more realistic agents, or if one is interested in modelling individual's language behaviour, other areas of linguistics need to be taken into account. General linguistics is of course the area in which most formal models of language are generated, and Jackendoff (2002) provides a good starting place. Jackendoff has interesting things to say about computer modelling as well as language evolution. However, for use in computer modelling, models of general linguistics might be too general, and therefore difficult if not impossible to implement.

For factual data about language learning and language performance, studies of infant development and psycholinguistics might provide interesting material. Infants learn to speak amazingly quickly and according to a rather fixed pattern. Also the parent-child interactions show cross-cultural similarities (i. e. Ferguson, 1964; Fernald et al., 1989). Good books that provide an introduction to the field of infant phonological development are Vihman (1996) and Jusczyk (1997). These provide detailed background information on the stages infants go through when learning how to speak, while being relatively theory-free. As for psycholinguistics, an easy-to-understand introduction to the field is provided by Field (2003) but admittedly, this is the only book I read on the subject, so there might be better ones around. A background in these subjects is relatively easy to acquire and will certainly increase the quality of the modelling work, as well as the ability of the modeller to convince an audience of linguists.

3 Overview of speech modelling[2]

Probably the first attempt at making a computer model to explain universals of speech sounds was made by Liljencrants & Lindblom (1972). This model optimized randomly initialized vowel systems with a fixed number of vowels. The optimization was based on a function that modelled the potential energy of repelling magnets (this potential energy is higher whenever the magnets are closer together). By shifting the individual vowels in the system, the energy function was minimized. Liljencrants and Lindblom found that vowel systems that were optimized in this way were remarkably similar to vowel systems found in human languages, although there were some discrepancies. Later re-implementations that used modified distance functions (e. g. Schwartz, Boë, Vallée, & Abry, 1997a; Vallée, 1994) have succeeded in making progressively better approximations of human vowel systems.

Subsequently, Lindblom et al. (1984) have tried to use an optimising model for explaining phonemic (that is combinatorial) coding of syllables. The syllables consisted of a simple consonant followed by a vowel. Although the systems that emerged were phonemically coded, their model has not had the success of the model for vowels, because there are many more parameters in it and because it is much more difficult to replicate the results.

Only in the mid-nineties did work on explaining sound systems with computer models get a new impulse with systems that were based on populations of sound systems and populations of agents. The first to make an agent-based implementation to investigate the emergence of vowel systems was Glotin (Berrah, Glotin, Laboissière, Bessière, & Boë, 1996; Glotin, 1995) of the Institut de Communication Parlée (ICP) in Grenoble, the same institute were Schwartz et al. (1997a) do their research. He made a model in which a population of talking agents tries to develop a shared repertoire of (a fixed number of) vowels. His agents have an acoustic as well as an articulatory representation of the

[2] This is a slightly reworked and updated version of the material presented in the history of modelling section of (de Boer, 2002).

vowels, and adapt their vowel systems on the basis of their interactions. The agents are also subject to a genetic algorithm, which is (according to Glotin, personal communication) not meant to be a model of actual biological evolution of the agents, but rather of the way sound systems are transferred from parents to children. This might be considered a weak point of the research, as the influence of the genetic algorithm and the interactions between the agents are difficult to separate. Another problem with the model was that it was computationally too complex, and that therefore only few simulations with small populations and small numbers of vowels could be run. In a way, this work was ahead of the computing power of the time.

It has been at the basis of a number of subsequent research efforts, however. In the first place those of Berrah (Berrah, 1998; Berrah & Laboissière, 1999) and myself (de Boer, 1997, 2000, 2001). Berrah's work was a direct continuation of Glotin's research. Berrah's model is a simplification of Glotin's model. The agents do no longer have an articulatory representation of the sounds they use, only an acoustic one. This reduces the computational load considerably and allows more experiments with larger populations and larger numbers of vowels to be run. Berrah extends Glotin's model by investigating what he calls the "Maximum Use of Available Features". By allowing the agents to use an extra feature (which could be length, nasalization etc. in human languages, but which he models as an extra abstract dimension of the acoustic space) he shows that this is only used whenever the number of vowels in the agents' repertoires exceeds a certain threshold. His simulations also contain a genetic component, which makes it sometimes hard to tell when a particular phenomenon is due to interactions between the agents and when it is due to the actions of the genetic algorithm.

My own work has concentrated on predicting vowel systems from interactions in a population. The agents have both an articulatory as well as an acoustic representation of their vowels, but use a much simpler articulatory model than the one used by Glotin. Also, the agents do not evolve, although experiments have been done with changing populations (de Boer & Vogt, 1999). They interact through language games (in this experiment called imitation games) only. It has been shown that vowel systems of human languages, and the relative frequencies with which they occur can be predicted quite accurately with this model.

Daniel Livingstone and Colin Fyfe of the university of Paisley have investigated the origins of linguistic diversity (Livingstone & Fyfe, 1999). They model a population of agents that has a spatial structure and monitor how linguistic diversity changes over time. The research question they are addressing is how it is possible that there are many different languages, and under what conditions such diversity can arise. Their work builds on work performed by Daniel Nettle on the emergence of linguistic diversity (Nettle, 1999).

More recently research has started to investigate syllable systems with genetic algorithms and population models. This work relates in a similar way to the optimizing simulation used by Lindblom et al. (1984) as Glotin's, Berrah's and my own work relates to Liljencrants' and Lindblom's (1972) model. Redford and colleagues of the university of Texas, Austin (Redford, Chen, & Miikkulainen, 2001) have made a model that is based on a genetic algorithm. The population consists of words, which in turn consist of a closed set of phonemes. Redford et al. use a number of rules that determine how hard it is to produce and perceive different combinations and sequences of phonemes. On the basis of this a fitness for all the words in the population is calculated and selection and recombination take place. They try out different combinations of rules and investigate which rules are most important to predict syllables that are like those found in human languages.

Other work on predicting properties of more complex utterances is being conducted and published at the moment. Pierre-yves Oudeyer of the Sony computer science laboratory in Paris, France is working on predicting repertoires of syllables using more realistic signals (e. g. Oudeyer, 2002; Steels & Oudeyer, 2000). Professor William Wang of the electronic engineering department of the City University of Hong Kong and co-workers Jinyun Ke and Mieko Ogura are working on modelling tone systems within the framework of genetic algorithms (Ke, Ogura, & Wang, 2003).

My own most recent work, in cooperation with Patricia Kuhl of the University of Washington is in investigating the role of mother-child interactions in the transfer and evolution of language. We have investigated with a computer model how infant-directed speech can facilitate learning (de Boer, to appear; de Boer & Kuhl, 2001) and how it can facilitate transfer of vowel systems from generation to generation (de Boer & Kuhl, to appear).

4 Modelling considerations

Making successful computer models of the evolution of speech is somewhat of an art. One needs to find the right balance between performance on the one hand and realism on the other. In order to do simulations that are relevant to linguists, it is necessary to make models that work with something that is close to real speech. Such models can become computationally intractable, though. It is often better to sacrifice excessive realism in order to have better performance. Thus one can do multiple repeated experiments, work with larger populations, or work with larger repertoires of sounds. It is easy to get carried away (especially for phoneticians trying to build computational models of articulation) and to want to use the most accurate models of speech production and perception available. This is not always the best strategy, in my opinion, as the bottleneck determining the realism of the model is usually not in the articulatory or perceptual model, but in the implementation of the more cognitive aspects of the model. These include, but are not limited to, the way the system recognises speech sounds, the way in which it learns them, or the way it coordinates movements of the vocal tract.

Simplifying too much is not a good idea, either. Although interesting work can be done by highly abstract models, it is always easier to convince linguists if one stays close to real speech. If one does choose to work with abstract signals, it is crucial to point out how these map to real speech and how the results need to be translated for and interpreted by linguists. Of course, showing results that are directly understandable to linguists increases the probability that one's work is accepted in their community enormously.

One should try to minimise the number of parameters in the system. When there are many parameters, the impression could arise that the model's results are caused by tuning, rather than that it captures something essential. This is probably the reason why Liljencrants and Lindblom's (1972) model on vowels was much more successful than Lindbom et al.'s (1984) later model on syllables. In any case one should try to derive parameter values from independent (physical, physiological, perceptual, articulatory or cognitive) considerations, and show that the performance of the model does not depend sensitively on the exact parameter values.

Modellers of speech can derive their inspiration from models used by engineers working in speech recognition and speech synthesis, but should be aware that the aims of the engineering approach are totally different from the aim of computer modelling of speech. Engineers are satisfied when their models achieve better recognition rates, or more realistic speech production, no matter how cognitively implausible their models are. When modelling speech from a cognitive perspective, one should always keep in mind that in the end one is trying to understand the workings of the human brain. This is not to say that one should always use models that are directly based on the architecture of the brain, such as neural networks. Often it is necessary to use higher level models in order to achieve more complex behaviour, or in order to keep the model's behaviour transparent. Nonetheless, it is always a good idea to remember that one is trying to understand cognition and not trying to build more and more fancy computer models.

There are a few other, more general things one has to keep in mind when doing computer modelling of speech. First of all, it is important that the statistical significance of the results be established. Although this may appear to be a bit of a superfluous statement, it is amazing how many computational modelling papers are written (and accepted for publication) without a statistical analysis of the

results. The author pleads guilty in this respect, too. However, the fact that results are accepted on good faith in the computational modelling community does not mean that they are accepted on this basis in other communities. Secondly, as has been mentioned above, it is a good idea to do a sensitivity study of one's model. This means that all parameters are varied and that it is reported how these changes do (or do not) influence the behaviour of the model. And finally, one should always be careful about bias in the model. Many modelling studies depend on random initialisations and random changes. Sometimes it is not straightforward to make sure that such initialisations and changes are unbiased. It is therefore always a good idea to test this before enthusiastically reporting results that are caused by systematic error instead of the model's dynamics.

5 Discussion

Speech is an interesting part of language that is well-known and easy to model, but that nevertheless has the potential of providing insights on language evolution that might have repercussions on the understanding of syntax as well. Speech sounds form sequences that are combinatorial and in this respect they resemble syntax. By some researchers (e. g. Carstairs-McCarthy, 1999; Jackendoff, 2002) speech is even considered a possible precursor to syntax. Also, learning of sequences is something that is necessary for both speech and syntax. Apart from this, many interesting and linguistically relevant results have already been achieved within the domain of the study of speech proper.

A lot of problems are still open, and many have to do with extending the models to more complex speech sounds and sequences of speech sounds. Working with more complex signals involves more complex articulatory and perceptual models, as well as learning time sequences. These are problems that have only been very partially solved in the research on computer speech production and processing. However, it is not necessary to wait for progress in these fields before interesting new experiments can be done. First of all, a lot of the more sophisticated production- and processing models have not yet been used in the modelling of the evolution of language. It is also possible to do many interesting experiments into the evolution of speech sounds with simpler abstract models. The transition from holistic to phonemically coded signals, for example, is still poorly understood. Also, the role erosion of phonetic forms of words and morphemes plays in grammaticalisation could be investigated with more abstract signals. At the moment, models of emergence of grammar generally assume that language consists of strings of discrete symbols and that words are separated by silence. In real speech this is clearly not the case, and it would be interesting to investigate the combination of a grammar-learning model and a realistic phonetic component. Finally, realistic sound change has not been modelled successfully, as far as the author is aware. A model that could produce realistic chain shifts and context-dependent sound changes would be extremely interesting. To achieve this, a component of meaning probably needs to be integrated. At the moment, most models investigating sound systems only have a very rudimentary implementation of meanings of utterances.

Apart from these examples, many other experiments can be conceived that would shed light on evolution of speech sounds as well as problems that exceed the domain of speech sounds alone. Therefore it is not just interesting to continue research into speech sounds, but it is also necessary to open the dialogue between modellers of speech sounds and modellers of other aspects of language.

References

Berrah, A.-R. (1998). *Évolution Artificielle d'une Société d'Agents de Parole: Un Modèle pour l'Émergence du Code Phonétique*. Grenoble: Thèse de l'Institut National Polytechnique de Grenoble, Spécialité Sciences Cognitives.

Berrah, A.-R., Glotin, H., Laboissière, R., Bessière, P., & Boë, L.-J. (1996). From Form to Formation of Phonetic Structures: An evolutionary computing perspective. In T. Fogarty & G. Venturini (Eds.), *ICML '96 workshop on Evolutionary Computing and Machine Learning, Bari* (pp. 23–29).

Berrah, A.-R., & Laboissière, R. (1999). SPECIES : An evolutionary model for the emergence of phonetic structures in an artificial society of speech agents,. In D. Floreano, J.-D. Nicoud & F. Mondada (Eds.), *Advances in Artificial Life, Lecture Notes in Artificial Intelligence* (Vol. 1674, pp. 674–678). Berlin: Springer, 1999.

Carstairs-McCarthy, A. (1999). *The origins of complex language: an inquiry in the evolutionary beginnings of sentences, syllables, and truth*. Oxford: Oxford University Press.

Catford, J. C. (1977). Mountain of tongues: the languages of the Caucasus. *Annual Review of Anthropology, 6*, 283-314.

de Boer, B. (1997). Generating vowel systems in a population of agents. In P. Husbands & I. Harvey (Eds.), *Fourth European Conference on Arti.cial Life.* (pp. 503-510). Cambridge (MA): MIT Press.

de Boer, B. (2000). Self organization in vowel systems. *Journal of Phonetics, 28*(4), 441-465.

de Boer, B. (2001). *The origins of vowel systems*. Oxford: Oxford University Press.

de Boer, B. (2002). Evolving Sound Systems. In A. Cangelosi & D. Parisi (Eds.), *Simulating the Evolution of Language* (pp. 79-97). Berlin: Springer Verlag.

de Boer, B. (to appear). Infant directed speech and the evolution of language. In M. Tallermann (Ed.), *Proceedings of the evolution of language conference (working title)*. Oxford: Oxford University Press.

de Boer, B., & Kuhl, P. (2001). Infant-directed vowels are easier to learn for a computer model. *Journal of the Acoustical Society of America, 110*(5, pt. 2), 2703.

de Boer, B., & Kuhl, P. (to appear). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*.

de Boer, B., & Vogt, P. (1999). Emergence of Speech Sounds in Changing Populations. In D. Floreano, J.-D. Nicoud & F. Mondada (Eds.), *Advances in Artifical Life, Lecture Notes in Artificial Intelligence 1674* (pp. 664-673). Berlin: Springer Verlag.

Everett, D. L. (1982). Phonetic rarities in Piraha. *Journal of the International Phonetic Association, 12*(2), 94-96.

Ferguson, C. A. (1964). Baby talk in six languages. *American Anthropologist, 66*(6, part 2), 103-114.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language, 16*, 477-501.

Field, J. (2003). *Psycholinguistics: A resource book for students*. London: Routledge.

Firchow, I., & Firchow, J. (1969). An abbreviated phoneme inventory. *Anthropological Linguistics, 11*, 271-276.

Glotin, H. (1995). *La Vie Artificielle d'une société de robots parlants: émergence et changement du code phonétique*. Grenoble: DEA sciences cognitives-Institut National Polytechnique de Grenoble.

Jackendoff, R. (2002). *Foundations of language*. Oxford: Oxford University Press.

Jusczyk, P. W. (1997). *The Discovery of Spoken Language*. Cambridge, MA: The MIT Press.

Ke, J., Ogura, M., & Wang, W. S.-Y. (2003). Optimization Models of Sound Systems Using Genetic Algorithms. *Computational Linguistics, 29*(1), 1-18.

Kirby, S. (2002). Natural language from artificial life. *Artificial Life, 8*(2), 185-215.
Ladefoged, P., & Maddieson, I. (1996). *The Sounds of the World's Languages*. Oxford: Blackwell.
Liljencrants, L., & Lindblom, B. (1972). Numerical simulatons of vowel quality systems. *Language, 48*, 839-862.
Lindblom, B., MacNeilage, P., & Studdert-Kennedy, M. (1984). Self-organizing processes and the explanation of language universals. In M. Butterworth, B. Comrie & Ö. Dahl (Eds.), *Explanations for language universals* (pp. 181-203). Berlin: Walter de Gruyter & Co.
Lindblom, B., & Maddieson, I. (1988). Phonetic universals in consonant systems. In L. M. Hyman & C. N. Li (Eds.), *Language, speech and mind* (pp. 62-78). London: Routledge.
Livingstone, D., & Fyfe, C. (1999). Modelling the Evolution of Linguistic Diversity,. In D. Floreano, J.-D. Nicoud & F. Mondada (Eds.), *Advances in Artificial Life, Lecture Notes in Artificial Intelligence* (Vol. Volume 1674, pp. 704-708). Berlin: Springer.
Maddieson, I. (1978). Universals of Tone,. In J. H. Greenberg, C. A. Ferguson & E. A. Moravcsik (Eds.), *Universals of Human Language* (Vol. Volume 2 Phonology, pp. 335-365). Stanford: Stanford University Press.
Maddieson, I. (1984). *Patterns of sounds*. Cambridge: Cambridge University Press.
Maddieson, I., & Precoda, K. (1990). Updating UPSID. *UCLA Working Papers in Phonetics, 74*, 104-111.
Nettle, D. (1999). *Linguistic Diversity*. Oxford: Oxford University Press.
Oudeyer, P.-y. (2002). Phonemic coding might be a result of sensory-motor coupling dynamics. In J. Hallam (Ed.), *Proceedings of the International conference on the simulation of adaptive behavior (SAB)* (pp. 406-416). Edinburgh: MIT Press.
Redford, M. A., Chen, C. C., & Miikkulainen, R. (2001). Constrained emergence of universals and variation in syllable systems. *Language and Speech, 44*, 27–56.
Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997a). The Dispersion-Focalization Theory of vowel systems. *Journal of Phonetics, 25*, 255-286.
Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997b). Major trends in vowel system inventories. *Journal of Phonetics, 25*, 233-235.
Sheldon, S. N. (1974). Some morphophonemic and tone rules in Mura-Pirahã. *International Journal of American Linguistics, 40*, 279-282.
Snyman, J. W. (1970). *An introduction to the !Xu (!Kung) language*. Cape Town: Balkema.
Steels, L., & Oudeyer, P.-y. (2000). The cultural evolution of syntactic constraints in phonology. In M. A. Bedau, J. S. McCaskill, N. H. Packard & S. Rasmussen (Eds.), *Proceedings of the VIIth Artificial life conference (Alife 7)*. Cambridge (MA): MIT Press.
Traill, A. (1985). *Phonetic and phonological studies of !Xóõ bushman*. Hamburg: Helmut Buske Verlag.
Vallée, N. (1994). *Systèmes vocaliques: de la typologie aux prédictions* (Vol. T). Grenoble: hèse préparée au sein de l'Institut de la Communication Parlée (Grenoble-URA C.N.R.S. no 368).
Vennemann, T. (1988). *Preference Laws for Syllable Structure*. Berlin: Mouton de Gruyter.
Vihman, M. M. (1996). *Phonological Development: The Origins of Language in the Child*. Cambridge MA: Blackwell.

Phonemic Coding: Optimal Communication Under Noise?

Bart de Boer
Artificial Intelligence Lab
Vrije Universiteit Brussel
Pleinlaan 2, B-1050 Brussels, Belgium
bartb@arti.vub.ac.be

Willem Zuidema
Language Evolution and Computation Research Unit
School of Philosophy, Psychology and Language Sciences
and Institute of Animal, Cell and Population Biology
University of Edinburgh
40, George Square
Edinburgh EH8 9LL, United Kingdom
jelle@ling.ed.ac.uk

http://arti.vub.ac.be/ ∼ bartb
http://www.ling.ac.uk/ ∼ jelle

Abstract

Human languages are universally phonemically coded, whereas many animal signal systems are not. A number of theories and models have been developed to explain this evolutionary transition, but some major problems remain. We present a simulation to investigate the hypothesis that phonemic coding is an side effect of optimizing signal systems for success in imitation. Crucially, signals in our model are trajectories in an (abstract) acoustic space. Hence, both holistic and phonemically coded signals have a temporal structure. Using both qualitative inspection of emerged systems of trajectories and a statistical analysis of a measure of phonemicity, we find that phonemically coded systems are indeed preferred. The model thus provides a new explanations for the evolutionary pathway to the emergence of phonemic coding.

1 Introduction

One of the universal properties of human language is the fact that it is phonemically coded. Linguistic utterances can be split into units that can be recombined into new linguistic utterances. For instance, the words “we”, “me”, “why” and “my” as pronounced in standard British English are built-up from the units “w”, “m”, “e” and “y”, which can all be used in many different combinations.

There is some controversy about the exact level at which combination takes place. In the traditional view the atomic units are phonemes: minimal speech sounds that can make a distinction in meaning. An increasingly popular alternative view is that the atoms are syllables, or the possible onsets, codas and nuclei of syllables. Nevertheless, there is general agreement that in natural languages, atomic units are combined into larger wholes. For the purposes of this paper, we do not need to take

sides in the debate about the exact nature of the combinatorial elements of human language. Instead, we study signals that occur in an abstract acoustic space, and address the question of why and how phonemically coded sets of signals have emerged.

The combinatorial nature of human speech is in contrast with many animal calls and non-linguistic human utterances, which generally cannot be split into smaller units. The songs of some songbirds and whales, however, do seem to have combinatorial structure. The fact that in evolutionary unrelated lineages combinatorial systems have emerged indicates that such systems can be considered as evolutionary attractors. Recombination apparently has major evolutionary advantages. Two views on the advantages that recombination offers are:

1. It makes it possible to transmit an infinite number of messages over a noisy channel (the "noisy coding argument", an argument from information theory, e.g. Nowak & Krakauer 1999).

2. It makes it possible to create an infinitely extensible set of signals with a limited number of building blocks. Such productivity provides a solution for memory limitations, because signals can be encoded more efficiently, and for generalization, because new signals can be created by combining existing building blocks (the "productivity argument", a point often made in the generative syntax tradition, e.g. Jackendoff 2002);

These advantages are a good starting point for answering the questions of *why* combinatorial coding would emerge, and *how* initially holistic systems (which seem to be the default for smaller repertoires of calls) can change into phonemically coded systems. In this paper we will address both questions. In the following we will discuss some existing formal models of phonemic coding, discuss which open problems remain and then develop a model of our own that addresses some of these problems.

2 Previous work

2.1 Natural selection for combinatorial phonology

Several mathematical and computational models have shown that under noisy transmission, digital, combinatorial coding is more efficient than continuous coding. Nowak & Krakauer (1999) apply this insight in the context of the evolution of language, and derive an expression for the "fitness of a language". Imagine a population of individuals that all agree on which signals to use for which objects or events. The fitness of a language is now given by the expected success of a random individual to communicate about a random object or event with a random other individual. Nowak et al. show that when communication is noisy and when just a single sound is used for every meaning, the fitness is limited by an "error limit": only a limited number of sounds can be used — and thus a limited of meanings be expressed — because by using more sounds the successful recognition of the current signals would be impeded. Nowak et al. further show that in such noisy conditions, fitness is higher when (meaningless) sounds are combined into longer words. When the environment is combinatorial (i.e. objects and actions occur in many combinations) the fitness is highest when meaningful words are combined into longer sentences.

These results are essentially particular instantiations of Shannon's more general results on "noisy coding" (Shannon, 1948), as is explored in a later paper by the same group (Plotkin & Nowak, 2000). More interesting is the question how natural selection could favor a linguistic innovation in a population where that innovation is still very rare. Nowak & Krakauer (1999) do a game theoretic analysis of

"compositionality". They consider all mixed strategies where both holistic and compositional signals are used, and show that strategies that use more compositionality can invade strategies that use less. This means that the adaptive dynamics of languages under natural selection should lead to compositionality. For combinatorial phonology a similar analysis can be given.

Although this model is a useful formalization of the problem and gives some important insights, as an explanation for the evolution of phonemic coding and compositionality it is still insufficient. The main problem is that the model only considers the advantages of combinatorial strategies, and ignores two obvious disadvantages: (1) by having a "mixed strategy" individuals have essentially two languages in parallel, which one should expect to be costly because of memory and learning demands and additional confusion; (2) combinatorial signals that consist of two or more sounds take longer to utter and are thus more costly. A fairer comparison would be between holistic signals of a certain duration (where repetition of the same sound decreases the effect of noise) and combinatorial signals of the same duration (where the digital coding decreases the effect of noise). This is the approach we take in this paper.

2.2 Crystallization in the perception–imitation cycle

A completely different approach to phonemic coding is based on "categorical perception". Categorical perception (Harnad, 1987) is the phenomenon that categorization influences the perception of stimuli in such a way that differences between categories are perceived as larger and differences within categories as smaller that they really are (according to an "objective" similarity metric). For instance, infants already perceive phonemes as closer to the closest prototype phoneme from their native language than it is according to an "objective" (cross-linguistic) acoustical metric (Kuhl *et al.*, 1992). Hence, when presented vowels as stimuli ranging from /o/ to /a/ in fixed increments, British subjects will hear the first stimuli as o's or almost o's, and the last as a's or almost a's. Apparently, the frequency and position of acoustic stimuli gives rise to particular phoneme prototypes, and the prototypes in turn distort the perception.

Oudeyer (2002) studies a model that yields such a perceptual distortion. In this model, signals are modeled as points in an acoustic space, and are thus instantaneous. Oudeyer considers that signals survive from generation to generation because they are perceived and imitated. Oudeyer shows that categorical perception *shapes* a signal repertoire such that it conforms more and more to the prototype phonemes. Thus, emitted signals shape perception, and distorted perception shapes the repertoire of signals in the cycle from emission to perception to emission (the perception–imitation cycle; see also Westermann 2001 for a model of sensori-motor integration and its relevance for imitation and categorical perception). Oudeyer calls the collapse of signals in a small number of clusters "crystallization".

Oudeyer's model is fascinating, because it gives a completely non-adaptive mechanism for the emergence of phonemic coding. However, it is not clear how well it would work if signals have a time structure rather than being instantaneous[1]. Moreover, even if the mechanism works also in these conditions, it remains an important question whether phonemic coding increases the functionality of the language, and thus the fitness of the individual that uses it. If not, one would expect selection to work against it. In particular, in Oudeyer's model, where signals are instantaneous, a large repertoire of signals is collapsed into a small number of clusters. A functional pressure to maintain the number of distinct signals would thus have to either reverse the crystallization, or combine signals from different clusters. This aspect, which seems the core issue in understanding the origins of phonemic coding, is

[1] Oudeyer has also tested the model for sequences of sounds (Oudeyer, p.c.), but, as far as we know, not for continuous trajectories. It seems that in this version of the model the "combinatorial" aspects of phonemic coding is imposed and only the "categorical" (see section 2.3) aspect is emergent, such that our criticism still holds.

not modeled by Oudeyer. In our model, we ensure that the number of distinct signals remains at least at the same level; i.e. the functionality increases rather than decreases.

2.3 Aspects of phonemic coding

Other models of phonemic coding assume the sequencing of phonetic atoms into longer strings as given. They concentrate rather on the structure of the emerged systems (Lindblom *et al.*, 1984; de Boer, 2001; Redford *et al.*, 2001) or on how conventions on specific combinatorial signal systems can become established in a population through cultural transmission (Steels & Oudeyer, 2000). These models are interesting, and, importantly, bridge the gap with empirical evidence on how phonemic coding is implemented in the languages of the world.

It appears from this discussion that there are 4 related, but distinct aspects to phonemic coding:

1. Phonemically coded systems are *categorical*, in that they allow only a small number of basic sounds and not all feasible sounds in between;
2. they are also *superficially combinatorial*, in that all parts of each signal overlap with parts of other signals;
3. they are also *productively combinatorial*, in that the cognitive mechanism that produces and interprets signals uses the common parts of signals as building blocks that can be combined in all sorts of combinations;
4. the possible sets of categories and combinatorial rules show particular (cross-linguistic) constraints.

These aspects form a hierarchy, where the aspects further down the list imply the aspect above it. Oudeyer (2002) shows a non-adaptive mechanism that can yield aspect 1 (and gives a starting point for 4), but does not explain how the other aspects come about and how the functionality of the signal system is preserved. Nowak & Krakauer (1999) show how natural selection could favor 2, but ignore the temporal aspects of holistic signals. Zuidema & Hogeweg (2000) and Zuidema (2003) can be viewed as assuming aspects 1 and 2, and addressing the emergence of aspect 3 under natural selection and cultural evolution respectively (but the models are not discussed in these terms). Lindblom *et al.* (1984); de Boer (2001); Redford *et al.* (2001); Steels & Oudeyer (2000) all address aspect 4.

The question thus remains open as to under what circumstances a system of holistically coded signals with finite duration would change into a phonemically coded system of signals. In the paper we study a single mechanism that can yield aspects 1 and 2.

3 The model

In our model, we do not assume combinatorial structure, but rather study the gradual emergence of phonemic coding from initially holistic signals. We do take into account the temporal structure of both holistic and phonemically coded signals. We view signals as continuous movements (“gestures”, “trajectories”) through an abstract acoustic space. We assume that signals can be confused, and that the probability of confusion is higher if signals are more similar, i.e. closer to each other in the acoustic space according to some distance metric. We further assume that a functional pressure on distinctiveness maximizes the distance between trajectories.

3.1 Representing trajectories

The model is based on part-wise linear trajectories in a bounded 2-D continuous space (of size 15.0×15.0 in all simulations reported here). Trajectories are sequences of points with fixed length (here: 20). Each point has a fixed distance of 1.0 to the immediately preceding and following points in the sequence. The following and preceding points to a point can lay anywhere on a circle of radius one with that point at the center. Trajectories always stay within the bounds of the defined acoustic space.

Signals in the real world are continuous trajectories, but in the model we need to discretize the trajectories. However, to ensure that we do not impose the phonemic structure we are interested in, we discretize at a much finer scale than the phonemic patterns that will emerge. Hence, the points on a trajectory are not meant to model atomic units in a complex utterance.

3.2 Measuring distances

The distance between two trajectories t and r is defined as the sum of the distances between all corresponding points in the best possible alignment of the two trajectories. In finding the best possible alignment, one point from t can be mapped on several neighboring points in r and vice versa. In this way trajectories that resemble each other in shape, but that do not align perfectly still are considered close. This models the way humans perceive signals. The distances are calculated using "dynamic time warping", an efficient method that before the advent of statistical models, has been used with reasonable success in computer speech recognition (e.g. Sakoe & Chiba, 1978).

3.3 Maximizing the total mutual distance

In the first set-up of the model, we consider an idealized single repertoire of trajectories that, in a sense, repel each other. That is, the total distance between trajectories is optimized using a simple hill-climbing algorithm. The model goes through a large number of iterations. At every iteration, the sum of all mutual distances is calculated. Then a random change is applied to a random trajectory t, and the total distance is measured again. If this second measurement is larger than the first, the change is kept. If not, the change is reverted.

Random changes always respect the constraints on well-formed trajectories. Hence, a random point, t_x, is moved to a new random position (from a Gaussian distribution around the old position, provided it falls within the boundaries of the acoustic space). The two points on both sides of the moved point, t_{x+1} and t_{x-1}, are moved closer or further away such that the distance to t_x is again 1. The direction from t_x to t_{x+1} or t_{x-1} remains the same, unless the point would cross the boundary of the space, in which case it is rotated to the closest point within the boundary at distance 1 from t_x. The same procedure is applied recursively to the neighbors of t_{x+1} and t_{x-1} until the ends of the trajectory are reached.

In the second set-up of the model, we investigate what kind of repertoires of trajectories emerge in a *population* of agents that try to imitate each other in noisy conditions. The model is very similar, but now each agent in the population has its own repertoire, and it tries to optimize its own success in imitating and being imitated by other agents of the population.

This version of the model is like the imitation games of de Boer (2000). These only modeled holistic signals (vowels) and did not investigate phonemic coding. The game implemented here is a slight simplification of the original imitation game. First, all agents in the population are initialized with a random set of a fixed number of trajectories. Then for each game, a speaker is randomly selected from the population. This speaker selects a trajectory, and makes a random modification to it. Then it plays a number of imitation games (50 in all simulations reported here) with all other agents in

the population. In these games, the *initiator* utters the modified trajectory with additional noise. The *imitator* finds the closest trajectory in its repertoire and utters it with noise. Games are successful if the imitator's signals is closest to the modified trajectory in the initiator's repertoire. If it turns out that the modified trajectory has better imitation success than the original trajectory, the modified trajectory is kept, otherwise the original one is restored.

4 Results

4.1 Optimizing a single repertoire

We ran the model under the single repertoire condition with a number of different parameters. In all simulations the initial trajectories are random sequences of positions, where the only constraints are that neighboring points are at distance 1 from each other and that all points are within the permitted space.

In simulations with few trajectories (up to 4), we find that the trajectories "bunch-up" and remain within a very small area at maximum distance from the areas used by the other trajectories. Each of these signals is thus a holistic signal, but the signals are "categorical".

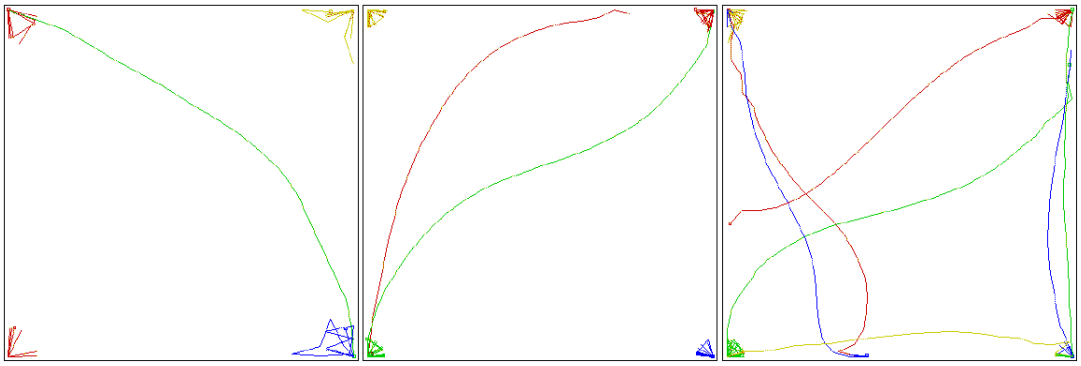


Figure 1: Comparison of optimized systems of 5, 6 and 10 trajectories. Note the reuse of start- and endpoints (squares indicate start points).

In simulations with 5 trajectories, 4 occupy the corners areas of the acoustic space in the same way as in simulations with 4 trajectories. However, the fifth trajectory stretches from one corner to another, and thus shares the areas for its begin and end points with two different other trajectories (see fig. 1, leftmost panel). This can be interpreted as a rudimentary phonemic code.

With more trajectories, the reuse of beginning and end points becomes more pronounced. In the simulation with 6 trajectories, the first 5 are similarly organized, but the sixth is essentially the inverse of the fifth. In the simulation with 10 trajectories, 3 trajectories are still bunched up in a small area of the acoustic space, but the other 7 are stretched out, sharing begin and end points with one another. Frequently one can find trajectories that are more or less the inverse of another trajectory.

In order to perform a statistical analysis, a numerical measure of the extent to which emerged systems were phonemically coded had to be defined. This measure, called the phonemicity $\mathcal{P}$, is defined as the ratio between the average distance between the start and end points of all trajectories and the average distance between all other corresponding points of all trajectories. Corresponding points are defined as points that are an equal number of steps away from either a start or an end point

(i.e., two points that are at position 3 are corresponding points, but so are a point at position 3 and position L-2). The details for this measure are in the appendix.

In a phonemically coded system of trajectories, start and end points are expected to be closer together than the other points on a trajectory, while in a holistically coded system of trajectories, the average distance is expected to be approximately equal. Therefore, the measure should give lower values for phonemically coded systems. It is quite likely that better measures of phonemicity can be defined, but this measure does make a distinction between holistically and phonemically coded systems, and was therefore adopted for the analysis.

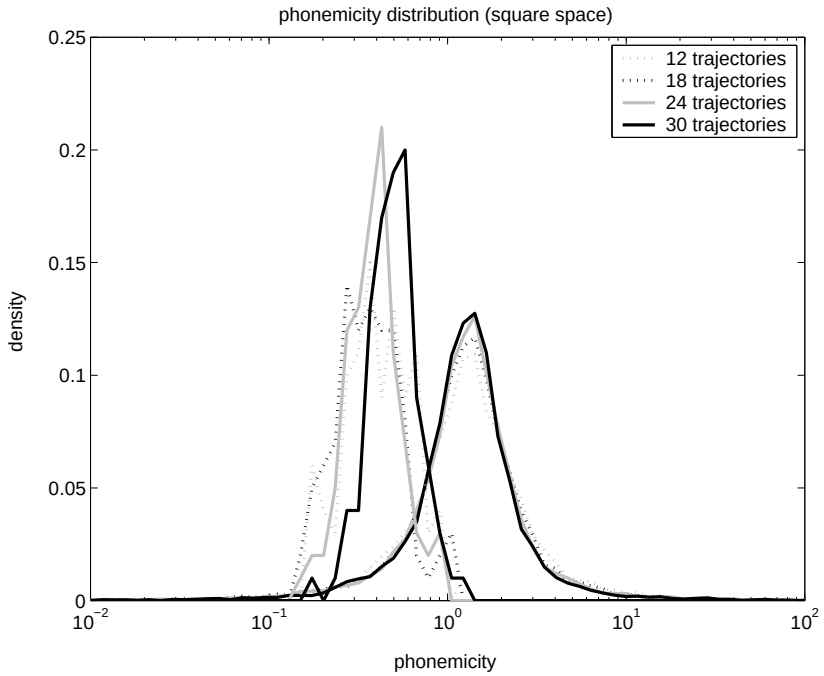


Figure 2: Distributions of phonemicity of random systems (right peak) and optimized systems (left peak). Note that the phonemicity measure for optimized systems is lower, indicating that the optimized systems are more phonemically coded than random systems.

Two conditions were compared. In both conditions the systems of trajectories were initialized randomly; only in the second condition were systems of trajectories optimized for distance first using 30,000 optimization steps. The results were measured for systems of many different sizes, but are presented for systems of 12, 18, 24 and 30 trajectories in figure 2. 10,000 random systems were evaluated, but for computational reasons only 100 optimized systems, as the amount of computation needed for optimization precluded larger numbers of systems to be evaluated. Note that the horizontal axis (showing the phonemicity) is logarithmic. This has the advantage of both making the peaks more distinct and making the distributions more similar to the normal distribution. When using the t-test, on both the phonemicity and its logarithm, it turns out the difference between the distribution of the random systems and the optimized systems is significant with $p < 0.05$ (the t-test is less appropriate for the non-log measure, because of the highly skewed distribution).

This result indicates that optimization for acoustic time-warped distance between trajectories results in more phonemically coded systems.

4.2 Optimizing repertoires in a population

For vowel systems, it has been shown that optimizing a single repertoire leads to similar systems as a population-optimization system (compare de Boer, 2000; Liljencrants & Lindblom, 1972). It can be shown that for trajectories the same is true, under the condition that noisy distortions of trajectories do not distort the shape of these trajectories too much. This is illustrated in figure 3.

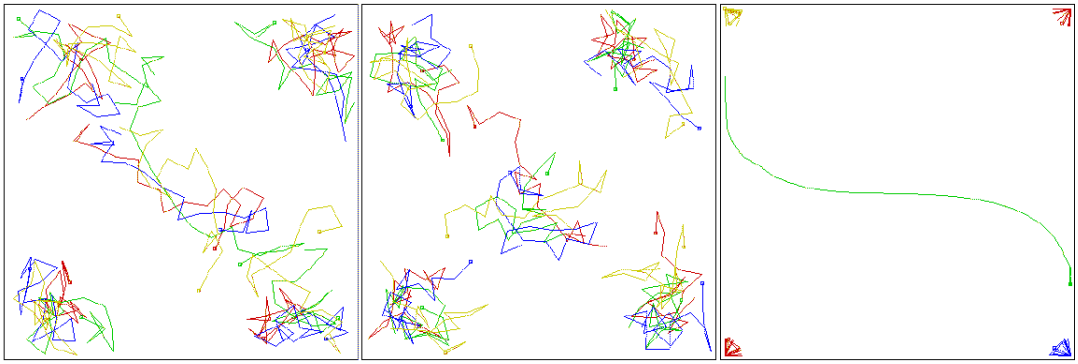


Figure 3: From left to right: emerged system with five trajectories in a population of ten agents (four agents shown), emerged system with five trajectories and uncorrelated noise, and optimized system of five trajectories. Small squares indicate the starting point of trajectories.

In this figure the left frame shows the system of five trajectories that resulted from playing imitation games in a population, using form-preserving noise. The right frame shows a system of five trajectories that resulted from optimizing distance. It can be observed that in both cases, the corners are populated by four trajectories, which are bunched up. The fifth trajectory, in contrast, follows the diagonal. As before, an analysis in terms of phonemes suggests itself: the four corners are basic phonemes, while the fifth trajectory uses one as the corners as a starting phoneme and the opposite corner as the ending phoneme. Both models result in similar systems of trajectories.

The middle frame, on the other hand, shows that when noise does not preserve shape of trajectories, a system results in which all trajectories are bunched up and an analysis in terms of phonemes is therefore not possible. As noise in real signals is band limited, it follows that shape will always be preserved to some extent. Therefore the shape-preserving model is indeed the correct model. Instead of investigating computationally extremely costly population models, it is therefore possible to investigate emergence of phonemic coding using the optimization model. For computational reasons, we have not performed simulations in the population condition with more than 5 trajectories.

5 Conclusion

We have investigated whether systems of trajectories that are used for imitation in a population would tend toward phonemic coding when agents tried to maximize their imitation success. It was found that running simulations of populations directly was too time-consuming. However, it was also found that direct optimization of time-warped distance between trajectories resulted in systems of trajectories that were similar to those found in preliminary experiments with imitation in a population. For this to be true, it was necessary to assume that in the population case, shape of trajectories was preserved under noise. This is a realistic assumption, as it turns out to be true for all noise that is band-limited, i.e. for which the energy of higher frequencies tends to zero. This is the case for all real-world noise.

When systems of trajectories were optimized for time-warped distance, it turned out that start- and endpoints were reused and that there were no trajectories (at least for limited numbers of trajectories) that had the same start- and endpoint and that only differed in the shape of the trajectory in between. This is indicative of phonemic coding. A measure of phonemicity was defined and it was found that optimized systems had significant lower values for this measure than random systems, indicating that they were more phonemically coded.

The conclusion to be drawn from this is that systems of complex articulations (trajectories) that have maximum distance to each other tend to show aspects of phonemic coding. Systems that have trajectories that are maximally distant from each other are most robust to noise. This means that optimizing systems of large numbers of complex articulations for robustness to noise, which is likely to happen when they are used for communication in a population, would result in systems of trajectories that can be analyzed in terms of phonemes.

The relevance for the evolution of speech is clear. When populations of agents start to communicate using small numbers of signals, it is unlikely that they would use phonemic coding, or be able to use it if it occurred. However, when extending the number of signals, the most robust systems would be the ones that can be analyzed as phonemically coded. Agents that have adaptations to detect and use this property would have an evolutionary advantage, as they would be able to learn faster, and probably to communicate more accurately as well. This provides a cultural beginning of a possible biological adaptation for using phonemically coded signals. This adaptation in the area of speech could later be exapted for use in combining words, in other words, for syntax.

6 Future work

The results described in this paper are preliminary, and need to be extended in several ways. Firstly, the model, especially in the population condition, should be tested with larger number of trajectories, and with trajectories of longer length. Presumably, the “phonemic coding” would then not just apply to the start and end points of the trajectories. Consequently, another measure of phonemicity needs to be defined.

Further, the model can be altered such that it allows trajectories of varying length in a single repertoire, and perhaps varying distances between the points of a trajectory.

Finally, and most ambitiously, the model should be extended to incorporate the aspects of phonemic coding that are currently not addressed: productive combinatorics and realistic constraints on the categories and rules of combination.

References

DE BOER, B. (2000). Self organization in vowel systems. *Journal of Phonetics* **28**, 441–465.

DE BOER, B. (2001). *The origins of vowel systems*. Oxford, UK: Oxford University Press.

HARNAD, S. (1987). *Categorical Perception*. Cambridge, UK: Cambridge University Press.

JACKENDOFF, R. (2002). *Foundations of Language*. Oxford, UK: Oxford University Press.

KUHL, P., WILLIAMS, K., LACERDA, F., STEVENS, K. & LINDBLOM, B. (1992). Linguistic experience alters phonetic perception in infants by 6 month of age. *Science* **255**, 606–608.

LILJENCRANTS, J. & LINDBLOM, B. (1972). Numerical simulations of vowel quality systems: the role of perceptual contrast. *Language* **48**, 839–862.

LINDBLOM, B., MACNEILAGE, P. & STUDDERT-KENNEDY, M. (1984). Self-organizing processes and the explanation of language universals. In: *Explanations for language universals* (Butterworth, M., Comrie, B. & Dahl, ., eds.), pp. 181–203. Berlin: Walter de Gruyter & Co.

NOWAK, M. A. & KRAKAUER, D. C. (1999). The evolution of language. *Proc. Nat. Acad. Sci. USA* **96**, 8028–8033.

OUDEYER, P.-Y. (2002). Phonemic coding might be a result of sensory-motor coupling dynamics. In: *Proceedings of the 7th International Conference on the Simulation of Adaptive Behavior* (Hallam, B., Floreano, D., Hallam, J., Hayes, G. & Meyer, J.-A., eds.), pp. 406–416. Cambridge, MA: MIT Press.

PLOTKIN, J. B. & NOWAK, M. A. (2000). Language evolution and information theory. *Journal of Theoretical Biology* pp. 147–159.

REDFORD, M. A., CHEN, C. C. & MIIKKULAINEN, R. (2001). Constrained emergence of universals and variation in syllable systems. *Language and Speech* **44**, 27–56.

SAKOE, H. & CHIBA, S. (1978). Dynamic programming optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* **26**, 43–49.

SHANNON, C. E. (1948). A mathematical theory of communication. *The Bell Systems Technical Journal* **27**, 379–423 and 623–656.

STEELS, L. & OUDEYER, P.-Y. (2000). The cultural evolution of syntactic constraints in phonology. In: *Proceedings of the VIIth Artificial life conference (Alife 7)* (Bedau, M. A., McCaskill, J. S., Packard, N. H. & Rasmussen, S., eds.). Cambridge (MA): MIT Press.

WESTERMANN, G. (2001). A model of perceptual change by domain integration. In: *Proceedings of the 23d Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

ZUIDEMA, W. (2003). How the poverty of the stimulus solves the poverty of the stimulus. In: *Advances in Neural Information Processing Systems 15 (Proceedings of NIPS'02)* (Becker, S., Thrun, S. & Obermayer, K., eds.). Cambridge, MA: MIT Press. (forthcoming).

ZUIDEMA, W. & HOGEWEG, P. (2000). Selective advantages of syntactic language: a model study. In: *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society* (Gleitman & Joshi, eds.), pp. 577–582. Mahwah, NJ: Lawrence Erlbaum Associates.

Appendix: Measuring phonemicity

The average distance $\mathcal{E}$ between the extreme points (start and end points) is given by:

$$\mathcal{E} = \frac{1}{2N(N-1)} \sum_{i=0}^{N} \sum_{j=i+1}^{N} \left(D\left(i_1, j_1\right) + D\left(i_1, j_L\right) + D\left(i_L, j_1\right) + D\left(i_L, j_L\right) \right) \tag{1}$$

where N is the number of trajectories and L is the length of each trajectory. The function D is a measure of distance between points and will be explained below. The average distance between all other corresponding points is given by:

$$\mathcal{C} = \frac{1}{N(N-1)(L-2)} \sum_{i=0}^{N} \sum_{j=i+1}^{N} \sum_{k=2}^{L-1} \left(D\left(i_k, j_k\right) + D\left(i_{L-k+1}, j_k\right) \right) \tag{2}$$

The phonemicity $\mathcal{P}$ is then simply:

$$\mathcal{P} = \frac{\mathcal{E}}{\mathcal{C}} \tag{3}$$

The distance function $D(i_a, j_b)$ is the inverse, squared Euclidean distance between point a of trajectory i and point b of trajectory j:

$$D\left(i_a, j_b\right) = \frac{1}{\epsilon + |p_a(i) - p_b(j)|^2} \tag{4}$$

where $p_a(i)$ is the position of point a of trajectory i. The term ϵ ($\epsilon = 0.01$ throughout this paper) is added to avoid division by zero. Note that this is a different distance function than was used in the optimization of the distances between trajectories.

Learning biases and language evolution

Kenny Smith*
Language Evolution and Computation Research Unit,
School of Philosophy, Psychology and Language Sciences,
The University of Edinburgh,
Adam Ferguson Building, 40 George Square, Edinburgh EH8 9LL

http://www.ling.ed.ac.uk/∼kenny

Abstract

Structural hallmarks of language can be explained in terms of adaptation, by language, to pressures arising during its cultural transmission. Here I present a model which explains the compositional structure of language as an adaptation in response to pressures arising from the poverty of the stimulus available to language learners and the biases of language learners themselves.

1 Introduction

The goal of evolutionary linguistics is to explain the origins and development of human language — how did language come to be structured as it is? Recent research attempts to answer this question by appealing to cultural evolution (Batali 2002; Brighton 2002; Kirby 2002). Language is culturally transmitted to the extent that language learners acquire their linguistic competence on the basis of the observed linguistic behaviour of others. A key contribution of those working within the cultural framework is to show that the cultural transmission of language leads to an adaptive dynamic — the adaptation, by language itself, to pressures acting during its cultural transmission. This cultural evolution can lead to the emergence of at least some of the characteristic structure of language.

This process of cultural evolution must be dependent on some biological endowment. What is not clear is what form this endowment takes, or to what extent it is language-specific. In this paper I will present a series of experiments, using a computational model of the cultural transmission of language, which allow us to refine our understanding of the necessary biological basis for a particular structural characteristic of language, compositionality. In this model a learner's biological endowment consists of a particular way of learning, with an associated learning bias.

2 Elements of the model

I will present an Iterated Learning Model (ILM) which allows us to investigate the role of stimulus poverty and learning bias in the evolution of compositional language. The ILM is based around a simple treatment of languages as a mapping between meanings and signals (see Section 2.1). Linguistic agents are modelled using associative networks (see Section 2.2). These agents are slotted

*The author is supported by ESRC Research Grant No. R000223969.

into a minimal population model to yield the ILM. For the purposes of this paper, we will consider an ILM with the simplest possible population dynamic — the population consists of a set of discrete generations, with each generation consisting of a single agent. The agent at generation n produces some observable behaviour (in this model a set of meaning-signal pairs), which is then learned from by the agent at generation $n + 1$.

2.1 Compositionality and a model of languages

Compositionality relates semantic structure to signal structure — in a compositional system the meaning of an utterance is dependent on the meaning of its parts. For example, the utterance "John walked" consists of two words, a noun ("John") and a verb ("walked"), which further consists of a stem ("walk-") and a suffix ("-ed"). The meaning of the utterance as a whole is dependent on the meaning of these individual parts. In contrast, in a non-compositional or *holistic* system the signal as a whole stands for the meaning as a whole. For example, the meaning of the English idiom "bought the farm" (meaning died) is not a function of the meaning of its parts.

The simplest way to capture this is to treat a language as a mapping between a space of meanings and a space of signals. In a compositional language, this mapping will be neighbourhood-preserving. Neighbouring meanings will share structure, and this shared structure will result in shared signal structure — neighbouring meanings in the meaning space will map to neighbouring signals in signal space. Holistic mappings are not neighbourhood-preserving — since the signal associated with a meaning does not depend on the structure of that meaning, shared structure in meaning space will not map to shared signal structure, unless by chance.

For the purposes of this model, meanings are treated as vectors and signals are strings of characters. Meanings are vectors in some F-dimensional space, where each dimension takes V possible values. F and V therefore define a meaning space $\mathcal{M}$.[1] The world, which provides communicatively relevant situations for agents in the model, consists of a set of objects, where each object is labelled with a meaning drawn from the meaning space $\mathcal{M}$.[2] Signals are strings of characters of length 1 to l_{max}, where characters are drawn from the character alphabet Σ.[3] l_{max} and Σ therefore define a signal space $\mathcal{S}$.

Given these representations of meanings and signals, we can now define a measure of compositionality. This measure is designed to capture the notion given above, that compositional languages are neighbourhood-preserving mappings between meanings and signals, and is based on a measure introduced in Brighton (2000). Compositionality (c) is based on the meaning-signal pairs that an agent produces, and is the Pearson's Product-Moment correlation coefficient of the pairwise distances between all the meanings and the pairwise distances between their corresponding signals.[4] $c = 1$ for a perfectly compositional system and $c \approx 0$ for a holistic system.

[1] The structure of this meaning space has been shown to have consequences for the cultural evolution of compositional structure (Brighton 2002). However, I will not vary this parameter. All results reported here are for the case where $F = 3$ and $V = 5$.

[2] All results presented here are for the case where the world contains 31 objects, each object is labelled with a distinct meaning, and those meanings are drawn from a hypercube subspace of the space of possible meanings — a *structured* world, in the terms of Smith *et al.* (forthcoming).

[3] For the results reported here, $l_{max} = 3$ and $\Sigma = \{a, b, c, d, e, f, g, h, i, j\}$.

[4] Distance in the meaning space is measured using Hamming distance. Distance in the signal space is measured using Levenstein (string edit) distance.

2.2 A model of a linguistic agent

We now require a model of a linguistic agent capable of manipulating such systems of meaning-signal mappings. I will describe an associative network model of a linguistic agent. This model is based upon a simpler model of a linguistic agent, used to investigate the cultural evolution of vocabulary systems (Smith 2002a). The main advantage of this model is that it allows the biases of language learners to be manipulated and investigated. For full details of the network model, the reader is referred to Smith (2003), Chapter 5.[5]

Representation Agents are modelled using networks consisting of two sets of nodes $\mathcal{N}_M$ and $\mathcal{N}_S$ and a set of bidirectional connections $\mathcal{W}$ connecting every node in $\mathcal{N}_M$ with every node in $\mathcal{N}_S$. Nodes in $\mathcal{N}_M$ represent meanings and partial specifications of meanings, while nodes in $\mathcal{N}_S$ represent partial and complete specifications of signals.

As summarised above, each meaning is a vector in F-dimensional space where each dimension has V values. *Components* of a given meaning are (possibly partially specified) vectors, with each feature of the component either having the same value as the meaning, or a wildcard. Similarly, components of a signal of length l are (possibly partially specified) strings of length l. Each node in $\mathcal{N}_M$ represents a component of a meaning, and there is a single node in $\mathcal{N}_M$ for each component of every possible meaning. Similarly, each node in $\mathcal{N}_S$ represents a component of a signal and there is a single node in $\mathcal{N}_S$ for each component of every possible signal.

Learning During a learning event, a learner observes a meaning-signal pair $\langle m, s \rangle$. The activations of the nodes corresponding to all possible components of m and all possible components of s are set to 1. The activations of all other nodes are set to 0. The weights of the connections in $\mathcal{W}$ are adjusted according to some weight-update rule. In Section 4 this weight-update procedure will be a parameter of variation. However, initially, we will consider the rule

$$\Delta W_{xy} = \begin{cases} +1 & \text{if } a_x = a_y = 1 \\ -1 & \text{if } a_x \neq a_y \\ 0 & \text{if } a_x = a_y = 0 \end{cases} \quad (1)$$

where $W_{xy} \in \mathcal{W}$ gives the weight of the connection between nodes x and y and a_x gives the activation of node x. The learning procedure is illustrated in Fig. 1 (a).

Production During the process of producing utterances, agents are prompted with a meaning and required to produce a meaning-signal pair. Production proceeds via a winner-take-all process. An *analysis* of a meaning or signal is an ordered set of components which fully specifies that meaning or signal. In order to produce a signal for a given meaning $m_i \in \mathcal{M}$, every possible signal $s_j \in \mathcal{S}$ is evaluated with respect to m_i. For each of these possible meaning-signal pairs $\langle m_i, s_j \rangle$, every possible analysis of m_i is evaluated with respect to every possible analysis of s_j. The evaluation of a meaning analysis-signal analysis pair depends on the weighted sum of the connections between the relevant nodes. The meaning-signal pair which yields the analysis pair with the highest weighted sum is returned as the network's production for the given meaning. The production process is illustrated in Fig. 1 (b).

[5] Available for download at http://www.ling.ed.ac.uk/∼kenny/thesis.html

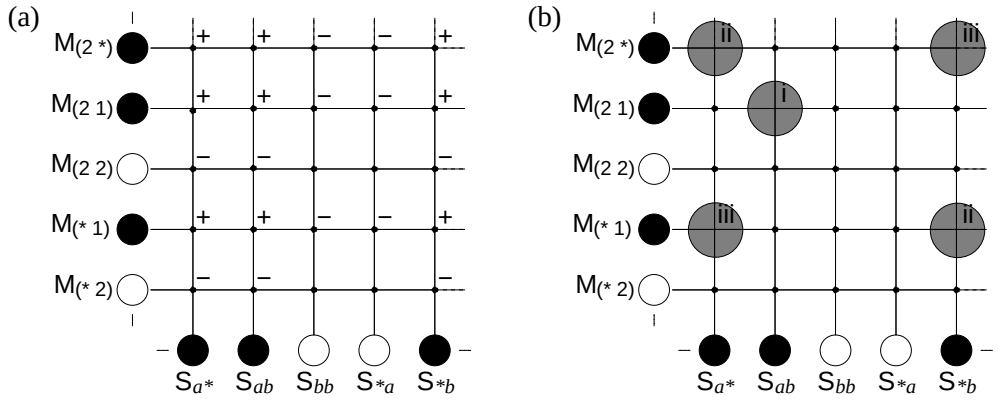


Figure 1: (a) Learning the meaning-signal pair $\langle (2\ 1), ab \rangle$. Nodes are represented by large circles and are labelled with the component they represent. For example, $M_{(2\ *)}$ is the node which represents the meaning component $(2\ *)$, where $*$ is an unspecified feature value. Nodes with an activation of 1 are represented by large filled circles. Small filled circles represent weighted connections. During the learning process, nodes representing components of $(2\ 1)$ and ab have their activations set to 1. Connection weights are then either incremented $(+)$, decremented $(-)$ or left unchanged. (b) Retrieval of three possible analyses of $\langle (2\ 1), ab \rangle$. The relevant connection weights are highlighted in grey. The weight for the one-component analysis $\langle \{(2\ 1)\}, \{ab\} \rangle$ depends on the weight of the connection between the nodes representing the components $(2\ 1)$ and ab, marked as i. The weight for the two-component analysis $\langle \{(2\ *), (*\ 1)\}, \{a*, *b\} \rangle$ depends on the weighted sum of two connections, marked as ii. The weight of the alternative two-component analysis $\langle \{(2\ *), (*\ 1)\}, \{*b, a*\} \rangle$ is given by the weighted sum of the two connections marked iii.

3 A familiar result

I will begin by replicating a familiar result: the emergence of compositional structure through cultural processes depends on the presence of a *transmission bottleneck* (Brighton 2002; Kirby 2002). Recall that a learner in the model acquires their linguistic competence on the basis of a set of observed meaning-signal pairs. That set of meaning-signal pairs is drawn from the linguistic behaviour of some other individual, which is a consequence of that individual's linguistic competence. I will investigate two possible conditions. In the *no transmission bottleneck* condition, this set of observed meaning-signal pairs contains examples of the signal associated with every possible meaning, and each learner is therefore presented with the complete language of the agent at the previous generation. In the *transmission bottleneck* condition, the set of observed behaviour does not contain examples of the signal associated with every meaning, therefore each learner is presented with a subset of the language of the agent at the previous generation.[6] The transmission bottleneck constitutes one aspect of the poverty of the stimulus problem faced by language learners — they must acquire knowledge of a large (or infinite, in the real-world case) language on the basis of exposure to a subset of that language.

In both conditions, the initial agents in each simulation run have all their connections weights set to 0, and therefore produce every meaning-signal pair with equal probability. Subsequent agents have connection weights of 0 prior to learning. Runs were allowed to progress for a fixed number of generations (200).[7] Figs. 2 (a) and (b) plot compositionality by frequency for the initial and final

[6]For all simulations involving a transmission bottleneck described in this paper, the number of utterances produced by agents was set so that language learners observed approximately 60% of the language of the previous agent.

[7]In the no bottleneck condition, the system of meaning-signal mappings is stable long before this point, in the sense that agents at generation n and $n+1$ produce identical sets of meaning-signal pairs. Absolute stability is impossible when there

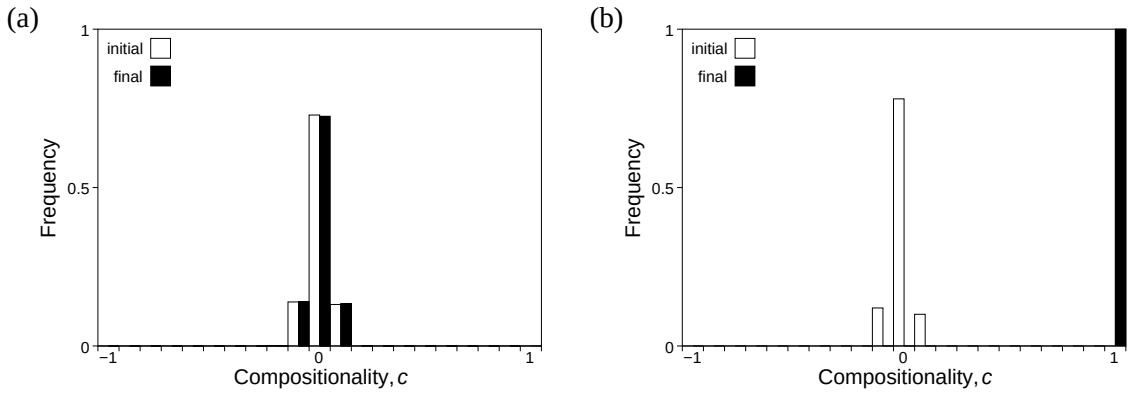


Figure 2: The impact of the transmission bottleneck. (a) gives frequency by compositionality for runs in the no bottleneck condition — both the initial and final systems are holistic. (b) gives frequency by compositionality for runs where there is a bottleneck on transmission — while the initial systems are again holistic, the final systems are all highly compositional.

languages, for the no bottleneck and bottleneck conditions respectively.[8]

As can be seen from the figure, when there is no bottleneck on transmission there is no cultural evolution and compositional languages do not emerge. In contrast, when there is a bottleneck on transmission highly compositional systems emerge with high frequency — cultural evolution leads to the emergence of compositional language from initially holistic systems. This confirms, using a rather different model of a language learner, previously established results (Brighton 2002; Kirby 2002).

In the absence of a transmission bottleneck, the initial, random assignment of signals to meanings can simply be memorised. Consequently, there is no pressure for compositionality and the holistic mapping embodied in the initial system persists. However, holistic systems cannot survive in the presence of a bottleneck. The meaning-signal pairs of a holistic language have to be observed to be reproduced. If a learner only observes a subset of the holistic language of the previous generation then certain meaning-signal pairs will not be preserved — the learner, when called upon to produce, will produce some other signal for that meaning, resulting in a change in the language. In contrast, compositional languages are generalisable, due to their structure, and remain relatively stable even when the learner observes a subset of the language of the previous generation. Over time, language adapts to this pressure to be generalisable. Eventually, the language becomes highly compositional, highly generalisable and consequently highly stable.

4 Exploring learning biases

To what extent is this fundamental result, that the transmission bottleneck leads to a pressure for compositional language, dependent on the model of a language learner? There is indirect evidence that this result is to some extent independent of the model of a language learner — a wide range of learning models all produce this fundamental result (Hurford 2000; Batali 2002; Kirby 2002; Kirby

is a bottleneck on cultural transmission — depending on the sample of observations each learner receives, an apparently stable system can change at any time. However, the *distribution* of systems is stable after 200 generations — allowing the simulation runs to proceed for longer gives the same result.

[8]The results for the no bottleneck condition are based on 1000 independent runs of the ILM. The results for the bottleneck condition are based on 100 runs — fewer runs are required as there is less sensitivity to initial conditions.

& Hurford 2002). However, do these models share a common element? Is there some *learner bias* common across all these models which is required for compositional language to evolve culturally?

In order to investigate this question, further experiments were carried out, in which the parameter of interest is the weight-update rule used to adjust network connection weights during learning. Different ways of adjusting connection weights will potentially lead to different learning biases — different ways of changing weights will make certain systems easier or harder to learn than others. The general form of the weight-update rule is as follows:

$$\Delta W_{xy} = \begin{cases} \alpha & \text{if } a_x = a_y = 1 \\ \beta & \text{if } a_x = 1 \wedge a_y = 0 \\ \gamma & \text{if } a_x = 0 \wedge a_y = 1 \\ \delta & \text{if } a_x = a_y = 0 \end{cases} \tag{2}$$

For the results described in the previous Section, $\alpha = 1$, $\beta = \gamma = -1$, $\delta = 0$. I will now consider a wider range of weight-update rules, restricting myself to rules where $\alpha, \beta, \gamma, \delta \in \{-1, 0, 1\}$. This yields a set of $3^4 = 81$ possible weight-update rules. In order to ascertain the biases of the different weight-update rules, each weight-update rule is subjected to three tests:[9]

Acquisition test: Can an isolated agent using the weight-update rule acquire a perfectly compositional language, based on full exposure to that language? To evaluate this, an agent using the weight-update rule was trained on a predefined perfectly compositional ($c = 1$) language, being exposed once to every meaning-signal pair included in that language. The agent was judged to have successfully acquired that language if it could reproduce the meaning-signal pairs of the language in production and reception.

Maintenance test: Can a population of agents using the weight-update rule maintain a perfectly compositional language over time in an ILM, when there is a bottleneck on transmission? To evaluate this, 10 runs of the ILM were carried out for the weight-update rule, with the agent in the initial generation having their initial connection weights set so as to produce a perfectly compositional language. Populations were defined as having maintained a compositional system if c remained above 0.95 for every generation of ten 200 generation runs.

Construction test: Can a population of agents using the weight-update rule construct a highly compositional language from an initially random language, when there is a bottleneck on transmission (as happened in the results outlined in the previous Section)? To evaluate this, 10 runs of the ILM were carried out for the weight-update rule, with the agent in the initial generation having initial connection weights of 0 and therefore producing a random set of meaning-signal pairs. Populations were defined as having constructed a compositional system if c rose above 0.95 in each of ten 200 generation runs.

The results of these sets of experiments are summarised in Table 1. Only a limited number of weight-update rules (two of 81) support the evolution of compositional language through cultural processes. Why? What is it about the assignment of values to the variables α, β, γ and δ in these rules that make them capable of acquiring, maintaining and constructing a compositional system?

A full analysis is somewhat involved, and I will simply summarise the key point here — for full details the reader is referred to Smith (2003). The two weight-update rules which pass the acquisition,

[9] A similar technique has been applied to the investigation of learning biases required for the cultural evolution of functional vocabulary systems (Smith 2002a).

Test result			Number of rules
Acquire?	Maintain?	Construct?	
no	no	no	63
yes	no	no	16
yes	yes	yes	2

Table 1: Summary of the results of the three tests. The table gives the three types of performance exhibited, and the number (out of 81) of weight-update rules fitting that pattern of performance.

maintenance and construction tests satisfy three conditions: 1) $\alpha > \beta$; 2) $\delta > \gamma$; 3) $\alpha > \delta$. These two rules[10] are the only weight-update rules from the sample of 81 which satisfy these conditions. By returning to the network and examining the way in which connection weights change on the basis of exposure to individual meaning-signal pairs, we can identify the consequences of these restrictions.

1. $\alpha > \beta$ ensures that, if an agent is exposed to the meaning-signal pair $\langle m_i, s_j \rangle$, they will in future tend to prefer produce s_j when presented with m_i, rather than $s_{k \neq j}$.

2. $\delta > \gamma$ ensures that, if an agent is exposed to $\langle m_i, s_j \rangle$, they will prefer *not* to produce s_j when presented with $m_{k \neq i}$.

3. $\alpha > \delta$ ensures that, if an agent is exposed to $\langle m_i, s_j \rangle$, they will tend to reproduce this meaning-signal pair in a manner which involves the maximum number of components.

Points 1 and 2 in combination lead to a preference for *one-to-one* mappings between meanings and signals — agents with the appropriate weight-update rules are biased in favour of learning languages which map each meaning to a constant signal (one-to-many mappings are avoided, see Point 1), and which map each distinct meaning onto a distinct signal (many-to-one mappings from meanings to signals are avoided, see Point 2). Point 3 corresponds to a bias in favour of memorising associations between elements of meaning and elements of signal, rather than between whole meanings and whole signals. This tendency to exploit regularities is presumably a general property of learning devices which are capable of generalising beyond their training data.

5 The learning bias elsewhere

How important are these two elements of bias? They are evident in all other models of the cultural evolution of linguistic structure, as a consequence of a learner preference for extracting meaningful, recurring chunks from the utterances they observe, coupled with production and learning constraints, as summarised in Table 2. This suggests that the two components of bias (a bias in favour of one-to-one mappings between meanings and signals, and a bias in favour of exploiting regularities in the meaning-signal mapping) are a prerequisite for the cultural evolution of compositional structure.

This then constitutes a testable hypothesis: if we believe that compositional language evolved in humans through cultural processes, we should expect that human language learners bring these two biases to the language acquisition task. I assume that the ability to extract regularities, and thus learn

[10] To be explicit, the two rules are:
$\alpha = 1, \quad \beta = -1, \quad \gamma = -1, \quad \delta = 0$
and $\alpha = 1, \quad \beta = 0, \quad \gamma = -1, \quad \delta = 0.$

Paper	Learning model	Structure emerges?	Against synonymy?	Against homonymy?
Hare & Elman (1995)	NN (m→s)	no	yes (architecture)	no (architecture)
Batali (1998)	NN (s→m)	yes	yes (deterministic production)	yes (architecture)
Kirby & Hurford (2002)		yes	yes (deterministic production)	yes (architecture)
Hurford (2000)	rule induction	yes	yes (deterministic production)	? (but homonymy unlikely)
Kirby (2002)		yes	yes (deterministic production)	yes (no learning of homonyms)
Batali (2002)	exemplar induction	yes	yes (cost reduction for reused expressions)	yes (cost increase for homonyms)

Table 2: Summary of the biases in models of the cultural evolution of linguistic structure, organised by learner model (NN = neural network, m→s = mapping from meanings to signals, s→m = mapping from signals to meanings). All models which lead to the emergence of structure build in biases against synonymy and homonymy, either during learning or production. See Smith (2002) for an explanation of the biases of different network architectures.

mappings from parts of meanings to parts of signals, is present in humans, and probably other species besides. Additionally, there is evidence that human language learners bring a one-to-one bias to the language acquisition task, at all levels of linguistic structure.

It has been proposed that, as a general principle, human language learners have a preference for one-to-one mappings between underlying meaning and surface form. This has been termed variously as a *maxim of clarity* (Slobin 1977), a preference for *transparency* (Langacker 1977), or a bias in favour of *isomorphism* (Haiman 1980). Table 3 summarises some of the relevant literature on the biases which human language learners bring to the acquisition of morphological, lexical and syntactic systems. As can be seen from this table, various authors have proposed that children bring biases against one-to-many and many-to-one mappings to the acquisition of all levels of linguistic structure — human language learners appear to possess a bias in favour of one-to-one mappings between meanings and surface forms.

Level	Study		
	Paper	Method	Conclusion
morphological	Mańczak (1980)	etymological dictionary survey	Bias against synonymy: paradigms lose synonymous morphemes.
	Slobin (1977)	observation	Bias against homonymy: widespread homonymy contributes to difficulty of acquiring inflectional system in Serbo-Croat.
lexical	Markman & Wachtel (1988)	experimental	Bias against synonymy: each object will have only one label.
	Macnamara (1982)	observation	Bias against homonymy: children avoid cross-categorial homonyms.
syntactic	Pinker (1984)	theoretical	Bias against synonymy: each deep structure maps to a single surface structure.
	Bever & Langendoen (1971)	theoretical/ historical	Bias against homonymy: change in OE relative clause structure due to avoidance of ambiguous constructions.

Table 3: Summary of the literature on biases of human language learners, organised according to level of linguistic representation.

6 Conclusions

I have presented an Iterated Learning Model of the cultural evolution of compositional structure. This model has been used to replicate a familiar result — the poverty of the stimulus available to language learners (as imposed by the transmission bottleneck) leads to the emergence of compositional structure. However, novelly, this cultural evolution has been shown to be dependent on language learners possessing two biases:

1. a bias in favour of one-to-one mappings between meanings and signals.
2. a bias in favour of exploiting regularities in the input data, by acquiring associations between parts of meanings and parts of signals.

Both these biases are present in most computational models of the evolution of linguistic structure. Significantly, there is also evidence to suggest that human language learners bring these biases to the language acquisition task. Compositionality, a fundamental structural property of language, can therefore be explained in terms of cultural evolution in response to two pressures — a pressure arising from the poverty of the stimulus, and a pressure arising from the biases of language learners. The source of this learning bias in humans is a topic for further research — is the bias a consequence of some general cognitive strategy, or a specific biological adaptation for the acquisition of language?

References

BATALI, J. 1998. Computational simulations of the emergence of grammar. In *Approaches to the Evolution of Language: social and cognitive bases*, ed. by J. R. Hurford, M. Studdert-Kennedy, & C. Knight, 405–426. Cambridge: Cambridge University Press.

—— 2002. The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In Briscoe (2002), 111–172.

BEVER, T. G., & D. T. LANGENDOEN. 1971. A dynamic model of the evolution of language. *Linguistic Inquiry* 2.433–463.

BRIGHTON, H. 2000. Experiments in iterated instance-based learning. Technical report, Language Evolution and Computation Research Unit.

—— 2002. Compositional syntax from cultural transmission. *Artificial Life* 8.25–54.

BRISCOE, E. (ed.) 2002. *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge: Cambridge University Press.

HAIMAN, J. 1980. The iconicity of grammar: Isomorphism and motivation. *Language* 56.515–540.

HARE, M., & J. L. ELMAN. 1995. Learning and morphological change. *Cognition* 56.61–98.

HURFORD, J. R. 2000. Social transmission favours linguistic generalization. In *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, ed. by C. Knight, M. Studdert-Kennedy, & J.R. Hurford, 324–352. Cambridge: Cambridge University Press.

KIRBY, S. 2002. Learning, bottlenecks and the evolution of recursive syntax. In Briscoe (2002), 173–203.

——, & J. R. HURFORD. 2002. The emergence of linguistic structure: An overview of the iterated learning model. In *Simulating the Evolution of Language*, ed. by A. Cangelosi & D. Parisi, 121–147. Springer Verlag.

LANGACKER, R. W. 1977. Syntactic reanalysis. In *Mechanisms of Syntactic Change*, ed. by C. N. Li, 57–139. Austin, TX: University of Texas Press.

MACNAMARA, J. 1982. *Names for things: a study of human learning*. Cambridge, MA: MIT Press.

MAŃCZAK, W. 1980. Laws of analogy. In *Historical Morphology*, ed. by J. Fisiak, 283–288. The Hague: Mouton.

MARKMAN, E. M., & G. F. WACHTEL. 1988. Children's use of mutual exclusivity to constrain the meaning of words. *Cognitive Psychology* 20.121–157.

PINKER, S. 1984. *Language Learnability and Language Development*. Cambridge, MA: Harvard University Press.

SLOBIN, D. I. 1977. Language change in childhood and history. In *Language Learning and Thought*, ed. by J. Macnamara, 185–221. London: Academic Press.

SMITH, K. 2002a. The cultural evolution of communication in a population of neural networks. *Connection Science* 14.65–84.

—— 2002b. Natural selection and cultural selection in the evolution of communication. *Adaptive Behavior* 10.25–44.

——, 2003. The Transmission of Language: models of biological and cultural evolution. PhD Thesis, The University of Edinburgh.

——, S. KIRBY, & H. BRIGHTON. forthcoming. Iterated learning: a framework for the emergence of language. In *Self-organization and Evolution of Social Behaviour*, ed. by C. Hemelrijk. Cambridge: Cambridge University Press.

Modeling language acquisition, change and variation

Willem Zuidema
Language Evolution and Computation Research Unit
School of Philosophy, Psychology and Language Sciences
and Institute of Animal, Cell and Population Biology
University of Edinburgh
40, George Square
Edinburgh EH8 9LL, United Kingdom
jelle@ling.ed.ac.uk

http://www.ling.ac.uk/ ∼ jelle

Abstract

The relation between Language Acquisition, Language Change and Language Typology is a fascinating topic, but also one that is difficult to model. I focus in this paper on the question how theories of language acquisition constrain theories of language change and typology. In the generative tradition and "Learnability Theory" this problem is approached by assuming that all linguistic variation can be described in terms of a relatively small number of parameters of a universal, innate core, the Universal Grammar. In this view, language acquisition is parameter setting, and language change is parameter change. I review some simple acquisition models and their consequences for language change, and discuss some problems with this approach. I will then discuss an alternative approach that is based on "Explicit Induction" algorithms for grammatical formalisms. I discuss which approach is most useful for which problems.

1 Language acquisition, change and typology

Every healthy human infant is capable of acquiring any one of a dazzling variety of human languages. This simple fact poses two fundamental challenges for linguistics: (1) understanding how children are so extremely successful at this apparently complex task, and (2) understanding how, although all humans have such similar linguistic abilities, such a wide variety of languages has emerged. These challenges are intricately linked: the languages that we observe today, are the result of thousands of years of cultural transmission, where every generation has acquired its language from the observed use by previous generations. That makes the acquisition of language a rather unique learning problem for learning theory, because what is being learned is itself the result of a learning process. Conversely, the structure of a language (say modern English) at any one time (say, 2003) is the result of perhaps millions of individuals learning from examples from a language with a very similar structure (say, the English of the 1960s).

This so-called *circular causality* (Steels, 1999) makes the relation between Language Acquisition, Language Change and Language Typology a fascinating topic, but also one that is difficult to model. I will focus here on the question how theories of language acquisition constrain theories of language change and typology. In the generative tradition and "Learnability Theory" this problem is approached by assuming that all linguistic variation can be described in terms of a relatively small number of parameters of a universal, innate core, the Universal Grammar. In this view, language acquisition is parameter setting, and

language change is parameter change. In the following I will review some simple acquisition models and their consequences for language change, and discuss some problems with this approach. I will then discuss an alternative approach from the emerging field of computational modeling of the evolution of language (Kirby, 2002b), that is based on "Explicit Induction" algorithms for grammatical formalisms. I will argue that the differences between the two approaches have been exaggerated, and will discuss for which sort of problems which sort of approach is most useful.

2 Parameter models

The "Parameter change" approach to this problem is based on *parameterizing* linguistic structure, such that we can characterize all differences between possible human languages by a vector of a small number of parameters. E.g., in the Principles and Parameters approach (Chomsky, 1981; Bertolo, 2001), language acquisition is described in terms of parameter settings for a universal core, the Universal Grammar. With such a description of language in hand, we can reformulate the challenges as follows: (1) how can learning, given primary linguistic data that conforms to any particular set of parameters, find that set of parameters? (2) given a set of learning procedures that are capable of finding the correct parameters, which ones predict the type of language change and statistical distributions (universals tendencies, Kirby 1999) that we can actually observe?

2.1 Parameter setting

In the "parameter setting" models of language acquisition, one assumes a finite number N of possible grammars. If all variation can be described by n different, Boolean and independent parameters, such that the total number of possible grammars is $N = 2^n$. Such parameters determine, for instance, whether or not an object precedes the main verb in a sentence, or whether or not the subject can be left out. Typically, although the number of parameters is estimated at around 30, concrete examples are only worked out for the 2 or 3 least controversial proposed parameters. A lot of work in parameter setting works with rather simplified models that can be studied analytically, and that depend only on the finiteness of N. Examples of such models are "memory-less learning", "batch learning" (e.g. Nowak *et al.*, 2001) and "learning by enumeration" (Gold, 1967). It is useful to look in a bit more detail at these models.

Memory-less learning (Niyogi, 1998) is arguably the simplest language acquisition model. The algorithm works by choosing a random grammar from the set of possible grammars each time the input data shows that the present hypothesis is wrong. The algorithm obviously is not very efficient, because it can arrive at hypotheses it has already rejected before; i.e. each time it randomly chooses a new grammar, it forgets what it has learned from all data it has received before. This algorithm is only of interest because it is simple and provides a lower bound on the performance of any reasonable algorithm (Nowak *et al.*, 2001).

The *batch learner*, in contrast, memorizes all received sentences and finds all grammars from the set of possible ones that are consistent with these sentences. Equivalently, it keeps track of all possible grammars that are still consistent with the received data. In any case, for any reasonably large set of possible grammars, the batch learner has monstrous memory and processing requirements. Its value lies in the fact that it is simple, and provides an upper bound on the performance of any reasonable learning algorithm, as long as there is no a-priori reason to prefer one grammar that is consistent with the data over another.

As exemplified by appendix A, we can, with a bit of effort, derive explicit formulas that describe the probability of success q as a function of the number of input sentences for both the memory-less and the batch learner. Under the assumption that every wrong grammar is equally similar to the right grammar (described with a similarity parameter a), we can in fact give a complete transition matrix T, where all

diagonal values are $q_{memoryless}$ and all off-diagonal values are $(1 - q_{memoryless})/(N - 1)$. This transition matrix plays an important role in models of language change described in the next section.

It is important to realize that these algorithms only work because a finite (and in fact, relatively small) number of possible grammars is assumed. Moreover, calculations such as in appendix A are relatively easy due to some important assumptions: (1) that the algorithms are not biased at all to favor certain possible grammars over others; (2) that (in the case of the memory-less learner) the probability of jumping to a wrong or right grammar remains constant throughout the learning process; and (3) that all grammars are equally similar to each other. Without these assumptions, similar calculations quickly get rather complex.

For instance, *learning by enumeration* (Gold, 1967), as the name suggests, proceeds by enumerating one at a time, and in prespecified order all possible grammars. Only if a grammar is inconsistent with incoming data ("text"), does the algorithm move on to the next grammar. The procedure is of interest, because it can be used as a criterion for learnability (Gold, 1967)[1]. Calculating q is more difficult than before, because the probably of changing to a wrong grammar *decreases* over time.

The *trigger learning algorithm* (Wexler & Culicover, 1980) is a popular model that is of (slightly) more practical interest. Rather than picking a random new grammar, as the memory-less learner does, or enumerating grammars in a random order, as in learning by enumeration, it changes a random parameter when it finds an input sentence that is inconsistent with the present hypothesis. If with the new parameter setting the sentence can be parsed, the change is kept, otherwise it is reverted. The trigger learning algorithm thus implements a kind of hill-climbing (gradient ascent), by keeping parameters that do well and only making a small change when it improves performance. The probability of the trigger algorithm to give the right grammar after b sentences is even more tricky to calculate, because the probability to reject a wrong hypotheses *decreases* as more and more parameters get correctly set.

Many other parameter setting models exists. E.g. Briscoe (2002a) develops a variant of the trigger learning algorithm, where parameters are no longer independent, but fall into linguistically motivated inheritance hierarchies. Further, rather than choosing a single parameter at random and changing it, as in the TLA, Briscoe's algorithm selects several random parameters and keeps track of their most likely setting in a Bayesian, statistical fashion. Yang (2000) argues that language acquisition is best viewed as a selectionist process, where many different parameter sets are considered in parallel. Niyogi & Berwick (1995) and Yang (2000) consider the further complication that children learn from input sentences that are drawn from different languages, and explore the expectations on what grammar settings they will end up with. In all these models, calculating the probabilities of the outcome of learning gets very complex and results are typically obtained by using computer simulations.

2.2 Parameter change

Niyogi & Berwick (1995), as well as neural network modelers Hare & Elman (1995), argue that a theory of language acquisition – and the mistakes children make when confronted with insufficient or ambiguous input – implies a theory of language change. Similarly, Kirby (1999) explores the idea that a theory on language use and processing – which alter the primary linguistic data – leads to specific expectations on language change and the resulting linguistic variation. Hence, by working out the consequences for language change and comparing them to empirical data, theories on language use, processing and acquisition can be

[1] Learning by enumeration can, within finite time, find the target grammar from a class of grammars if the following conditions hold: (1) the class of grammars is finite (enumerable), (2) for every two grammars in the class, there exists a sentence that distinguishes between two grammars (i.e. that is grammatical according to one, and ungrammatical according to the other), and (3) the distinguishing sentence will occur within a finite amount of time in the the text, generated by the target grammar. It follows that the class of grammars is then learnable from text. It can be shown that superfinite classes of grammars, such as the context-free or context-sensitive grammars, are not learnable in this sense (Gold, 1967). Principles & Parameters-models, in contrast, are learnable (Wexler & Culicover, 1980) and so are many other classes (Angluin, 1980).

tested. Formally, a class of grammars $\mathcal{G}$, a learning algorithm $\mathcal{A}$ and a model of the primary linguistic data (a probability distribution $\mathcal{P}_i$ over the possible sentences of language i) together constitute the main ingredients of a dynamical system that describes the change in numbers of speakers of each language[2].

Several general results have been obtained. For instance, Niyogi & Berwick (1995) and Yang (2000) find that with different choices for $\{\mathcal{G}, \mathcal{A}, \mathcal{P}\}$, the change in the number of speakers of a particular language tends to follow an S-shaped curve, consistent with observed patterns in historical data. More interestingly, Nowak *et al*. (2001) derive a *coherence threshold*. In their model, natural selection selecting for more frequent grammars, helps a population to converge on a specific grammar. Mistakes in learning, on the other hand, lead to divergence, because it essentially randomizes the choice of grammars. Nowak et al. find that if the accuracy in learning is below a precise threshold, all coherence in the population is lost and all languages are spoken with equal probability[3].

Niyogi and Berwick apply their methodology to a number of case studies. For instance, they look at a simple 3-parameter system where the parameters determine whether or not specifiers (1) and complements (2) come before the head of a phrase, and whether or not the verb is obligatorily in second position (3). In this system, there are 8 different possible grammars (languages). By making assumptions on the frequency with which triggers for each of the parameters are available to the child, they can estimate the probability a specific learning algorithm can learn each language. They numerically determine the probabilities of transitions between each of the 8 language over 30 generations with 128 triggers per generation. They find that languages with the third parameter set to "0" ($V2-$) are extremely unstable and that the $V2+$ parameter therefore quickly gets fixed in all simulations. This observation is contrary to observed trends in historical data, where $V2+$ is typically lost. Niyogi and Berwick argue that this falsifies their preliminary model, and thus illustrates the feasibility of testing the diachronic accuracy of the assumptions on $\{\mathcal{G}, \mathcal{A}, \mathcal{P}\}$.

2.3 Some features of parameter change models

Several other parameter change models have been studied. They have in common the emphasis on the uniformity of languages, i.e. all possible languages (grammars) are of equal quality. Hence, children acquiring a language do not go from a simple grammar to a more complex one, but rather jump from one grammar to an equally complex alternative. Not the quality of the language, but the uncertainty about which is the correct one changes over time.

Moreover, in all these models the acquisition of syntax is studied independently from the acquisition of phonology, semantics, pragmatics and the lexicon, and, usually, independent from the particularities of the child's parsing algorithm. The training data are "triggers", i.e. strings of grammatical categories. The problems of learning the syntactic categories of words and their meaning, and learning to recognize the phonological form and the boundaries between words are all ignored.

Further, the models fit into a tradition that is much mathematically oriented. Although many results are obtained through numerical simulations, the models are formulated at a rather abstract level. Generations are typically discrete, the number of parameters small (2, 3, 5), number of training samples and the number of individuals in a population very small or, alternatively, infinite.

The models are valuable, because they give a *general* insight in how linguistic conventions can change and spread in a population. However, the problem with this approach is that its potential for explaining *specific* aspects of language acquisition and language typology depends completely on the successful parametrization of linguistic descriptions. That dependence has advantages, because it makes the relation with other linguistic theories very clear, but it has some major disadvantages as well.

[2] In addition to the triple $\{\mathcal{G}, \mathcal{A}, \mathcal{P}\}$ (Niyogi & Berwick, 1995), one needs assumptions on population and generation structure and the number of training sentences the algorithm receives.

[3] Presumably, a similar mechanism explains the lack of coherence in the simulations of Niyogi & Berwick (1995).

First, there is, as for now, no such parametrization available. If efficient parametrization (i.e. with 20 or 30 parameters) turns out to be impossible, models that depend on them will be inadequate. Second, even if it is possible in principle, without a complete theory available on what each parameter means, solutions in terms of these parameters give little insight on why children learn certain things with more ease than others, or why languages tend to show certain patterns more often than others. Finally, parameter-models might give an adequate description of the variation in languages in a quasi-stable state, but that does not necessarily mean that they also give an adequate description of language variety when languages are changing. In particular, observed trends in language change regarding the interaction between phonology, syntax, semantics and pragmatics seem hard to capture in available parameter models.

3 Explicit Induction

3.1 Grammar Induction: impossible and irrelevant?

Grammar Induction algorithms are usually based on the intuition that the frequency of occurrence of substring in the training sentences, and the contexts in which they appear, contain information on what the underlying constituents and the rules of combination of the target grammar are. E.g. Zellig Harris, in describing the methods linguists use to infer the grammar of an unknown language, defines the crucial concept of "substitutability" as follows: "If our informant accepts DA'F as a repetition of DEF, and if we are similarly able to obtain E'BC as equivalent to ABC, then we say that A and E are mutually substitutable" (Zellig Harris, 1951, quoted in van Zaanen 2001).

It is possible to design induction algorithms that, just like Harris's linguist, use observed patterns in training sentences to induce the underlying grammar. However, due to initial negative results on the theoretical possibility of learning a grammar from positive data (Gold, 1967) and developments in linguistic theory (e.g. Chomsky, 1965), the *induction* of grammar has been widely viewed as both impossible and irrelevant.

The supposed impossibility of grammar induction is based on a widespread misinterpretation of negative learnability results. Gold (1967) showed that e.g. the class of context-sensitive languages is not *identifiable in the limit*. Even we if accept identification in the limit as the appropriate criterion for learnability, Gold's results mean nothing more than, in his own words:

> *"The class of possible natural languages is much smaller than one would expect from our present models of syntax. That is, even if English is context-sensitive, it is not true that any context-sensitive language can occur naturally. Equivalently, we may say that the child starts out with more information than that the language it will be presented is context-sensitive. In particular, the results on learnability from text imply the following: The class of possible natural languages if it contains languages of infinite cardinality, cannot contain all languages of finite cardinality." (Gold, 1967)*

In other words, a class of context-sensitive grammars needs to be constrained to make it learnable. Angluin (1980) has shown that very non-trivial classes of formal languages are learnable. Nothing in the formal results, however, proves that the necessary restrictions are due to an extensive, innate, language-specific Universal Grammar; they could be simply generic properties of the human brain[4].

The supposed irrelevance of grammar induction algorithms is based on the fact that the dominant linguistic theories of the last decades assume extensive innate knowledge. If children don't do grammar induction, why design computer programs that do? Evidence for this view comes – in addition to the learnability

[4] Although it is of course true that learnability is a valid test for judging the validity of a (grammatical) theory, and that few proposed theories other than those from the nativist tradition pass it. However, one can argue that nativist theories, rather than solving the learnability problem, simply shift it to the domains of evolutionary theory and cognitive neuroscience.

results – from empirical observations in child language acquisition. Typically, such arguments have the form: the child correctly uses construction X very early in life, even though the primary linguistic data it has received up to that point does not provide enough evidence to choose between X and several alternative logical possibilities. Thus, it is concluded, the child must have prior (innate) knowledge of X.

More and more it is now recognized that this "knowledge of X" might be an emergent result of the interaction between not necessarily language-specific cognitive and learning abilities, and the structure, meaning and pragmatics of the linguistic data the child received (MacWhinney, 1999). Consequently, the need to postulate language-specific adaptations might be limited (Jackendoff, 2002; Hauser *et al.*, 2002).

3.2 Induction Algorithms

Wolff (1982), and similarly Stolcke (1994), Langley & Stromsten (2000) and Zuidema (2003), presents a model based on the idea that a grammar is a compressed representation of a possibly infinite language (string set). These algorithms all use context-free grammars as the grammar formalism, learn from text and run through three phases that can be termed "incorporation", "compression" and "generalization". I will refer to these algorithms as "compression-based induction".

In the incorporation phase, input sentences s are stored as idiosyncratic rewrite rules $S \mapsto s$. In the compression phase (or "syntagmatic merging"), the most frequent substrings z in the right-hand sides of the stored rules are replaced by a unique non-terminal symbol N. Rules of the form $N \mapsto z$ are added to the grammar. In the generalization phase (or "paradigmatic merging"), two nonterminals N and N' are considered *substitutable* if they occur in the same context; all occurrences of N' are then replaced by N. Different variants of the basic algorithm differ in how *greedy* they are, and in whether or not they are *incremental*. Kirby (2000), and later papers, uses a algorithm were the context-free grammars are enriched with a predicate-logic based semantics.

A related framework based on substitutability is developed by van Zaanen (2001) and termed "Alignment Based Learning" (ABL). Van Zaanen develops a number of algorithms for the two phases of the ABL framework: Alignment learning and selection learning. In the alignment learning phase input sentences are compared, aligned and common substrings are identified. The *unequal* parts z and z' of the two sentences are labeled with a non-terminal. The non-terminal is unique if neither z nor z' was labeled already, but the algorithm reuses the existing label if available, and equates the two non-terminals if both z and z' were labeled already. In the latter two conditions a form of generalization occurs. Each labeling is a hypothesis on a possible constituent of the target language, and very many such hypotheses are generated.

In the selection learning phase, a subset of the generated hypotheses is selected. That subset is chosen such that it is concise (each hypothesis can be used to analyze many sentences), and that it is internally consistent (hypotheses do not overlap). The ABL algorithm yields a tree-bank: an annotated version of the input corpus (it thus implements automated tagging). From the tree-bank, context-free grammars can be trivially induced.

3.3 Language Evolution

In the "Explicit Induction" approach to modeling language change and evolution, language change is studied based on similar induction algorithms, i.e. learning algorithms that produce an explicit grammar based on training sentences (see Hurford, 2002, for a review). Such an approach avoids the problems of parameter models, because they can incorporate any available linguistic formalism. However, they have two major disadvantages as well: (1) language induction is very challenging problem that is far from solved, even for simplified and well understood grammar formalisms; (2) models that incorporate a full-blown linguistic formalism, including procedures for language production and interpretation, quickly get very complex.

Two recent models by Kirby (2002a) and Batali (2002) show that there is reason for optimism for progress on bl problems. Kirby presents a model that is very clear in its set-up. It uses first-order predicate logic with a small set of entities and predicates to represent semantics, and a extension of context-free grammars to represent syntax and the syntax-semantics mapping. The model thus uses well-understood and conventional linguistic formalisms and a simple learning procedure. However, by using the output of one learning cycle as input for the next Kirby was able to get some unconventional results: the spontaneous emergence of a recursive, infinite but learnable language. However, the learning algorithm used is very brittle, and it's difficult to extend the model to domains with more diverse semantics and a more heterogeneous syntax.

In contrast, Batali's model is very difficult to understand. It also uses a form of predicate logic to represent semantics, but it uses "exemplars" as the basic representation of the grammar, and "argument maps" to guide the combination of exemplars into meaningful sentences. The results show the emergence of a complex language, with properties similar to case marking and subordinate clause marking in natural languages. The emergent languages are essentially infinite but nevertheless learnable (from meaning–form pairs). The learning algorithm is successful and robust in this complex domain presumably because of the redundancy it allows.

3.4 Some features of explicit induction models

Several other explicit induction models have been studied. They have in common that no uniformity of languages are assumed. Typically, individuals in these models start with an empty grammar and empty lexicon, and gradually add new rules and lexical items based on the received sentences and observed patterns. Individuals are, however, equipped with an invention procedure, such that they can generate new sentences when required.

Further, in these models learning is typically from form–meaning pairs and a lexicon is built-up in parallel with the grammar. The recognition of phonemes and the pragmatics of dialogs are built-in as assumptions of the models.

The models are all implemented as computer programs. Typically, the models are rather concrete: they consist of a population of individuals, with procedures for production, invention, interpretation and induction, and a set of possible message to communicate. The languages studied in these models are still relatively simple, and exhibit just some basic word orders or morphological markers for the semantic roles of agents, patients and action. Empirical data from historical linguistics has so far played no role in these studies.

4 Discussion

I have reviewed some models of language acquisition and language change from two different traditions. The crucial question – which approach is best? – is still largely open to discussion. The following issues are important in comparing both approaches:

Learnability - Theoretical arguments. From the field of learnability theory it has sometimes been argued that grammar induction is impossible. In section 3.1 I have argued that this position is based on a misunderstanding of the negative learnability results. Learnability, however, is an important test for the validity of a grammar formalisms and induction algorithms. The challenge is to find a combination of a formalism that is as expressive as human languages are (i.e. mildly context-sensitive), and a learning algorithm that can induce it from the available primary linguistic data. In my view, parameter setting

models meet this challenge, but only by making unsatisfactory assumptions on the prior knowledge the algorithms start with. Explicit induction models, on the other hand, present considerable progress (i.e. most work with context-free grammars), but more work still needs to be done.

Learnability - Empirical arguments. From the field of psycholinguistics it has been argued that children have prior knowledge of syntactic constructions, because they choose, from apparently many logical possibilities that are consistent with the received evidence, the correct, seemingly arbitrary option. Grammar induction models, in this view, are – if not impossible – irrelevant, because children do not do induction. I believe that explicit induction algorithms have already shown that the logic of this argument is false. There is no need for assuming explicit prior knowledge, because the outcome of the interaction between learning biases and training data is subtle and often unexpected. Moreover, because languages are transmitted culturally from generation to generation, seeming arbitrary choices are likely to be the correct ones, because previous generations have used the same arbitrary learning algorithm to learn their language (Deacon, 1997; Kirby, 2000; Briscoe, 2002a; Zuidema, 2003).

Equivalence More subtly, it has been suggested that explicit induction models might in some sense be equivalent to parameter setting models. If the space of grammars that induction algorithms explore is finite, then that space could in principle be parametrized and hence described by a finite number of parameters. The induction algorithm can then be described, albeit possibly in a clumsy and complicated way, as a parameter setting procedure. If this is true – and it presumably is for the context-free grammar and finite-state machine inducers – the crucial issue is parsimony and clarity. Presumably, for some purposes the representation in terms of parameters is more useful, but for comparison with psycholinguistic, neurological and historical data the explicit grammar representation seems more appropriate. Further, the parameterized representation leads naturally to the uniformity assumptions, whereas the explicit grammar representation leads naturally to the view that grammars grows over time. Finally, stochastic grammar formalisms can not be parametrized in the concise way that parameter setting models usually assume. Worse, lexicalized, exemplar-based models can not be parametrized because there are infinitely many probability distributions that can be assigned to the string set (Bod, 1998).

In conclusion, the two approaches to modeling of language change are rooted in different theoretical positions on the nature of language and language acquisition. If one adopts the Principles and Parameters framework, the parameter change approach is the appropriate way to conceptualize language change. However, this approach requires more work to make explicit how each parameter is to be interpreted, which triggers for each parameter are available, how the child learns her lexicon and recognizes syntactic categories in the sentences it receives, how parameters depend on each other, etc. Moreover, it requires a satisfactory explanation for the evolution and development of the Universal Grammar in the child's brain. However, some Explicit Induction models might, even if one adopts this approach, still be useful as an equivalent representations that can be more easily compared to empirical data.

If one rejects the Uniformity Hypothesis and conceptualizes grammar acquisition as the gradual built-up of a grammar in the mind of the child, explicit induction models are the appropriate approach. Parameter change models are still useful as simple, but mathematically sophisticated models of how conventions spread in a population.

References

ANGLUIN, D. (1980). Inductive inference of formal languages from positive data. *Information and Control* **21**, 46–62.

BATALI, J. (2002). The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In: Briscoe (2002b).

BERTOLO, S., ed. (2001). *Language Acquisition and Learnability*. Cambridge University Press.

BOD, R. (1998). *Beyond Grammar: An experience-based theory of language*. Stanford, CA: CSLI.

BRISCOE, T. (2002a). Grammatical acquisition and linguistic selection. In: Briscoe (2002b).

BRISCOE, T., ed. (2002b). *Linguistic evolution through language acquisition: formal and computational models*. Cambridge University Press.

CHOMSKY, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

CHOMSKY, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

DEACON, T. (1997). *Symbolic species, the co-evolution of language and the human brain*. The Penguin Press.

GOLD, E. M. (1967). Language identification in the limit. *Information and Control (now Information and Computation)* **10**, 447–474.

HARE, M. & ELMAN, J. (1995). Learning and morphological change. *Cognition* **56**, 61–98.

HAUSER, M., CHOMSKY, N. & FITCH, W. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science* **298**, 1569–1579.

HURFORD, J. R. (2002). Expression / induction models of language. In: Briscoe (2002b).

JACKENDOFF, R. (2002). *Foundations of Language*. Oxford, UK: Oxford University Press.

KIRBY, S. (1999). *Function, selection and innateness: The emergence of language universals*. Oxford University Press.

KIRBY, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In: *The Evolutionary Emergence of Language: Social function and the origins of linguistic form* (Knight, C., Hurford, J. & Studdert-Kennedy, M., eds.). Cambridge, UK: Cambridge University Press.

KIRBY, S. (2002a). Learning, bottlenecks and the evolution of recursive syntax. In: Briscoe (2002b).

KIRBY, S. (2002b). Natural language from artificial life. *Artificial Life* **8**, 185–215.

KOMAROVA, N., NIYOGI, P. & NOWAK, M. (2001). The evolutionary dynamics of grammar acquisition. *J. Theor. Biology* **209**, 43–59.

LANGLEY, P. & STROMSTEN, S. (2000). Learning context-free grammars with a simplicity bias. In: *Proceedings of the Eleventh European Conference on Machine Learning*, pp. 220–228. Barcelona: Springer-Verlag.

MACWHINNEY, B., ed. (1999). *The emergence of language*. Mahwah, NJ: Lawrence Erlbaum Associates.

NIYOGI, P. (1998). *The informational complexity of learning*. Boston, MA: Kluwer.

NIYOGI, P. & BERWICK, R. C. (1995). The logical problem of language change. Tech. rep., M.I.T.

NOWAK, M. A., KOMAROVA, N. & NIYOGI, P. (2001). Evolution of universal grammar. *Science* **291**, 114–118.

STEELS, L. (1999). The puzzle of language evolution. *Kognitionswissenschaft* **8**.

STOLCKE, A. (1994). *Bayesian Learning of Probabilistic Language Models*. Ph.D. thesis, Dept. of Electrical Engineering and Computer Science, University of California at Berkeley.

WEXLER, K. & CULICOVER, P. (1980). *Formal principles of language acquisition*. Cambridge MA: MIT Press.

WOLFF, J. G. (1982). Language acquisition, data compression and generalization. *Language & Communication* **2**, 57–89.

YANG, C. D. (2000). Internal and external forces in language change. *Language Variation and Change* **12**, 231–250.

VAN ZAANEN, M. (2001). *Bootstrapping Structure into Language: Alignment-Based Learning*. Ph.D. thesis, School of Computing, University of Leeds.

ZUIDEMA, W. (2003). How the poverty of the stimulus solves the poverty of the stimulus. In: *Advances in Neural Information Processing Systems 15 (Proceedings of NIPS'02)* (Becker, S., Thrun, S. & Obermayer, K., eds.). Cambridge, MA: MIT Press. (forthcoming).

A Memory-less learner and batch learner

To estimate the probability that memory-less learning finds the correct grammar after a certain number (b) of sample sentences, we need to consider the inverse: the probability that the algorithm still has a wrong hypothesis after b sample sentence.

$$P(right\ grammar\ after\ b\ samples) = 1 - P(wrong\ grammar\ after\ b\ samples) \tag{1}$$

The probability that the learner still has the wrong hypothesis, depends on the probability that it initially chose the wrong hypothesis (simply $(N-1)/N$) times the probability that it remained for all b sentences at a wrong hypothesis. If it makes no essential difference which wrong grammar is the present hypothesis and how long it has held it as the hypothesis[5], the probability that the algorithm remain for b sentences at a wrong hypothesis is simply $P(remain)^b$. Hence,

$$P(wrong\ grammar\ after\ b\ samples) = \frac{(N-1)}{N}\left(P(remain)\right)^b \tag{2}$$

The probability to remain at a wrong grammar for each random input sentence is given by the probability that that input sentence happens to be consistent with the present (wrong) grammar, plus the probability that the algorithm jumps to another wrong grammar:

$$P(remain) = P(consistent) + P(another\ wrong\ grammar) \tag{3}$$

The probability that a sentence is consistent with a wrong grammar is simply the similarity parameter a in Nowak *et al.* (2001). The probability that the algorithm jumps to another wrong grammar is given by the probability that the input sentence is inconsistent $(1-a)$ times the fraction of other wrong grammars $((N-2)/N)$.

Putting all this together, the probability (q) that the memory-less learner has found the correct grammar after b input sentences is given by (Komarova *et al.*, 2001)[6]:

$$\begin{aligned} q_{memoryless} &= 1 - \frac{(N-1)}{N}\left(a + \frac{(N-2)(1-a)}{N-1}\right)^b \\ &= 1 - \frac{(N-1)}{N}\left(1 - \frac{(1-a)}{N-1}\right)^b \end{aligned} \tag{4}$$

The probability that the batch learner has found the correct grammar after b input sentences is found by Nowak *et al.* (2001) to be

$$q_{batch} = \frac{\left(1 - \left(1 - a^b\right)^N\right)}{(Na^b)} \tag{5}$$

[5]That is the case, for the memory-less learner, under the assumption of Nowak *et al.* (2001) that all grammars are equally similar to each other. In contrast, in a Principles & Parameters model, we can calculate the expected similarity based on estimates of how many parameters are revealed in a single sentence. Under the assumption that every sentence reveals m parameters, that all parameters are Boolean and that all parameters are revealed with equal probability: $a \approx \left(\frac{1}{2}\right)^m$. $a \approx \left(\frac{1}{2}\right)^m$. a is then an expected value rather than a constant, and equation (2) needs to be adapted. For simplicity, we will here follow the assumption of Nowak et al.

[6]Note that there is an error in this equation in Nowak *et al.* (2001) that is corrected in Komarova *et al.* (2001)

Modelling the Emergence of Case

Joanna Moy and Suresh Manandhar
Department of Computer Science
University of York

http://www.cs.york.ac.uk/∼joanna

1 Introduction

A further investigation into the role of linguistic evolution as an alternative to biological evolution in the emergence of syntax is presented. This follows on from the idea that languages themselves are evolving entities, which adapt to be easily acquired by the human learner. It has already been shown that it is possible for the rudimentary elements of syntax, i.e. compositionality and recursiveness, to emerge in a population of language learning agents without any specific linguistic knowledge and in the absence of selection for communicative ability[8],[7]. Attempts are made here to extend this framework to cover another important aspect of natural language: the use of a case system to make distinctions between thematic roles.

1.1 The Origins of Human Language

Human language seems to be both qualitatively and quantitatively unlike any other form of animal communication. Unsurprisingly, then, the origins of this uniquely human ability are the subject of much debate, particularly the question of nature versus nurture. Is language innate? Or must it be "discovered" anew by each new human learner? Or is the real answer somewhere between these two extremes?

That we are in some respects predisposed to language is beyond doubt. Terrence Deacon suggests that human beings manifest an extensive array of perceptual, motor, learning, and even emotional predispositions towards the learning of language[5]. For example, the positioning of the human larynx is significantly lower in the throat than that of other primates, and we have a greater degree voluntary control over our respiratory function, enabling the long slow exhalations necessary for the production of speech. Although cognitive adaptations to language are less easy to identify than physiological ones, it is worth noting that "all normal children, raised in normal social environments, inevitably learn their local language, whereas other species, even when raised and taught in this same environment, do not". Attempts to teach symbolic communication to non-human primates have produced some interesting results, but it is clear that they do not have the same facility and natural aptitude for language that we do. Language emerges in all normal children all over the world at approximately the same age, no matter what culture they are growing up in and what language they are acquiring[1]. Furthermore, Briscoe notes that "failure [to acquire language] appears to correlate more with genetic defects or with an almost complete lack of linguistic input during the critical period, than with measures of general intelligence or the quality or informativeness of the learning environment"[3].

It is clear that there is "something special about human brains that enables us to do with ease what no other species can do even minimally without intense effort and remarkably insightful training"[5].

So perhaps the real question is this: what aspects of the ability to learn language are innate, and what form do they take?

1.2 Principles and Parameters

The Chomskyan position is generally stated to be that children are born with a large amount of innate knowledge regarding the structure of language (see Culicover[4] for review). This premise is based on the observation that we do not learn to speak by merely repeating utterances that we have heard, but instead internalize a set of rules for the construction of novel utterances, and yet the language to which children are exposed is not sufficient to make these rules explicit. If the structure of language is underspecified, then children must need additional guidance to successfully acquire the rules. In his earlier work, Chomsky suggests that this might take the form of information about linguistic universals, i.e. knowledge of the general outline of language, a kind of "blueprint". Acquisition simply becomes a problem of choosing the appropriate language from within the space of possible languages, for which the existence of a specific cognitive module or Language Acquisition Device is postulated. Chomsky's more recent work on the Principles and Parameters framework takes this idea further still, in proposing a Universal Grammar which constrains the space of possible languages even more tightly. Language acquisition is reduced to the setting of a "finite number of finite-valued parameters".

There are many objections to the Chomskyan theory. Among them is the idea that a Language Acquisition Device is impossible because it could never have arisen by Darwinian natural selection. Instead it must have required some kind of macro-mutation in order to come into existence. The debate rages on this issue, as Chomsky himself believes that natural selection cannot account for his theory, and wishes instead to invoke "alternative physical principles". However, Stephen Pinker[11], also strongly in favour of the nativist view of language acquisition, (although of a slightly different flavour), argues that natural selection, by virtue of its directedness, is the only process that *can* account for such a highly specialized organ. Another area of controversy is whether or not child-language learning data actually fits the patterns predicted by the theory. Aitchison[1] suggests that if the Principles and Parameters model is correct, then children should be aware of the innate constraints on the form of language and should never make an utterance that is outside the scope of all natural languages. Also, their acquisition should proceed in dramatic steps as the appropriate parameter settings are acquired. Yet there seems to be plentiful child-language data which suggests that this is not the case[13].

1.3 Language as an Adaptive System

An alternative to the Universal Grammar is the idea that, rather than having brains that are pre-wired with quite specific information about language, we are instead equipped with puzzle-solving apparatus that enables us to successfully process linguistic data as we come across it[1]. The stronger form of this hypothesis is that children possess language-specific puzzle-solvers, whereas in the weaker form the mechanisms used to acquire language are simply a facet of the general intelligence of human beings.

A yet more radical stance suggests that Chomsky has inverted cause and effect. Rather than suggesting that human brains have evolved to be better able to acquire language, could it be that languages themselves evolve to be more easily acquired by human brains? Can languages themselves be viewed as evolving entities, which have adapted to suit the cognitive processes present in the human mind? When, as Chomsky observed, children quickly master the basics of grammar despite the paucity of the input stimuli from which they learn, one could perceive them to be making "lucky guesses" about grammar and syntax and the way words work together. These might be all too easy

to attribute to some innate language-specific knowledge. However, Terrence Deacon[5] believes that learning is occurring by trial and error, but that a very large proportion of the guesses children make are correct – not because they possess innate knowledge about language, but because the languages they are learning have evolved in such a way as to fit in with the guesses that are likely to be made.

It is possible to view the relationship between a language and its speaker as a similar to that between a virus and its host. Just as the viral DNA uses its host cell in order to reproduce, the information encapsulated in a grammar of a language becomes integrated into the machinery of the human brain, and uses this to reproduce.[1] The metaphor can be taken further still by the observation that although a common language may link a social group, the *internal* grammar of each of those speakers is unlikely to be identical, but subject to variation – thus the language of their community is more like a collection of similar but not identical languages. And as a result of this variation, languages are able to evolve with respect to the selection pressures around them, just as variation within a viral gene-pool drives evolution.

The selection pressures in question are the biases of human learners. In making their "lucky guesses" about the way in which words work together, children actually neglect a large proportion of the hypothesis space – they fail to explore the full range of possible ways of organising words. Thus a language that organises its words in a way that falls outside the "lucky guesses" of children will not be easily learnt (and will be less likely to pass on to the next generation), whereas a language for the whom the "lucky guesses" are correct guesses will be much more easily acquired.

The emergence of simple languages and compositional structures as a result of this process has been demonstrated using computer simulations[2],[8]. In particular, Kirby has done a series of experiments in which he has demonstrated the evolution of syntax in a population of agents equipped with a set of simple grammar learning heuristics, but which are not themselves subjected to any selective pressures, [6],[7],[8],[9], showing that it is possible for a language to evolve independently of its "host".

Kirby's work incorporates what he refers to as the "iterated learning model" where agents learn observationally from the behaviour of others. Agents in the simulation are all identical, and there is no selection for communicative ability. Learners attempt to build a grammar by extracting regularities from the utterances they hear, with the result that they quickly converge on a very tidy, minimal, fully compositional grammar to express the meaning space required. Kirby explains this as a consequence of the "dynamics of language transmission", that is the fact that an agent cannot hope to sample utterances for the whole of the meaning space during its lifetime, and thus a language in which the meaning of a string can easily be predicted from its structure will stand a much better chance of being propagated from one generation to the next. Initially there is no such language, but if similarities between strings with similar meanings should occur by chance, and if the appropriate generalizations are made, structure within the language will be selected for.

The work described in this paper is a further investigation of the role of linguistic evolution, (as opposed to biological evolution), in the emergence of syntax. A key feature of the grammars that emerge from Kirby's simulations is that they use word order to specify meaning distinctions. However, it is worth noting that in many natural languages, meaning distinctions are not wholly specified by word order – even in English, some freedom of word order is allowed, and other languages allow much more. This is generally accompanied by a much richer case system than that found in English. The purpose of this work is to see if it is possible to use freedom of word order as a driving force towards the emergence of a case system.

[1] A key difference however, is that this relationship is symbiotic (unlike that between virus and host) because both host and language benefit from the other.

2 Modelling a Primitive Case System

The current work was carried out using an iterated learning model of language evolution based on that described by Kirby[7], to which the reader is referred for the details of the implementation.[2] At any given moment in time there exist two agents in the system, a learner and a speaker. The speaker is required to produce an utterance to express a meaning drawn from a simple meaning space. The utterance is made up of a string, composed of a number of alphabetic characters (intended to represent the irreducible phonemes of a language), plus a representation of that meaning. The meaning space is composed of "who did what to whom" type ideas, drawn from a set of five possible participants, plus five actions each requiring an agent and a patient. The speaker selects an appropriate string by consulting its own internal grammar, which, as in Kirby[7] is represented as context-free grammar enriched with simple semantics: non-terminal symbols have a single argument associated with them which conveys semantic information. If the agent's grammar does not specify an appropriate string for the required meaning, the agent will invent a new string (or a new part of a string) with a predefined probability (here simply set to 1). As the initial population has no grammar, this will be the only way that utterances can be produced in the early stages of the simulation. When presented with the speaker's utterance, the learning agent first attempts to see if it is covered by its own grammar, by attempting to parse the string. If the parse is successful and the correct meaning is returned, then the learner does nothing further. If not, the agent attempts to learn from the input according to the simple grammar learning heuristics described below. This process is repeated 100 times, at which point the current speaker is removed from the simulation, the current learner becomes the new speaker, and a new learner with an empty grammar is "born".

The grammar learning heuristics consist of two basic phases. The first is a simple incorporation step, whereby the utterance is added to the agent's grammar as a simple grammar rule of the form `s/[meaning]-->string`. The second stage is to make generalizations between this utterance and the others it has heard. This involves selecting pairs of rules and seeking to create a rule that will subsume them both. To this end, there are two operations available to the agent:

- If rules A and B differ only by non-terminals X and Y, and if changing Y to X would make them identical, then rule B is removed, and all other instances of Y in the grammar are changed to X.
- If the semantics of rules A and B differ by the value of a single element whose meanings are a and b, and their strings differ by substrings α and β, a and b are replaced by a variable x, and α and β are replaced by a non-terminal, N whose meaning is x. New production rules are created from N to strings α and β with meanings a and b respectively (effectively attributing the differences between the two meanings to the differences in the strings).

As in Kirby's system, the language spoken by the population of the simulation evolves over a number of generations from a simple vocabulary driven language, where each meaning is represented by an idiosyncratic string with no internal structure, to a fully compositional language, in which the meaning of the string is derived from the meaning of its parts, and the way they are assembled. In particular, separate syntactic categories for nouns and verbs emerge, and there is use of word order to distinguish between thematic roles.

Free word order languages do not emerge, and nor does the use of inflection to specify the distinction between thematic roles. This is not entirely surprising, given the nature of the heuristics that are used in grammar induction. These heuristics work by comparing strings and looking for common

[2] Although this system does not deal with recursion as Kirby's does.

prefixes and suffixes. When the parts of the strings which are the same have been identified, those which are different can be attributed to differences in the meanings of the two strings (as described above). Thus if presented with the strings *abcdef* meaning "John loves Mary", and *abcdgh* meaning "John loves Kate", the grammar inducer would identify the common prefix *abcd*, whilst noting that the final sections of the two strings differ. Thus the difference in meaning would be ascribed to this, resulting in the conclusion that *ef* means "Mary", whilst *gh* means "Kate". Suppose however, that the second string had been *ghabcd*, as might occur in a language that allows freedom of word order. This shares neither a common prefix nor suffix with the string *abcdef*, so the current grammar inducer would fail to notice any similarity between the two. As a result it would not pick out the relevant differences either. Thus if chance regularities between strings did occur such that a free word order language might be induced, the current grammar inducer would not be able to induce it.

However, natural languages do not tend to exhibit such rigid word order as those emerging from the simulation. Even English, which has a relatively strict ordering, allows a small degree of word order freedom, for example when the speaker wishes to topicalize the object. Other languages, such as German allow a lot more, and still others exist which allow almost complete freedom of word order. Clearly, in such languages, it is no longer possible to use word order to distinguish between thematic roles: if the language allows both SVO and OVS sentences, for example, and the string *johnlovesmary* is heard, how is the hearer to distinguish between the two possible meanings "John loves Mary" and "Mary loves John"? Instead, inflection is commonly used – different forms of the nouns *john* and *mary* to determine their case, i.e. whether they are subject or object of the sentence. Thus the two possible meanings can be distinguished by the form of each noun used.

It is noted that in the original system, occasionally a grammar with two noun categories rather than one emerges, so that agents are effectively using different subject and object versions of each noun, despite having a fixed word order. The current work is an attempt to see if it is possible to create a selective pressure for languages of this type. Will a degree of word order freedom cause the emergence of case for distinguishing thematic roles once word order is no longer a reliable cue? A number of changes were made:

- Firstly a function was added to "shuffle" the utterance made by the speaking agent with a fixed probability p, i.e. the parts of the utterance are re-ordered randomly. In the early stages of the simulation, where the grammar is composed entirely of idiosyncratic strings, this will have little effect, whereas later, once separate syntactic categories have begun to evolve, it will have the result of re-ordering the subject, object and verb components of the sentence. This is intended to model the occasional "mistake" or use of word order to provide emphasis by the speaker. Once an utterance with an alternative word order has been made, if it is incorporated into the learner's grammar, the rules producing that word order may well be used again, when that agent becomes a speaker. Thus it may be propagated down the generations. It is hoped, that by generating potentially ambiguous word orders, this will act as a driving force towards distinguishable subject and object versions of each noun, i.e. a primitive case system.

- Secondly it was necessary to enable the system to make use of multiple rules with different word orders. As the original system is deterministic, given a choice between two possible rules to generate an utterance for a given meaning, the same one will be used every time. Clearly this will not allow the propagation of alternative word orders, so the parser/generator was changed to work on a probabilistic basis. A count was associated with each rule in the grammar, according to the number of times that rule has been used. When generating utterances, if there is more than one rule that can produce a string with the required meaning, the probability of choosing a given rule is weighted according to its count.

- Finally, the fact that ambiguous utterances will be perfectly tolerated was addressed. Tolerance of ambiguity works against the emergence of case because there is no need for a means of disambiguation to become established. In order to overcome this, a *distinguishability* flag was added, which could be set to either 0 or 1 depending on whether or not the agents require utterances to be distinguishable. When the flag is set to 0, the system behaves as it did in its original incarnation, and is completely tolerant of ambiguity. The learner simply looks to see if it can parse the string to give the correct meaning, and if so, it is satisfied, and does nothing further. If the string cannot be parsed at all, or if the meaning returned is incorrect, it adds it to its grammar associated with the intended meaning. This means that a single string can be associated with any number of meanings. However, when the flag is set to 1, agents assume that each utterance is completely unambiguous. When presented with an utterance-meaning pair by the speaker, the learner again tries to parse that utterance, but this time is satisfied if it can parse the string to give *any meaning*. If a parse is possible, it assumes it already knows that utterance, *even if the meaning returned is incorrect*. It only incorporates a string into its grammar if it cannot parse it at all. So, according to the revised rules, if the speaker agent only has a single noun category, used to express both subjects and objects, and it then shuffles its output in such a way that the subject and object are inverted, the learner will never acquire the rule required to generate this sentence. However, if the speaker has two separate noun categories, one to represent each of subjects and objects, the learner will be able to acquire this rule. This should result in a selection pressure for separate subject and object noun categories.

3 Results

The results obtained from these changes to the simulation appear very promising at first. Running the system with p set to 0.01 and the distinguishability flag set to 1, the resulting grammars can be subdivided into two types. The first (Type A) has two separate noun categories for expressing subject and object and these grammars generally allow a large range of the possible word orders.[3] The second type of grammar (Type B) has only one noun category, used for both subject and object, but a more restricted range of word orders. However, word order is not entirely fixed: in general, approximately half of the possible word orders are represented. The set of possible word orders can be subdivided into pairs, which are identical but for the fact that the subject and the object have been transposed. If subject and object forms of any given noun are identical, this makes it impossible to determine which of the two rules is being applied and thus distinguish which noun is the subject and which is the object. Therefore, orderings from these pairs are mutually exclusive in the Type B grammars. If a sentence is made up of two nouns plus a verb and no other characters, this allows a total of six permutations. Generally three of these are expressed, without any loss of distinguishability of subject and object. For example, in the case where the word order SOV is allowed, SVO will not be, because this would cause confusion. Other rules such VSO can be perfectly easily distinguished from this however, due to the different positioning of the verb.

A typical Type B grammar looks like this:[4]

[3] Interestingly, although two separate noun categories have emerged, and within any given rule one of them is used for the subject of the sentence and one for the object, there is no consistency *between* rules: in the grammar above, three of the top level rules for building a sentence use category 1 to represent the subject and category 3 to represent the object, whilst one rule uses category 3 for subject and category 1 for object.

[4] Where P, X and Y are variables over predicates, subjects and objects, respectively.

```
s/[P, X, Y] --> [3/Y, 2/P, 3/X]
s/[P, X, Y] --> [2/P, 3/X, 3/Y]
s/[P, X, Y] --> [3/X, 3/Y, 2/P]
3/john   --> [q]
3/mary   --> [t]
3/pete   --> [u, f]
3/anna   --> [z, e]
3/kath   --> [r]
2/loves  --> [c]
2/hates  --> [r, a]
2/adores --> [i, t]
2/kisses --> [i]
2/sees   --> [m, j, g]
```

It can be seen that this grammar exhibits three of the six possible word orders for sentences sentences made up of two nouns, plus one verb: OVS, VSO and SOV. From each of the mutually exclusive pairs, SOV-OSV, VSO-OSV and SVO-OVS, only one is present. And of the three that do occur, the position of the verb makes it quite clear which rule is being applied, and thus it is possible to differentiate which noun is subject and which is object. This can be contrasted with the typical Type A grammar shown below. This one exhibits four possible word orders, OVS, SOV, VSO and SVO:

```
s/[P, X, Y] --> [3/Y, 4/P, 1/X]
s/[P, X, Y] --> [1/X, 3/Y, 4/P]
s/[P, X, Y] --> [4/P, 3/X, 1/Y]
s/[P, X, Y] --> [1/X, 4/P, 3/Y]
1/john   --> [i]
1/mary   --> [f, z, x]
1/pete   --> [h, n, v]
1/anna   --> [j]
1/kath   --> [y]
3/john   --> [a, t]
3/mary   --> [d]
3/pete   --> [l]
3/anna   --> [i, u]
3/kath   --> [q]
4/loves  --> [c]
4/hates  --> [k, h, k]
4/adores --> [h, i, x]
4/kisses --> [f]
4/sees   --> [l]
```

It should be noted that although only four of the possible six word orders are displayed in this grammar, it includes both members of one of the pairs of orderings which are mutually exclusive in the Type B grammar: OVS and SVO. This is made possible by the existence of two noun categories. Thus the meaning “John loves Mary”, using the OVS rule would be *dci* which is perfectly distinguishable from “Mary loves John” using the SVO rule, which is *fzxcat*, even though both sentences involve a word for “Mary” followed by a word for “loves” followed by a word for “John”.

Type A grammars seem to appear at a slightly higher frequency that Type B: 56% of the runs carried out exhibited a Type A grammar, and 44% a Type B. This is in contrast to the occurrence of separate noun categories in the original system, which only happened in approximately 24% of the runs. It would appear that the attempts to create a selective pressure for some form of case marking have been successful.

However, further investigation shows that things are not quite as they may seem. Re-running the simulation with p set to 0 (no shuffling) and the distinguishability flag also set to 0 (ambiguity is tolerated), i.e. as in the original system, only using the probabilistic parser resulted in 2-noun category grammars in 60% of the runs. There is also occasional spontaneous re-ordering of the sentence structure.

Returning to a p value of 0.01 (1% of utterances are shuffled) but with the distinguishability flag still set to 0 (ambiguity still tolerated), the number of grammars exhibiting case is reduced slightly to about 54% which is close to the value found with the distinguishability flag set to 1. However, the results are still quite different. For both grammars with case and grammars without case, all of the possible word order permutations are at some point added to the grammar. Whether or not they survive and are propagated to future generations is purely a matter of chance: with the probability of shuffling set at 0.01, it is unlikely that a single speaker will spontaneously produced the same "shuffled" sentence more than once, therefore the count associated with the rule specifying that particular word order will be very low. If that rule is chosen, it will be incorporated in the grammar of the next learner, but again it is unlikely to be chosen more than once, so again will result in a rule with a low probability of being chosen. Therefore most of the alternative word orders are quickly lost. However over the course of several thousand generations, rules for alternative word orders are learnt and propagated, and gradually many of the possible permutations are added to the grammar.[5] The pattern of word order restriction seen in the Type B grammar does not occur.

4 Conclusions and Further Work

The current work has shown that it is possible for a primitive case system (i.e. separate noun forms to represent subject and object) to emerge from a population of learners equipped with a simple learning algorithm for grammar induction, but no language specific knowledge. However, it would appear that it is not variability of word order that drives this emergence, as initially predicted. Nonetheless, once case has emerged, it certainly facilitates the use of alternative word orders, by helping to resolve some of the ambiguities that they may generate. This makes it possible for a language with a case system to support a wider range of word orders than one without. However, in those populations in which case does not emerge, the language need not be restricted to just a single word order, but to a very specific subset of those available: those which are easily distinguishable from each other by the positioning of other elements of the utterance.

This is an interesting result because such a pattern of restricted word order is quite unlike natural language, where it is generally the case that word order is either fairly strict, as in English, or allows a full range of word orders, e.g. languages such as Turkish and Serbo-Croatian[12]. Even when languages do exhibit a restricted set of word orders, such as Italian, which is predominantly SVO, but in which OVS, VSO and VOS are also relatively commonplace, this subset often includes pairs

[5] Generally no more than 5 different word orders are expressed at any one time however. This is because each alternative word order is added at the expense of the others – for any new order to get high enough counts to have a reasonable chance of survival, the others must have their counts reduced. If the count on any one of these is reduced too much, then it becomes likely that it will die out instead. Importantly though, there is no restriction on *which* word orders can appear together.

of word orders for which other cues are required to specify subject and object. Interestingly, Italian does not have a particularly rich case structure, and these alternative word orders are generally only allowed when the meaning can be disambiguated from the context.

As an extension to the present work, the "distinguishability flag" used to prevent agents from tolerating ambiguity has been removed, as this is clearly unrealistic. Instead, an additional phase has been added to the simulation in which teacher and learner converse with each other. No additional grammar rules are learnt during this phase, but if a string uttered by the conversational partner is parsed to give an incorrect meaning, then the rules used in that parse will be penalised, resulting in a decreased likelihood of those rules being used in the future. Early results from these simulations seem to show that in the case of grammars with only a single noun category, there is a much higher tendency towards a single dominant word order. Presumably this is because both versions of an ambiguous word order have an equal likelihood of being penalised, and are therefore equally likely to have their counts reduced to a very low level, and ultimately to disappear, as opposed to the situation where one rule gets established at the other's expense. However, this still does not result in a truly natural language-like distribution of word orders.

The results of this study suggest a number of possible future directions. First and foremost, an investigation of the mechanism by which the probabilistic parser drives agents to evolve case is required. Is the related to the occasional spontaneous re-ordering of sentences? And how does this spontaneous re-ordering occur? It was expected that the emergence of case would be driven by the need to disambiguate between possibly conflicting word orders, and yet adding this pressure does not appear to drive the system any further towards case-based languages, if anything it seems to slightly reduce their likelihood.

Secondly, it would appear that in natural language also, mechanisms other than the need to disambiguate are at work in the development of case. Otherwise there would be no reason why languages should exhibit completely fixed word order, and patterns such as that found in Italian would be somewhat unlikely. This may perhaps be related to sentence processing demands: Lupyan and Christiansen[10] have done studies using simple recurrent networks, which demonstrate that certain word orders are easier to learn than others. In particular, SVO and OSV are more readily acquired than SOV. In this case, the addition of case markings aids acquisition. This is in keeping with the fact that in natural language, SOV languages generally exhibit case markings, whereas most caseless languages are SVO or VSO. Therefore it appears that case markings may originally appear in order to facilitate the acquisition of fixed word order languages whose underlying word order is difficult to learn, but that their existence might enable more freedom in the word order used. In the current system, all possible word orders are equally easy to learn and equally likely. It would be interesting, therefore, to try and reproduce the effects of word order on learnability, and to see if this results in patterns of word order restriction more akin to those found in natural language.

Finally, the "case system" that emerges in the current study is far from representative of the type of case found in natural languages, which normally takes the form of inflectional affixes. In the present results, subject and object forms of the same noun are completely unrelated. Whilst this might occur for certain irregular wordforms that are very frequently used in a given language (such as "we" and "us"), it is not the norm. A planned follow-up, therefore, is to extend the current system to generalize across any chance regularities that may occur between subject and object forms of a given noun, or across different nouns of the same case, in the hope that a truly inflectional case system may be derived.

In conclusion then, although follow-up work is clearly required, the present study has demonstrated the emergence of a primitive form of case within the evolutionary framework described above.

References

[1] Jean Aitchison. *The Articulate Mammal*. Routledge, fourth edition, 1998.

[2] John Batali. Computational simulations of the emergence of grammar. In James Hurford, Chris Knight, and Michael Studdert-Kennedy, editors, *Approaches to the Evolution of Language: Social and Cognitive Bases*. Cambridge University Press, 1998.

[3] Ted Briscoe. Grammatical acquisition and linguistic selection. In E. J. Briscoe, editor, *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge University Press, 1999.

[4] Peter W. Culicover. *Principles and Parameters: An Introduction to Syntactic Theory*. Oxford University Press, 1997.

[5] Terrence Deacon. *The Symbolic Species: The Co-evolution of Language and the Human Brain*. Penguin Books, 1997.

[6] Simon Kirby. Language evolution without natural selection: From vocabulary to syntax in a population of learners. Technical report, University of Edinburgh, 1998. EOPL-98-1.

[7] Simon Kirby. Learning, bottlenecks and the evolution of recursive syntax. In Ted Briscoe, editor, *Linguisitic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge University Press, 1999.

[8] Simon Kirby. Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In Chris Knight, Michael Studdert-Kennedy, and James Hurford, editors, *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*. Cambridge University Press, 2000.

[9] Simon Kirby. Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions of Evolutionary Computation*, 5(2):102–110, 2001.

[10] Gary Lupyan and Morten H. Christiansen. Case, word order, and language learnability: Insights from connectionist modeling. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, to appear.

[11] Steven Pinker. *The Language Instinct*. Penguin, 1994.

[12] Dan I. Slobin and Thomas G. Bever. Children use canonical sentence schemas: A crosslinguistic study of word order and inflections. *Cognition*, 12:229–265, 1982.

[13] B Wilson and A M Peters. What are you cookin' on a hot? *Language*, 64:249–73, 1988.

Simulating language change with Functional OT

Gerhard Jäger
University of Potsdam

http://www.ling.uni-potsdam.de/~jaeger

1 Introduction

The research reported here is a reaction to recent work by Judith Aissen on the typology of case marking systems within Optimality Theory (OT). Aissen (2000) explains certain linguistic universals by assuming universal sub-hierarchies of OT constraints. I found this intriguing but unsatisfactory because these rankings are obviously much better adapted to the statistical patterns of actual language use than their inverse. Paul Boersma's Gradual Learning Algorithm for Stochastic OT is well-suited to establish a connection between frequencies of utterance types and constraint rankings, but it has to operate in generative and interpretive direction simultaneously. Intuitively, Aissen's sub-hierarchies are easier to acquire for such a bidirectional learning algorithm than possible (but non-existent) alternatives. This can be made precise by employing Iterated Learning in the sense of Kirby and Hurford (2002)—certain constraint ranking patterns are universal because they are evolutionary invariant. After briefly reviewing the empirical phenomena to be dealt with, Aissen's OT account, and the assumptions on OT learning that I employ in my version of Iterated Learning, I will report a series of simulations that link Aissen's findings with results of quantitative studies. A fuller account can be found in Jäger (2002).

2 Differential Case Marking

It is a common feature of many case marking languages that some but not all objects are case marked. However, it is usually not entirely random which objects are marked and which aren't. Rather, case marking only applies to a morphologically or semantically well-defined class of NPs. Bossong (1985) calls this phenomenon "Differential Object Marking" (DOM). A common pattern is that all NPs from the top section of the *definiteness hierarchy* or the *animacy hierarchy* are case marked while those from the bottom section are not.

(1) a. personal pronoun > proper noun > definite full NP > indefinite NP
b. human > animate > inanimate

Differential case marking also frequently occurs with subjects. In contradistinction to DOM, DSM ("Differential Subject Marking") means that only instances of some *lower* segment of the definiteness/animacy hierarchy are case marked. DSM usually co-occurs with DOM within one language. This phenomenon is called *split ergativity*. These patterns of "Differential Case Marking" (DCM) can be represented as the result of aligning two scales—the scale of grammatical functions (subject vs. object) with some scale which classifies NPs according to substantive features like definiteness or animacy. Ranking the grammatical functions according to prominence leads to the binary scale

(2) Subj $>$ Obj

Harmonic alignment of two scales means that items which assume comparable positions in both scales are considered most harmonic. For alignment of the scale above with the definiteness hierarchy this means that pronominal subjects (+prominent/+prominent), as well as indefinite objects (-prominent/-prominent) are maximally harmonic, while the combination of a prominent position in one scale with a non-prominent position in the other scale is disharmonic. More precisely, harmonically aligning the hierarchy of syntactic roles with the definiteness hierarchy leads to two scales of feature combinations, one confined to subjects, and the other to objects. The subject scale is isomorphic to the definiteness hierarchy, while the ordering for objects is reversed.

(3) a. Subj/pronoun $\succ$ Subj/name $\succ$ Subj/def $\succ$ Subj/indef
b. Obj/indef $\succ$ Obj/def $\succ$ Obj/name $\succ$ Obj/pronoun

In this way DCM can be represented as a uniform phenomenon—case marking is always restricted to upper segments of these scales. This pattern becomes even more obvious if optional case marking is taken into account. As Aissen (2000) points out, if case marking is optional for some feature combination, it is optional or obligatory for every feature combination that is lower in the same hierarchy, and it is optional or prohibited for every point higher in the same hierarchy. Furthermore, if one looks at actual frequencies of case marking patterns in corpora, all available evidence suggests that the relative frequency of case marking always increases the farther down one gets in the hierarchy. What is interesting from a typological perspective is that there are very few attested cases of "inverse DCM"—languages that would restrict case marking to lower segments of the above scales. The restriction to upper segments appears to be a strong universal tendency.

Prince and Smolensky (1993) develop a simple method to translate harmony scales into OT constraints: for each element x of a scale we have a constraint *x ("Avoid x!"), and the ranking of these constraints is just the reversal of the harmony scale. The constraint sub-hierarchies obtained in this way are assumed to be universal, i.e. language specific total ranking respects them.

Generally, the common pattern of DCM is that non-harmonic combinations must be morphologically marked while harmonic combinations are unmarked. To formalize this idea in OT, Aissen employs the formal operation of *constraint conjunction*. If C_1 and C_2 are constraints, $C_1 \& C_2$ is another constraint which is violated iff both C_1 and C_2 are violated. Furthermore, two general constraints play a role: "*$\emptyset$" is violated if a morphological feature is not marked, and "*STRUC" is violated by any morphological marking. Each constraint resulting from harmonic alignment is conjoined with *$\emptyset$, and the ranking of the conjoined constraints is isomorphic to the ranking induced by alignment. The alignment of the animacy hierarchy with the scale of grammatical functions thus for instance leads to the following universal sub-hierarchies:

(4) *$\emptyset$ & *Subj/inanim $\gg$ *$\emptyset$ & *Subj/anim
*$\emptyset$ & *Obj/anim $\gg$ *$\emptyset$ & *Obj/inanim

Interpolating the constraint *STRUC at any point in any linearization of these sub-hierarchies leads to a pattern where morphological marking indicates non-harmony. The choice of the threshold for morphological marking depends on the relative position of *STRUC.

3 Statistical bias

In Zeevat and Jäger (2002) (ZJ henceforth) we attempt to come up with a functional explanation for the DCM pattern that are analyzed by Aissen. The basis for this approach is the observation that har-

monic combinations of substantive and formal features (like the combinations “subject+animate” or “object+inanimate”) are common in actual language use, while disharmonic combinations (like “subject+inanimate” or “object+animate”) are rather rare. This intuition has been confirmed by several corpus studies. Table 1 displays the relative frequencies of feature combinations in the corpus SAMTAL, a collection of everyday conversations in Swedish that was annotated by Oesten Dahl. (Only subjects and direct objects of transitive clauses are considered,)

	NP	+def	-def	+anim	-anim
Subj	3151	3098	53	2948	203
Obj	3151	1830	1321	317	2834

Table 1: Frequencies in the SAMTAL corpus of spoken Swedish

There are statistically significant correlations between grammatical function and each of the substantive features definiteness, pronominalization and animacy. The correlations all go in the same direction: harmonic combinations are over-represented, while disharmonic combinations are under-represented. If attention is restricted to simple transitive clauses, the chance that an arbitrarily picked NP is a subject is (of course) exactly 50%—exactly as high as the chance that it is a direct object. However, if an NP is picked at random and it turns out to be definite, the likelihood that it is a subject increases to 62.9%. On the other hand, if it turns out to be indefinite, the probability that it is a subject is as low as 3.9%. Analogous patterns obtain for all combinations. I henceforth assume the working hypothesis that these statistical biases are universal features of language use.

4 Bidirectional Stochastic Optimality Theory

Aissen (2000) and Aissen and Bresnan (2002) point out that there is not just a universal tendency towards DCM across languages, but that DCM can also be used to describe statistical tendencies within one language that has, in the traditional terminology, optional case marking. Structural DCM can actually be seen as the extreme borderline case where these probabilities are either 100% or 0%. Stochastic Optimality Theory (StOT henceforth) in the sense of Boersma (1998) is a theoretical framework that is well-suited to formalize this kind of intuition. As a stochastic grammar, a StOT-Grammar does not just distinguish between grammatical and ungrammatical signs, but it defines a probability distribution over some domain of potential signs (in the context of OT: **GEN**).

StOT deviates from standard OT in two ways:

- **Constraint ranking on a continuous scale:** Every constraint is assigned a real number rather than a position in an ordinal hierarchy.
- **Stochastic evaluation:** At each evaluation, the placement of a constraint is modified by adding a normally distributed noise value. The ordering of the constraint after adding this noise value determines the actual evaluation of the candidate set at hand.

So we have to distinguish between the value that the grammar assigns to a constraint, and its actual ranking during the evaluation of a particular candidate.

An OT system consists of several constraints, and the addition of a noise value is done for each constraint separately. After adding the noise values, the actual values of the constraints define a total ranking. This total ordering of constraints is then used to evaluate candidates in the standard OT fashion, i.e. the strongest constraint is used first as a decision criterion, if there is a draw resort is taken to the second highest constraint and so on.

The probability for C1 $>$ C2 depends on the difference between their mean values that are assigned by the grammar. Let us denote the mean values of C1 and C2 as c1 and c2 respectively. Then the probability that C1 outranks C2 is a monotonic function of the difference between their mean values, c1$-$c2. If c1 $=$ c2, both have the same chance to outrank the other. This corresponds to a scenario where there is free variation between the candidates favored by C1 and those favored by C2. If C1 is higher ranked than C2, there is a preference for the C1-candidates. If the difference is larger than 12 units, the probability that C2 outranks C1 is less than 10^{-5}, which means that it is impossible for all practical purposes. In such a grammar C1 always outranks C1, and candidates that fulfill C2 at the expense of violating C1 can be regarded simply as ungrammatical (provided there are alternative candidates fulfilling C1, that is). So the classical pattern of a categorical ranking is the borderline case of the stochastic evaluation. It obtains if the distances between the constraints are sufficiently large.

Paul Boersma's Gradual Learning Algorithm (GLA) is an algorithm for learning a Stochastic OT grammar. It maps a set of utterance tokens to a grammar that describes the language from which this corpus is drawn. As a stochastic grammar, the acquired grammar makes not just predictions about grammaticality and ungrammaticality, but it assign probability distributions over each non-empty set of potential utterances. If learning is successful, they converge towards the relative frequencies of utterance types in the training corpus.

GLA operates on a predefined generator relation GEN that determines what qualifies as possible inputs and outputs, and which input-output pairs are admitted by the grammatical architecture. Furthermore it is assumed that a set CON of constraints is given, i.e. a set of functions which each assign a natural number (the "number of violations") to each element of GEN.

At each stage of the learning process, GLA assumes a certain constraint ranking. As an elementary learning step, GLA is confronted with an element of the training corpus, i.e. an input-output pair. The current grammar of the algorithm defines a probability distribution over possible outputs for the observed input, and the algorithm draws its own output for this input at random according to this distribution. If the result of this sampling does not coincide with the observation, the current grammar of the algorithm is slightly modified such that the observation becomes more likely and the hypothesis of the algorithm becomes less likely. This procedure is repeated for each item from the training corpus.

Optimality Theoretic evaluation is speaker oriented. It operates on a set of possible realizations of a hidden state (in the present context: possible forms corresponding to a given meaning). However, the statistical bias discussed above is only operative in the hearer direction. A hearer that is confronted with two structurally ambiguous NPs in construal with a transitive verb has to decide which one is subject and which is object. For animate NPs for instance the odds are that it is a subject while inanimates are most likely objects. To match these biases in a StOT grammar, Optimality Theory—and OT learning—has to operate in hearer direction as well.

The **Bidirectional Gradual Learning Algorithm** ("BiGLA" henceforth) differs from the original version in two respects. First, during the generation step the algorithm generates an optimal output for the observed input on the basis of a certain constraint ranking. It is tacitly assumed that "optimal" here means "incurring the least severe pattern of constraint violations" in standard OT fashion. In BiGLA it is instead assumed that the optimal output is selected from the set of outputs from which the input is *recoverable*. The input is recoverable from the output if among all inputs that lead to this output, the input in question incurs the least severe constraint violation profile (i.e. we apply interpretive optimization). If there are several outputs from which the input is recoverable, the optimal one is selected. If recoverability is impossible, the unidirectionally optimal output is selected.

This modification can be called **bidirectional evaluation**. Besides BiGLA involves **bidirectional learning**. This means that BiGLA both generates the optimal output for the observed input, and the optimal input for the observed output. "Comparison" and "adjustment" apply both to inputs and

outputs as well. The pseudo-code for BiGLA is:

Initial state All constraint values are set to the *initial value*.

for ($i := 0; i <$ *NumberOfObservations*$; i := i + 1$) {

Observation A training datum is drawn at random from the training corpus, i.e. a fully specified input-output pair $\langle i, o \rangle$.

Generation

- For each constraint, a noise value is drawn from a normal distribution N and added to its current ranking. This yields the *selection point*.
- Constraints are linearly ranked by descending order of the selection points.
- Based on this ranking, the grammar generates an optimal output o' for the input i and an optimal input i' for the output o using bidirectional evaluation.

Comparison If $i = i'$ and $o = o'$, nothing happens. Otherwise, the algorithm compares the constraint violations of the learning datum $\langle i, o \rangle$ with the self-generated pairs $\langle i, o' \rangle$ and $\langle i', o \rangle$.

Adjustment

- All constraints that favor $\langle i, o \rangle$ over $\langle i, o' \rangle$ or $\langle i', o \rangle$ are *promoted* by some small predefined numerical amount ("plasticity").
- All constraints that favor $\langle i, o' \rangle$ or $\langle i', o \rangle$ over $\langle i, o \rangle$ are *demoted* by the plasticity value.

}

5 BiGLA and DCM

Suppose the BiGLA is confronted with a language that has the same frequency distribution of the possible combinations of subject vs. object with animate vs. inanimate as the spoken Swedish from the SAMTAL corpus and uses case marking in exactly 50% of all cases, but in a way that is totally uncorrelated to animacy. We only consider simple transitive clauses, and we assume that this toy language has no other means for disambiguation besides case marking. So a learning datum will always be a combination of two NPs with a transitive verb. (I also assume that there are no verb specific preferences for certain readings of morphological markings.) Let us call the first NP "NP1" and the second one "NP2".

To see how BiGLA reacts to this language, we have to specify **GEN** and a set of constraints. Strictly speaking, animacy plays a double function in this experiment: it is of course an aspect of the meaning of an NP, but I also assume that this specification for +anim or −anim can be read off directly from the form of an NP. So +anim and −anim are treated as formal features, and **GEN** only relates animate meanings to +anim forms and inanimate meanings to −anim forms. There are thus eight possible semantic clause types to be distinguished because NP1 can be subject and NP2 object or vice versa, and both subject and object can be either animate or inanimate.

Let us assume that **GEN** supplies just one case morpheme, which is optional. The linking of this morpheme to a grammatical function is governed by the constraints, so **GEN** imposes no restrictions in this respect. **GEN** thus admits four types of morphological marking within a clause: both NP1 and NP2 can be case marked or unmarked. If $+/-$anim is taken into account, we get 16 different forms in total. However, **GEN** is organized in such a way that the animacy specification of the forms is

completely determined by the meaning. So we end up with altogether 32 meaning-form combinations that are consistent with this **GEN**.

As mentioned above, we extract the frequencies of the possible meanings from the SAMTAL corpus. The absolute numbers are given in table 2.
Not surprisingly, the combination where both subject and object are harmonic is by far the most frequent pattern, and the combination of two disharmonic NPs is very rare.

	subj/anim	subj/inanim
obj/anim	300	17
obj/inanim	2648	186

Table 2: Frequencies of clause types

Table 3 gives a frequency distribution (in per cent of all clauses in the corpus) over this **GEN** which respects the relative frequencies of the different meanings from SAMTAL and treats the linking of NP1 or NP2 to the subject role as equally likely. The notation "case1-case2" indicates that NP1 is marked with case1 and NP2 with case2 (M and Z abbreviate "marked" and "zero" respectively). Likewise, the notation "su/a-ob/i" means that NP1 is interpreted as animate subject and NP2 as inanimate object etc.

As for the constraint inventory, I basically assume the system from Aissen (2000) (restricted to the animate/inanimate contrast). This means we have four marking constraints. Using the same notation as in the table above, we can write them as *(su/a/z), *(su/i/z), *(ob/a/z), and *(ob/i/z). They all enforce case marking. They are counteracted by *STRUC which is violated by a clause as often as there are case morphemes present in a clause. (The evaluation of the constraints is done per clause, not just per NP.) The case morpheme may be interpreted either as ergative or as accusative case, and both interpretations are favored by one constraint each, m⇒su for ergative and m⇒ob for accusative. Finally I assume that the grammar does distinguish between interpreting NP1 or NP2 as a subject.

	M-M	M-Z	Z-M	Z-Z
su/a-ob/a	1.19	1.19	1.19	1.19
su/a-ob/i	10.50	10.50	10.50	10.50
su/i-ob/a	0.07	0.07	0.07	0.07
su/i-ob/i	0.74	0.74	0.74	0.74
ob/a-su/a	1.19	1.19	1.19	1.19
ob/a-su/i	0.07	0.07	0.07	0.07
ob/i-su/a	10.50	10.50	10.50	10.50
ob/i-su/i	0.74	0.74	0.74	0.74

Table 3: Training corpus

In real languages there are many constraints involved here (pertaining to syntax, prosody and information structure). In the context of our experiment, I skip over these details by assuming just two more constraints, SO and OS. They are violated if NP2 is subject and if NP1 is subject respectively. Since all constraints start off with the initial value 0, there is no *a priori* preference for a certain linking—these two constraints simply equip UG with means to distinguish between the two possible linking patterns. Altogether we thus get the nine constraints in table 4.

*(su/a/z):	*Avoid unmarked animate subjects!*
*(su/i/z):	*Avoid unmarked inanimate subjects!*
*(ob/a/z):	*Avoid unmarked animate objects!*
*(ob/i/z):	*Avoid unmarked inanimate objects!*
m⇒su:	*Marked NPs are subjects.*
m⇒ob:	*Marked NPs are objects.*
*STRUC:	*Avoid case marking!*
SO:	*NP1 is subject and NP2 object.*
OS:	*NP2 is subject and NP1 object.*

Table 4: Constraint inventory

6 Iterated learning and DCM

Depending on the OT system that is used, the training corpus and the chosen parameters, the stochastic language that is defined by the acquired grammar may deviate to a greater or lesser degree from the

training language. Especially for BiGLA this deviation can be considerable. (It is perhaps misplaced to call BiGLA a "learning" algorithm; it rather describes a certain adaptation mechanism.) If a sample corpus is drawn from this language and used for another run of BiGLA, the grammar that is acquired this time may differ from the previously learned language as well.

Such a repeated cycle of grammar acquisition and language production has been dubbed the *Iterated Learning Model* of language evolution by Kirby and Hurford (2002).

The production half-cycle involves the usage of a random generator to produce a sample corpus from a stochastic grammar. In the simulations, I assumed that this sample corpus has the same absolute size than the initial corpus. Furthermore I assume that the absolute frequencies of the different *inputs* are kept constant in each cycle. What may change from cycle ("generation") to cycle are the relative frequencies of the different outputs for each input.

A given stochastic grammar G defines a probability distribution $p_G(\cdot|i)$ over the possible outputs o for each input i. Using a random generator, this probability distribution can be used to generate a new corpus that represents the acquired grammar. One cycle of learning and production represents one generation in the evolutionary process that is simulated. This cycle my be repeated arbitrarily many times, i.e. over an arbitrary number of generations.[1]

In the first simulation to be reported here I used the constraint inventory, generator relation and corpus frequencies given above as initial input for iterated learning. The successive constraint rankings that emerge in this way are plotted in figure 1. The learning procedure was repeated 500 times, and the generations are mapped to the x-axis, while the y-axis again gives the constraint rankings.

While there are no rough changes from one generations to the next, the grammar as a whole gradually changes its characteristics over time. Aissen' sub-hierarchies—*(su/i/z) $\gg$ *(su/a/z) and *(ob/a/z) $\gg$ *(ob/i/z)—are invariant though.

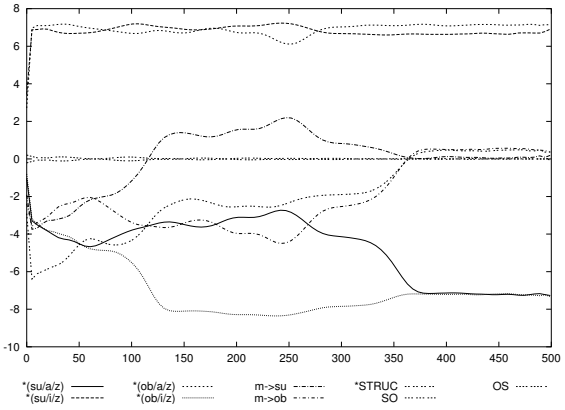


Figure 1: Diachronic development

We may distinguish four phases. During the first phase (generations 1–10), the constraints *(su/i/z) and *(ob/a/z) stay closely together, and they increase their distance from *STRUC. This amounts to an ever stronger tendency for case marking of disharmonic NPs. Simultaneously, *(su/a/z) and *(ob/i/z) stay close to *STRUC, i.e. we have optional case marking of harmonic NPs. This corresponds to a split ergative system with optional marking of harmonic and obligatory marking of disharmonic NPs. This characteristics remains relatively stable during the second phase (roughly generations 11–100). Then the system becomes unstable. After another thirty generations, it enters a relatively stable state where case marking of inanimate objects is completely lost while case marking of animate subjects is still optional. Case marking of disharmonic NPs remains obligatory. This situation remains constant (with some minor variation) for about 200 generations, when case marking of animate subjects is lost as well. The constraint ranking now reached is

$$\{\text{*(su/i/z), *(ob/a/z)}\} \gg \{\text{m}\Rightarrow\text{su, m}\Rightarrow\text{ob, *STRUC, SO, OS}\} \gg \{\text{*(su/a/z), *(ob/i/z)}\}$$

Needless to say that the diachronic development that is predicted by the BiGLA (together with **GEN**,

[1]The software package *evolOT* implements this version of the Iterated Learning Model. It is freely available from http://www.ling.uni-potsdam.de/~jaeger/evolOT.

the constraint set, and the probability distribution over meanings from SAMTAL) depends on the pattern of case marking that was used in the first training corpus. A full understanding of the dynamics of this system and the influence of the initial conditions requires extensive further research. In the remainder of this section I will report the results of some experiments that give an idea of the overall tendencies though.

If the first training corpus contains no case marking at all (a somewhat unrealistic scenario, given that the **GEN** supplies case morphemes—perhaps this models the development of a language immediately after some other device has been reanalyzed as case morpheme), the overall development is similar to the previous setup. The ranking that BiGLA induces from the initial corpus places *STRUC extremely high (at 55.79), while the constraints that favor case marking are placed much lower, thus reflecting the absence of case marking. Still, the Aissen sub-hierarchies are respected, with *(su/a/z) at -33.04, *(su/i/z) at 5.03, *(ob/a/z) at 1.04 and *(ob/i/z) at -29.03. However, case marking of disharmonic NPs is gradually acquired within a few generations, and after thirty generations the system already enters the steady state of split ergativity (see figure 2).

It was mentioned in the beginning that DCM is a strong universal tendency. There are very few languages with an inverse DCM pattern. This is predicted by the assumption of Aissen's universal sub-hierarchies: there cannot be a language that marks animate subjects with higher probability than inanimate ones, say. It is revealing to run the BiGLA on a training corpus with such an (allegedly impossible) pattern. I did a simulation with a training corpus where all and only the harmonic NPs were case marked. The development of the constraint ranking is given in figure 3.

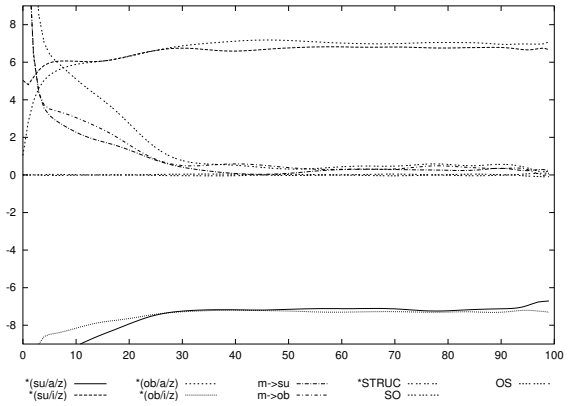


Figure 2: No case marking in the initial state

The BiGLA in fact learns the inverse pattern, i.e. it comes up with a grammar where the Aissen sub-hierarchies are reversed: *(su/a/z) $\gg$ *(su/i/z) and *(ob/i/z) $\gg$ *(ob/a/z). Accordingly, the language that is learned in the first generation marks almost all harmonic NP but nearly no disharmonic ones. So UG admits such a language, and it is also learnable. However, it is extremely unstable. Already after twelve generations the Aissen sub-hierarchies emerge and remain stable for the remainder of the simulation. Nonetheless, the case marking patterns changed dramatically after that. For about 100 generations after the emergence of the Aissen hierarchies, case marking is obligatory for disharmonic and optional for harmonic NPs. After that, the system dramatically changes its character and enters a state of a pure ergative system, i.e. all subjects and no objects are case marked. Around generation 1000 (not included in the graphics) the system switched into a split ergative state, as in the first two experiments.

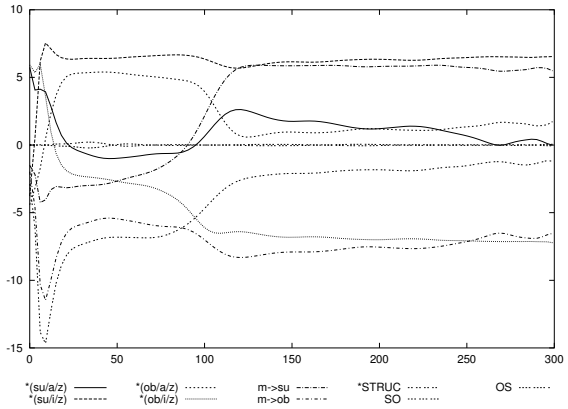


Figure 3: The future of anti-DCM

While these simulations establish a connection between the statistical patterns of language use and

the independently motivated constraint hierarchies postulated by Aissen, the experimental results are at odds with the actual typological tendencies. Languages with split ergativity are a minority among the languages of the world. The majority of languages follows a nominative-accusative pattern, often combined with DOM. It is a matter of dispute whether pure (morphological) ergative languages exist at all, and in any case they are very rare. How do these facts relate to the predictions of iterated learning? I will conclude this section with some speculations about the typology of case marking patterns within the paradigm of iterated learning using BiGLA.

The dynamics of the system is very sensitive to the relative frequencies of the different meanings. The emergence of Aissen's sub-hierarchies is due to the fact that there are much more clauses of the type "animate subject – inanimate object" than the inverse type. The clauses where both arguments are of the same animacy are irrelevant here. Their relative frequency is decisive for the precise nature of the steady states though. In the SAMTAL corpus, the number of clauses were both arguments are animate (300) has the same order of magnitude as the number of clauses with two inanimate arguments (186). If we look at definiteness instead, this is different. Here the frequencies are as in table 5. There are about sixty times as many clauses with two definite arguments as clauses with two indefinite NPs. Feeding a training corpus with these relative frequencies and 50% probability of case marking for each NP type into iterated BiGLA gives a qualitatively different trajectory than in the previous experiments. It is given in figure 4.

	subj/def	subj/indef
obj/def	1806	24
obj/indef	1292	29

Table 5: Frequencies of clause types with respect to definiteness

Here the system reaches a steady state after about 70 generations. The emerging ranking is the following (where "*(ob/d/z)" stands for "Avoid unmarked definite objects!" etc.):

$$\{*(obj/d/z),\ m{\Rightarrow}obj\} \gg *(obj/i/z) \gg \{*(subj/i/z),\ SO,\ OS\} \gg *STRUC \gg *(su/d/z) \gg m{\Rightarrow}su$$

This grammar seems to describe a language with obligatory object marking and DSM. However, recall that **GEN** only supplies one case morpheme here, and the sub-hierarchy m⇒obj ≫ m⇒su ensures that this morpheme is unequivocally interpreted as accusative. Thus ergative marking is impossible and the constraint ranking describes a language with obligatory object marking and no subject marking.

To sum up the findings from this section, we may distinguish several types of case marking patterns according to their likelihood. Most unlikely are languages that violate UG, i.e. where there is no constraint ranking that describes such a language. If we assume a UG as above (i.e the **GEN** and set of constraints discussed in the previous section), there can't be a language where either both subject and object or neither are case marked. (Feeding such a corpus into BiGLA leads to a language where about 60% of all clauses contain exactly one case marker.) Note that it is extremely unlikely but not impossible to find a corpus with this characteristics, because this language is a subset of many UG-compatible languages. Such a corpus would be highly un-representative though.

The next group consists of languages that correspond to some constraint ranking but are not learnable in the sense that exposing the BiGLA to a sample from such a language leads to a grammar of a substantially different language. The language without any case marking would fall into this category (provided **GEN** supplies case marking devices). There is a ranking which describes such a language, namely

$$*STRUC \gg \{OS,\ SO\} \gg \{*(su/a/z),\ *(su/i/z),\ *(ob/a/z),\ *(ob/i/z),\ m{\Rightarrow}su,\ m{\Rightarrow}ob\}$$

However, if the BiGLA is exposed to a sample from this language, it comes up with a substantially different ranking, namely

*STRUC ≫ {m⇒su, m⇒ob} ≫ *(su/i/z) ≫ *(ob/a/z) ≫ {OS, SO} ≫ *(su/a/z) ≫ *(ob/i/z)

11.1% of the NPs in a sample corpus drawn from this language do carry case marking.

The third group consists of languages that are both in accordance with UG and learnable, but diachronically instable. This means that the BiGLA acquires a language that is similar but not entirely identical to the training language, and that the deviation between training language and acquired language always goes into the same direction. Diachronically this leads to a change of language type after some generations. This can be observed most dramatically with languages with inverse DCM (compare figure 3). There the language type switches from inverse split ergativity to optional split ergativity within less than twenty generations.

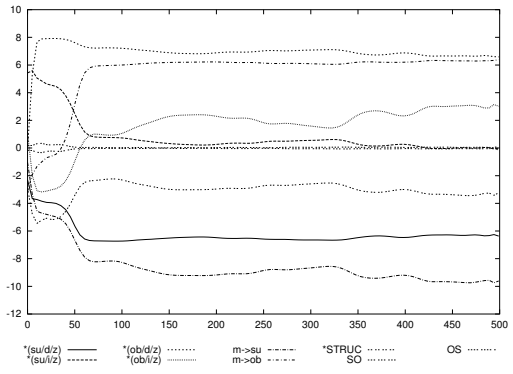


Figure 4: Stabilization at accusative system

The most likely language types are those that are diachronically stable and are additionally the target of diachronic change in many cases. The experiments conducted so far indicate that there is exactly one such steady state for each experimental setup—split ergativity in the first and nominative-accusative in the second scenario.

Given the extremely coarse modeling of the factors that determine case marking in our experiments and the fact that the experiments all depend on a probability distribution over meanings that is based on just one corpus study, these results have to be interpreted with extreme caution. They fit the actual patterns of typological variation fairly well though, so it seems worthwhile to pursue this line of investigation further.

References

Aissen, J. (2000). Differential object marking: Iconicity vs. markedness. Manuscript, UCSC.

Aissen, J. and J. Bresnan (2002). OT syntax and typology. course material from the Summer School on Formal and Functional Linguistics. University of Düsseldorf.

Boersma, P. (1998). *Functional Phonology*. Ph.D. thesis, University of Amsterdam.

Bossong, G. (1985). *Differentielle Objektmarkierung in den neuiranischen Sprachen*. Günther Narr Verlag, Tübingen.

Jäger, G. (2002). Learning constraint sub-hierarchies. The Bidirectional Gradual Learning Algorithm. manuscript, University of Potsdam.

Kirby, S. and J. R. Hurford (2002). The emergence of linguistic structure: An overview of the Iterated Learning Model. In A. Cangelosi and D. Parisi, eds., *Simulating the Evolution of Language*, pp. 121–147. Springer, London.

Prince, A. and P. Smolensky (1993). Optimality theory: Constraint interaction in generative grammar. Technical Report TR-2, Rutgers University Cognitive Science Center, New Brunswick, NJ.

Zeevat, H. and G. Jäger (2002). A reinterpretation of syntactic alignment. In D. de Jongh, M. Nielsenova, and H. Zeevat, eds., *Proceedings of the Fourth International Tbilisi Symposium on Language, Logic and Computation*. University of Amsterdam.

Modelling Zipfian Distributions in Language

Catriona Tullo and James R Hurford
University of Edinburgh
http://www.ling.ed.ac.uk/~jim

1 Introduction

G.K.Zipf famously discussed a number of patterns in the distributions of linguistic units, such as words and phonemes, in texts. We address several of these here, and attempt to explain their origins in terms of simple principles of language use, including, but going beyond, Zipf's own 'Principle of Least Effort'.

1.1 Rank/Frequency and Length/Frequency Correlations

The term "Zipfian distribution" refers to "a distribution of probabilities of occurrence that follows Zipf's Law". Zipf's law is an experimental law, not a theoretical one; i.e. it describes an occurrence rather than predicting it from some kind of theory. The observation that, in many natural and man-made phenomena, "The probability of occurrence of ... items starts high and tapers off. Thus, a few occur very often while many others occur rarely." The formal definition of this law is: $\mathbf{P_n = 1/n^a}$, where $\mathbf{P_n}$ is the frequency of occurrence of the **n**th ranked item and **a** is close to 1.

Applied to language, this means that the rank of a word (in terms of its frequency) is approximately inversely proportional to its actual frequency, and so produces a hyperbolic distribution. To put Zipf's Law another way: $fr = C$, where: r = the rank of a word, f = the frequency of occurrence of that word, and C = a constant (the value of which depends on the subject under consideration). Essentially this shows an inverse proportional relationship between a word's frequency and its frequency rank[1]. Zipf calls this curve the 'standard curve'. Texts from natural languages do not, of course, behave with such absolute mathematical precision. They cannot, because, for one thing, any curve representing empirical data from large texts will be a stepped graph, since many non-high-frequency words will share the same frequency.

[1]Note that this generalization is distinct from another frequency pattern also noted by Zipf, namely that $nf^2 = K$, where f is the frequency of some word, n is the number of words occurring f times in a text, and K is a constant (Zipf, 1935).

But the overall consensus is that texts match the standard curve significantly well. Li (1992:1842) writes "This distribution, also called Zipf's law, has been checked for accuracy for the standard corpus of the present-day English [Kučera & Francis, 1967] with very good results". See Miller (1951:91-95) for a concise summary of the match between actual data and the standard curve.

Zipf also studied the relationship between the frequency of occurrence of a word and its length. In *The Psycho-Biology of Language* (1935), he stated that "it seems reasonably clear that shorter words are distinctly more favoured in language than longer words." So a very few shorter words are used (spoken or written) very frequently, while others are used very rarely. Zipf did not specifically claim that the same Law which describes the connection between word rank and frequency also applies to frequency and length. He merely stated that there is a general tendency for word length to decrease as word frequency increases, "the magnitude of words tends, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of occurrences", (Zipf, 1935). Nor did he expand on any possible mathematical formula to model this. It seems clear that the general length/frequency correlation is realized more messily in texts than the rank/frequency correlation. Nevertheless it is clear that there is a gross similarity between the Rank/Frequency and the Length/Frequency curves observable in linguistic texts. Both are roughly 'J-shaped', and we will refer to both distributions under the broad heading of 'Zipfian distribution'.

1.2 What a Zipfian Distribution can Explain

Several studies (Kirby, 2001; Onnis et al., 2002) propose models which explain the emergence of irregularities in language. Thus, beside the Zipfian correlations, linguists are familiar with the fact that there is also a correlation between the frequency of words and constructions and their morphological or syntactic irregularity. For instance, the most frequent verbs in a language are also those most likely to be irregular. The cited studies report computer simulations in which successive generations learn their language from the statistically biased usage of previous generations. Agents in these simulations use words or grammatical patterns with varying frequency, determined by an assumed Zipfian distribution. Taking such a Zipfian distribution as given *a priori*, the observed correlation between frequency and irregularity emerges as a stable property of the language of a simulated population.

But of course, the assumed Zipfian distributions themselves remain to be explained.

1.3 What can Explain a Zipfian Distribution?

George Miller (1965) observed that a random text would be expected to exhibit Zipfian distributions. A random text is generated by iteratively emitting random letters from a given alphabet including the space character. Maximal strings of non-space characters are counted as 'words' in such a text. Miller's idea is taken up by Li (1992) who proves that such rank/frequency and rank/length correlations are indeed to be expected in any random text of reasonably large size.

So what? Miller thought that the fact that random texts give rise to Zipf-like distributions invalidated Zipf's own 'Least Effort' explanation for the Length/frequency correlation. "It seems, therefore, that Zipf's rule can be derived from simple assumptions that do not strain one's credulity . . . , without appeal to least effort" (Miller, 1957). "Zipf's curves are merely one way to express a necessary consequence of regarding a message source as a stochastic process" (Miller, 1965). In other words, Zipf, in proposing his Least Effort Hypothesis, had not eliminated the Null Hypothesis. Miller's argument is that if the distributions to be accounted for emerge from random texts, the Null Hypothesis can account for them, and there is no need of any further explanatory mechanism, such as a Law of Least Effort. But is this a reasonable Null Hypothesis? Of course, texts in natural languages are not generated by random emission of phoneme-sized elements. They are not even generated by emission of words randomly picked from a lexicon (a zero-order approximation to a natural language – Shannon, 1948). Why, for example, could it not be an equally reasonable Null Hypothesis that all words are equiprobable? Then, of course, the Zipfian distributions would appear strikingly significant, and in need of explanation.

What strains credulity is surely Miller's idea that human language results from a stochastic monkey-and-typewriter scenario. Mathematical derivations of Zipf-like distributions from random texts deliberately ignore the fact that natural language texts are produced by intentionally communicating agents. A reasonable explanation for Zipf-like distributions should be embedded in a theory which makes realistic assumptions about the causal factors which give rise to natural language texts. Here we present such a model. Our model retains Miller's idea that the message source is a stochastic process, but situates this process in a more realistic human culturo-linguistic scenario. Our work also dovetails with models, such as Kirby's and Onnis et al.'s, which assume Zipfian distributions among their given initial conditions. Thus the present study can be seen as complementing those studies by deepening their foundations.

There exist two sources of meaning choice for an average speaker. The first is the environment. A speaker may react to something in their surroundings by talking about it. A commonly used example of this is the habit many people have of starting a conversation by talking about the weather. For the purposes of this discussion we assume for the moment that, from the point of view of the environment, all meanings are equally likely to be chosen by the speaker (i.e. the environment has an even frequency distribution).

Once a dialogue has begun, however, another source of word choice is available – those words used, or heard, in the preceding conversation. For example if the first speaker in a conversation (S1) begins by mentioning that a person, Fred, has gone on holiday, it is likely that S1's conversational partner, S2, will carry on the discussion by talking about one of the two topics started by S1, i.e. Fred or holidays. Any attempt to change the topic to something completely unrelated would cause a certain amount of confusion on the part of S1 and possibly a breakdown in the conversation. This means that words related to Fred, or holidays, are likely to be used most often in this discussion. So from the entire vocabulary of the language there is a small subset of words which have a much higher chance of selection.

Our hypothesis is that it is the topics or meanings used in the preceding dialogue which give the Zipfian distribution to the frequency curve of words. Some words are used more frequently than others because language users hear them more frequently than other words. So if, for some reason, a word is spoken more frequently for a while, it will then become spoken even more frequently because it has been heard more often. So, in essence, the more frequent a word is, the more frequent it will become. This should happen even if, to start with, all words are given an even frequency distribution, and speakers initially choose words at random. By chance, some words will be selected more often than others will. This inevitably leads to an uneven frequency distribution developing in the language. Our model takes a lead from Harremoës & Topsøe (2001); they write "The child gets input from different sources: the mother, the father, other children, etc. Trying to imitate their language with frequencies which are closer to Zipf's laws than the sources. As a language develops during the centuries the frequencies will converge to a hyperbolic distribution." (A conversation with Jörg Rieskamp also helped to inspire this model.)

2.1 Testing the Model by Simulation: Rank/Frequency

A computer model was devised to test the coherence of this hypothesis. A set containing the numbers 1 to 1000 was created. This corresponds to the initial word-store of the language, with each number

representing a word, so this vocabulary is created with an even frequency distribution. From this first set 1000 ‘words’ are selected at random one at a time, and are used to create the word-store for the next generation. When a word is chosen it is copied, not moved, into the new word-store. This makes it possible for the new word-store to contain more than one instance of certain words whilst having failed to copy other words which appeared in the original set. This process is repeated, with each new word-store being selected from that of the generation before (not from the original 1000 numbers with even distribution).

Thus when a new word-store is being built some words will be selected more than once and others will not be selected at all for the new group. This means that over time, although the actual size of the corpus of words will remain constant at 1000, the size of the vocabulary (i.e. the number of different words) will tend to decrease, as some words are lost due to not being selected.

(Note that this model falls within the ‘iterated learning’ paradigm as developed by Kirby in several publications (e.g. Kirby & Hurford, 2000), also called the ‘E/I’ (Expression/Induction) class of models in Hurford (2002).)

We evaluate the results of our simulations impressionistically as follows. A true Zipfian curve is a simple hyperbola. A simple hyperbolic graph (of the equation $rf = C$), drawn on a double logarithmic scale, is a straight diagonal line from top left to bottom right. When the data obtained from the model is plotted on the same graph as this hyperbola it is possible to compare the two lines.

This basic model was run with various additional features, as described in the following subsections.

2.1.1 Drastic Vocabulary Loss in the Basic Model

The results from running the basic model showed that the distribution of words produced was not quite ‘J shaped’. There is a skewed distribution evident in the results, but it does not have the almost exponential shape of the true Zipfian distribution, as can be seen in figure 1. This shows the frequency of a word plotted against it’s rank after a single run of 100 generations.

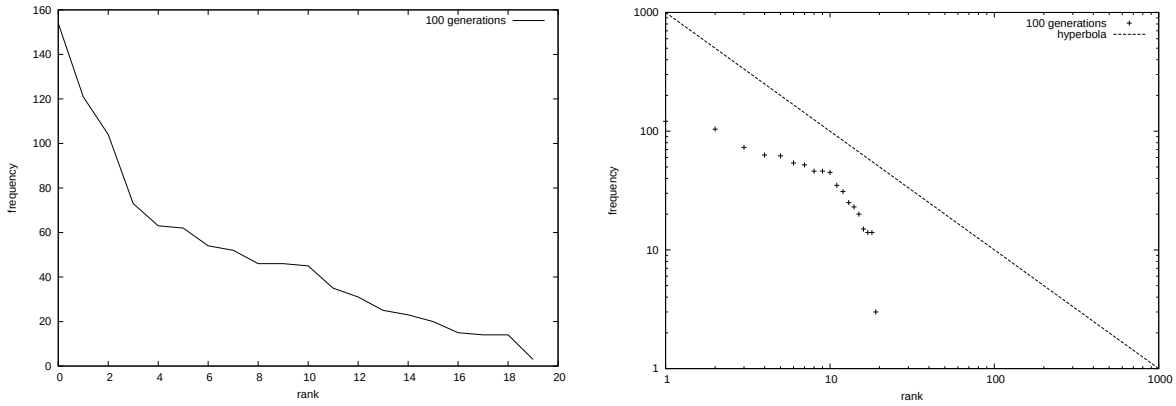


Figure 1. Left panel: The frequency of a word plotted against its rank after 100 generations. Note that there are only 19 words left in the vocabulary at this point.
Right panel: Frequency-rank distribution after 100 generations plotted on double log scale.

After 100 generations (i.e. 100 successive vocabularies) from a starting vocabulary of 1000 there are only on average only 20 different words left. This is a dramatic reduction and would have devastating effects on any real language if word-types were actually lost in this way. As the right panel of figure 1 shows, the word-stores resulting from this basic model do not have a perfectly hyperbolic, or Zipfian, distribution, although they approach it.

This different distribution could be due to the fact that in this basic model words die out completely, and there is no way of them ever being reintroduced into the language. In real languages some words are used a lot less than others, but they aren't necessarily lost from the language completely. A way of randomly re-introducing some words back into the language, or alternatively, preventing words from being lost completely, is the next step.

2.1.2 Meaning Choice Partly Triggered by the Environment

Our basic model investigated the effect of sampling previous discourse on word selection. But the environment clearly also has some influence on word choice. Speakers do not simply carry on talking about one topic for entire conversations; changes in subject matter do occur. Some of these may come from associations with the current discourse, but another source of topic is the speakers' surroundings.

To introduce this effect into the existing model a stable set containing all of the possible meanings in the language (i.e. here the numbers 1 to 1000) was preserved in the background during the runs. This represents the environment. At the beginning of a run a variable 'Environment Parameter', *E*, was

set, representing the probability of the next word to be put into the word-store coming directly from the environment, rather than from a sampling of the word-store of the previous generation. For example, where *E = 0.1*, each time a new word is selected to be added to the word-store, it has a 0.1 probability of being randomly selected from the complete environmental set of meanings, and a 0.9 probability of being randomly selected from the word-store of the previous generation, as in the basic version of the model..

It was found that as the probability of a new word being selected from the environment, rather than from the preceding conversation, was increased, the time taken for the vocabulary to decrease, and take on a Zipf-like distribution, increased. Figure 2 shows how the rank-frequency distribution after 100 generations changes with different weightings of the environment. From this it can be seen that this modification has achieved results approaching, but still somewhat far from, a hyperbolic distribution.

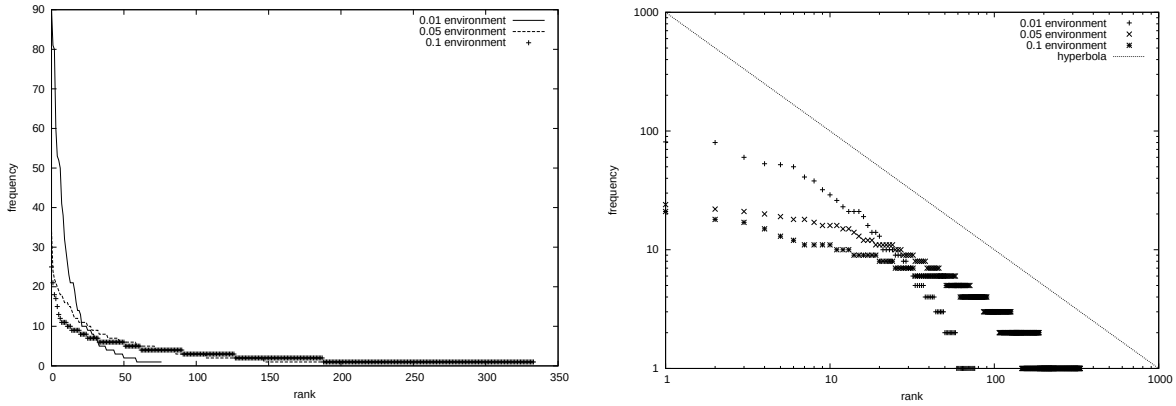


Figure 2: Rank-frequency distribution for differing environmental influence – *E* = 0.01, 0.05 and 0.1. The right panel is on a double log scale, for comparison with a hyperbolic curve (straight line).

2.1.3 Corpus Size

Thus far in this paper we have assumed that the speaker in each generation will select only 1000 words every time. The number of different word types available to the speaker initially may only be 1000, but a realistic assumption is that far more word tokens than this will actually be spoken (or put into each new word-store at each generation).

To model this the original set is left unchanged, hence the environment still contains only one instance of each word. However in every subsequent generation more than 1000 word tokens can be selected. This number is kept constant throughout every run of the model, therefore each word-store (after the original environment) will contain the same number of tokens. So although there are still only 1000

different lexical items in the language, from this, for example, 2000 word tokens can be presented at each generation. For the initial run of this, the level of the influence of the environment was 0.

It was found that increasing the corpus size again slowed the decrease in the size of the vocabulary. This is illustrated in figure 3 (left panel). Altering the corpus size causes the frequency distribution of the words to change markedly, as shown in figure 3 (right panel).

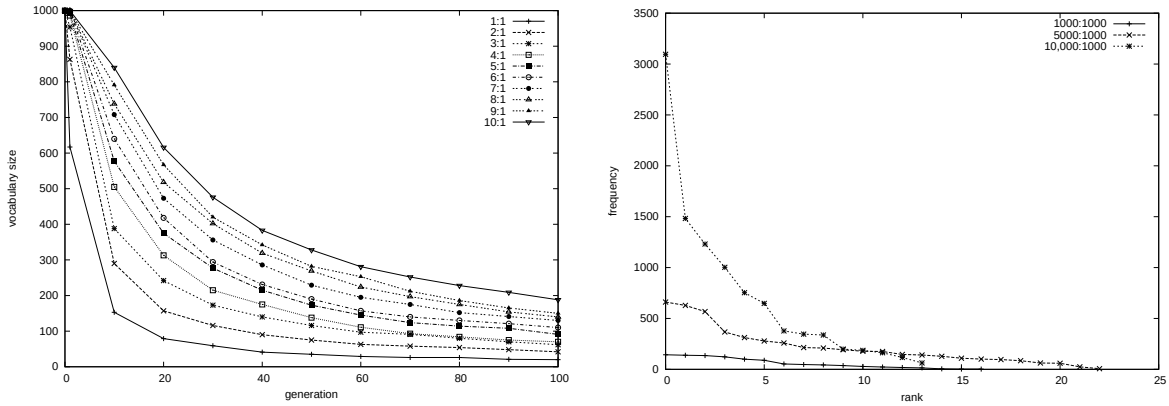


Figure 3. Results with different corpus sizes, selected from a vocabulary of 1000.
Right panel: Change in vocabulary size at each generation with differing corpus size.
Left panel: Rank-frequency distribution for corpus sizes of 1000, 5000 and 10,000.

This effect is seen because increasing the number of selections means that the initial selection from the environment contains a greater variety of words. Hence the size of the vocabulary decreases at a much lower rate than in previous runs. However, the Zipfian distribution can still emerge, albeit a little more slowly. No matter how many selections are made, some will still be selected more than others, and these words will have a greater chance of being selected in the next generation as well.

2.1.4 Combining Environmental Influence with Corpus Size

If this effect were combined with a more stable vocabulary (where words didn't die out so rapidly) a hyperbolic distribution may emerge, given that in earlier trials the environment had a positive effect on the rank-frequency distribution of a vocabulary. To this end the effect of the environment was re-introduced. The probability of a word being selected from the environment was kept small at 0.05.

It was found that allowing selections from the environment meant that the vocabulary size more or less stabilised after approximately 20 generations for each of the runs. Figure 4 (left panel) illustrates this effect (and the also the different sizes at which stabilisation occurred). Reintroducing the environmental

influence meant that the result of increasing the corpus size was even more apparent. The frequency distribution of the words continued to change. To test whether this change was towards a more Zipf-like distribution the same test was applied as before. The results were plotted on a double logarithmic scale alongside a hyperbolic line. The result of this is shown in figure 4 (right panel).

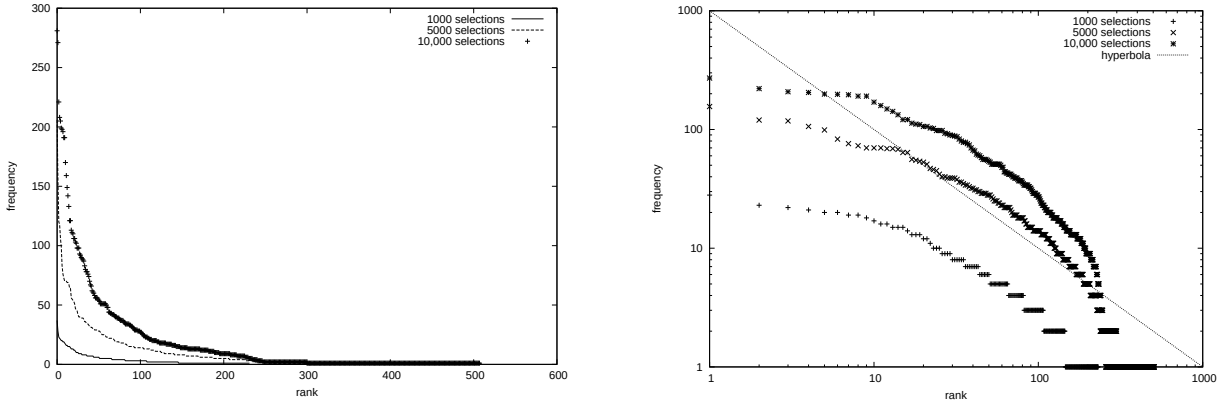


Figure 4: Combining environmental influence and varying corpus size, from a vocabulary of 1000.
Left panel: Vocabulary size against generation for corpus sizes of 1000, 5000 and 10,000 .
Right panel: Log-scale graph showing the rank-frequency distribution after 100 generations for corpus sizes of 1000, 5000 and 10,000.

This clearly shows that, although the lines are closer to the hyperbole than in earlier graphs, there is still not an exact hyperbolic distribution in the frequency of words in this model. Perhaps however this is an unrealistic expectation for a model which uses only a small number of words. Even with the highest corpus size (10,000) the size of the vocabulary is still only about one third of that used by Zipf in his original study (of the distribution of words in James Joyce's *Ulysses*).

2.2 Testing the Model: Length/Frequency

The model as developed so far makes no mention of word length. At the beginning, we undertook to explain the Zipfian length/frequency correlations as well as the rank/frequency correlations. In keeping with our general approach, modelling the cultural transmission of word-stores across generations by learning, we now introduce a factor of word-length, and apply a simple implementation of Zipf's own least effort principle. Recall that Zipf made less precise statistical claims about the length-frequency corelation.

We hypothesise that the more often a word is spoken the more likely it is that at some point the

signal passing between the two individuals involved in a conversation will be degraded somehow. For example, if two people are talking in a noisy environment it is possible that a word spoken by S1 will be heard differently by S2 because of the interference of the noise in their surroundings. Of course meaning must be conserved in the real world, so the change in the word must be recognised by all speakers in the community - they must still be able to communicate and be understood by each other.

We will continue to use the model so far developed looking at frequency and rank, with some slight modifications to add a length to each word. In the current model, the initial set, or the environment, contains the numbers 1 to 1000, each of which represents one word. To this a length will be added for each word. So the initial set will consist of 1000 word-length pairs. This means that every time a word is selected to be put into the new word-store, its length is copied over with it. All words start with the same length, arbitrarily determined, at the beginning of a run.

To simulate the effect of shortening another variable was created which sets the probability of 'noise' interfering and thus the word being shortened by 1 between the source vocabulary and its destination. By this means it should occur that words which are selected more often are shortened more often and so more frequent words will become the shortest ones. (This is the same random shortening mechanism as in Kirby's (2001) model.) Obviously with this model of word shortening it is possible to have the same word with two different lengths in the same vocabulary. This is because a word may be copied over to the new vocabulary once without its length being reduced, but then copied over again and have its length reduced by one. To normalise a difference in length occurring between two instances of the same word, once a vocabulary is complete it is searched for such an event. The length of each word is then changed to match the lowest length present for that word in the vocabulary.

There are two possible procedures for analysing the data obtained from this new model. The first would be to follow the method used on the results gained from looking at the relationship between frequency and rank. This is simply to look at an individual word's frequency and its length, and plot the two against each other. This produces graphs with multiple lengths for the same frequency.

The alternative method, used by Zipf, is to group together all words with the same frequency and then take the average length for words with that frequency. Figure 5, a graph illustrating the data from an investigation of R. C. Eldridge (1911) of the English words used in four American newspapers, demonstrates this method, (Zipf, 1935). There are many fewer examples of words of higher frequencies, therefore any difference in the lengths of words of these frequencies has a much greater impact on the average length. For this reason Zipf averaged length over sets of higher frequencies, rather than simply taking the average length of words with the same frequency. This paper is attempting to recreate the results

Zipf obtained from his study of natural language. Therefore the second method described will be used to illustrate the results of the new model.

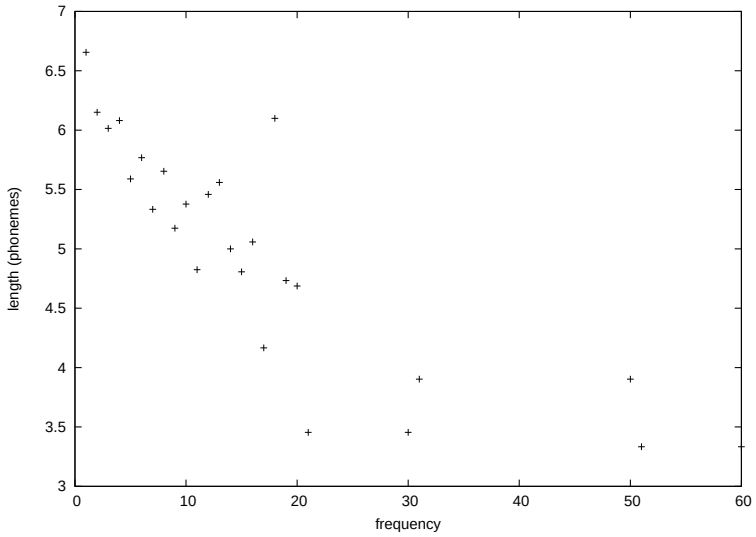


Figure 5. Eldridge's survey of frequency-length relationships of words in American newspapers.

2.2.1 Length/Frequency Results with Basic Settings

The first runs of the new model will use 'basic' settings: no environmental influence and a corpus size of 1000. This will give a basic shape for the graph of length against frequency. Parameters will then be altered to include environment and Kdifferent corpus sizes. We hypothesize that the relationship between frequency and length will require much the same conditions as were needed to gain a Zipf-like rank/frequency correlation in the previous section. This is due to the unavoidable link between these two relationships; if the frequency distribution is not like that of a natural language then the reported frequency-length relationship may not occur.

In this situation the rank-frequency distribution is much like it was before in the first few runs of the very first model. This meant that at the end of a run of 100 generations there were very few words left in the language (somewhere between 15 and 25 usually). This obviously has an effect on the frequency-length distribution. As so many words have been lost from the vocabulary it is very difficult to find a pattern on a graph of frequency versus length with so few points.

2.2.2 Adding Environmental Influence Again

To combat this loss of diversity the environment parameter was adjusted to try and stabilise the size of the vocabulary. This prevents words being lost so quickly and may allow the pattern of word shortening

to emerge as a result. The environment parameter was set to a relatively low level at 0.05. This was to ensure that there would be some stabilisation of the vocabulary size, whilst still allowing the vocabulary to develop to the Zipf-like shape.

The results graphed below show that the inverse relationship between frequency and length is beginning to emerge, but still not to the extent reported by Zipf. It may be that the vocabulary size, although somewhat stabilised, is not large enough for the length-frequency relationship to appear. As noise is increased to 0.5 the inverse shape of the graph begins to show through as the range of lengths at the 100th generation grows, as shown in figures 6.

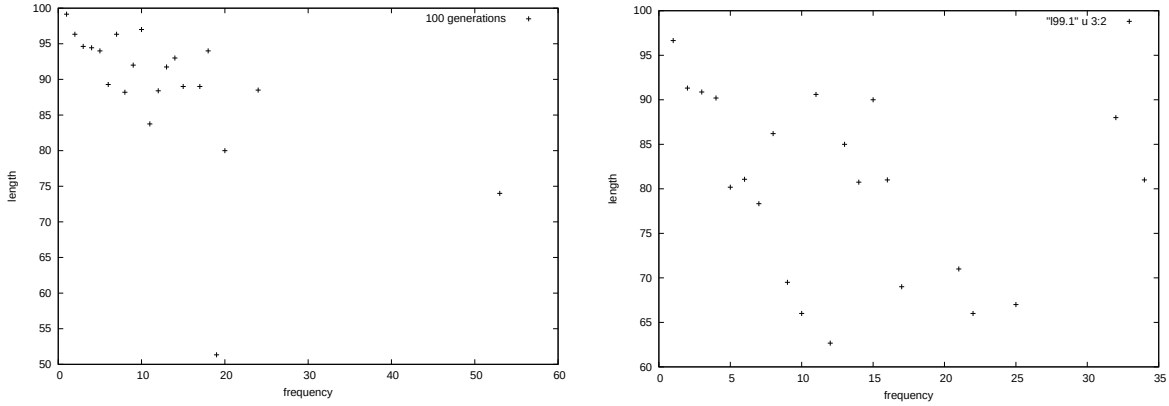


Figure 6: Frequency-length distribution after 100 generations, with environmental influence set at 0.05: **Left panel:** Noise = 0.1. **Right panel:** Noise = 0.5.

2.2.3 Varying Corpus Size Again

Corpus size is the parameter which had the greatest effect on the rank-frequency distribution. Therefore once this relationship is as Zipf found it to be, the inverse relationship between frequency and length should emerge. Runs were completed with 5000 and 10,000 selections to ascertain what effect increasing numbers of selections would have on the frequency-length distribution.

From the graphs below it can be seen that increasing the corpus size has produced a distribution closer to that Zipf obtained. Figure 7 shows the frequency-length relationship after 100 generations for corpus sizes of 5000 and 10,000. Less frequent words can be seen to have a greater length, although more frequent words tend to have a more variable length.

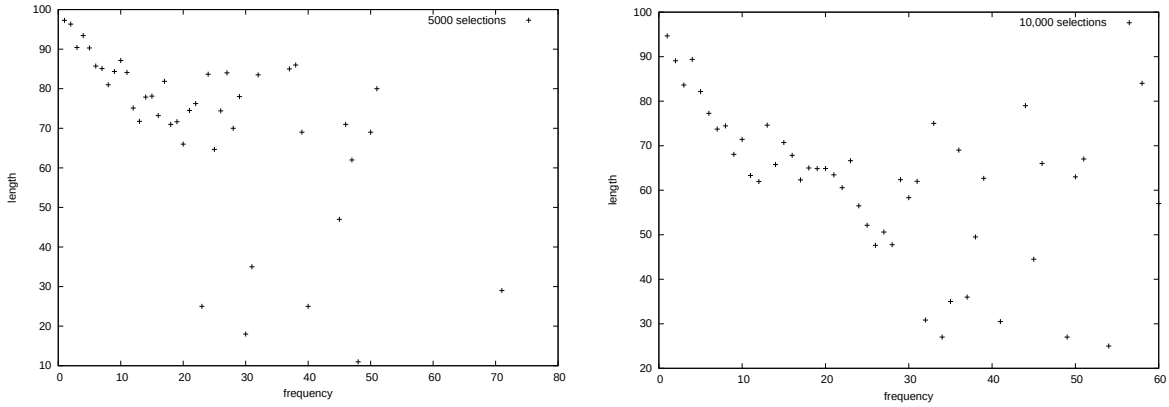


Figure 7: Frequency-length distribution after 100 generations, $E = 0.05$, noise = 0.1.
Left panel: Corpus size = 5000. **Right panel:** Corpus size = 10,000.

Eyeballing these results, the pattern in the right-hand panel is quite similar to Eldridges's data presented in figure 5.

3 Conclusion

Our simulations model the following factors in the transmission of vocabularies across generations:

- Storage of information on frequency of words heard,
- A major influence of stored word-frequency on production,
- A mild influence of non-discourse-related factors ('environment') on word-choice,
- Large corpus size relative to vocabulary size (token/type ratio),
- Noise affecting random shrtening of words.

Our results show interesting similarities with the rank/frequency and length/frequency distributions described by Zipf.

4 Bibliography

Eldridge, R. C. (1911) *Six Thousand Common English Words*, Buffalo: The Clement Press.

Harremoës, P. and Topsøe, F. (2001) "Maximum Entropy Fundamentals", *Entropy*, 3:191-226.

Hurford, James R. (2002) "Expression/induction models of language evolution: dimensions and issues".

In *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, edited by Ted Briscoe, Cambridge University Press. pp.301-344.

Kirby, Simon, (2001) “Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity”, *IEEE Transactions on Evolutionary Computation*, 5(2):102-110.

Kirby, Simon and Hurford, James R (2001) “The Emergence of Linguistic Structure: an Overview of the Iterated Learning Model”, in Parisi, Domenico and Cangelosi, Angelo, Eds. *Computational Approaches to the Evolution of Language and Communication.* Springer Verlag, Berlin.

Kučera, H., and W. Nelson Francis (1967) *Computational Analysis of Present-Day American English*, Brown University Press, Providence, Rhode Island.

Li, W. (1992), “Random texts exhibit Zipf’s-law-like word frequency distribution”, *IEEE Transactions on Information Theory*, 38(6):1842-1845

Miller, George A., (1951) *Language and Communication*, McGraw -Hill, New York.

Miller, George A., (1957) “Some effects of intermittent silence”, *American Journal of Psychology*, 70, pp.311-314.

Miller, George A., (1965) Introduction to republication of Zipf (1935), MIT Press, Cambridge, MA.

Onnis, Luca, Matthew Roberts, and Nick Chater (2002) “Acquisition and evolution of natural languages: Two puzzles for the price of one”, paper given at the Fifth International Conference on the Evolution of Language, Harvard.

Shannon, Claude E., (1948) “A mathematical theory of communication”, *Bell Systems Technical Journal*, 27:379-423, 623-656.

Zipf, G. K., (1935) *The Psycho-Biology of Language*, Houghton Mifflin, Boston.

Zipf, G. K. (1949) *Human Behaviour and The Principle of Least Effort*, Addison-Wesley, Cambridge, MA.

Iterated learning and grounding: from holistic to compositional languages

Paul Vogt
ILK/Computational Linguistics and AI
Tilburg University, Tilburg, The Netherlands.
Language Evolution and Computation Research Unit
University of Edinburgh, UK.
paulv@ling.ed.ac.uk

http://www.ling.ed.ac.uk/~paulv

Abstract

This paper presents a new computational model for studying the origins and evolution of compositional languages grounded through the interaction between agents and their environment. The model is based on previous work on adaptive grounding of lexicons and the iterated learning model. Although the model is still in a developmental phase, the first results show that a compositional language can emerge in which the structure reflects regularities present in the population's environment.

1 Introduction

Evolutionary computational linguistics has become a booming research area during the past decade, see, e.g., [5, 9] for overviews. One particular area that has gained increasing attention is the emergence of compositional languages, i.e. languages in which parts of an expression have a structured relationship to their semantics. Strikingly, but not surprisingly, almost every study that has investigated the emergence of compositional languages have assumed a predefined meaning space, e.g., [1, 8]. Consequently, these studies are subject to the *symbol grounding problem* [7], which relates to the question how symbols become meaningful to the agent who uses them. One of the reasons why we should try to avoid the symbol grounding problem is that a lot of linguistic structures may be induced from an agent's interaction with their environment.

The only known study that investigates the origins of grammatical structures in a grounded setting has been reported by Steels [10]. In this work, Steels proposes a model in which agents construct a procedural grammar of which the semantics are acquired through their interaction with their environment and the grammatical structures through a complex interplay between the semantics and linguistic utterances produced by the agents. The experimental framework on which Steels' model is based is called the *Talking Heads experiment* [11]. In this experiment, a population of agents attempt to develop a language with which they communicate about aspects of their environment. This environment contains geometrical coloured figures that the robotic agents see with their steerable camera heads.

This paper presents a new computational model based on a simulation of the Talking Heads experiment. This model combines the *iterated learning model* as proposed by Kirby [8] with aspects of symbol grounding aimed at researching the emergence of compositional languages. In particular,

the paper attempts to show how learners can induce syntactic and semantic structures by observing the linguistic behaviours of adult speakers and by discovering visual regularities in their environment. Through a process of invention and induction, a language is bootstrapped from scratch and transmitted culturally to subsequent generations.

The following section presents some background on the state-of-the-art in iterated learning and grounding. Section 3 presents the proposed model. Initial results are presented in Section 4, which are discussed in Section 5.

2 Iterated learning and grounding

The iterated learning model (ILM) has been proposed to study aspects of language evolution, in particular the way language is transmitted culturally from one generation to another [4, 8]. The ILM contains a population of adults and learners, where the adults teach their language to the learners through linguistic interactions such as language games. After each iteration, the adults are replaced by the learners and new learners enter the population. Kirby [8] has shown how the ILM could model a transition from initial holistic protolanguage into compositional languages. In this study, the holistic language was constructed from associations between predefined meanings (represented as predicate-argument structures) and unstructured signals. Using a number of heuristics, the agents were able to induce syntactic structures relating to the regularities of predicate-argument structures of the semantic space. By applying a bottleneck on the transmission of the language, Kirby was able to show how compositional languages emerged after a number of generations. One shortcoming of Kirby's simulations is that it was subject to the symbol grounding problem.

In [15], I applied the ILM to study the evolution of lexicons in a simulation of the Talking Heads experiment [11]. In this study, a holistic language emerged based on agents' observations of their environment – thus providing a way of grounding – and on the learners' observations of linguistic behaviours of adults. In these simulations, a shared lexicon was constructed by processing numerous language games in which agents tried to convey the meanings of observed objects. These meanings were adaptively formed using discrimination games. If the agents failed to convey a meaning, new (holistic) word-forms were invented, adopted or associations were weakened. If the agents were successful, used associations were strengthened. In this paper the Talking Heads simulation – implemented in the toolkit *THSim* [16][1] – is used as the starting point for studying the emergence of compositional languages.

The model proposed in this paper is based on finding conceptual spaces. In line with Gärdenfors [6], I will use the term *conceptual space* as a space where concepts (or meanings) can be stored and observations can be conceptualised. A conceptual space is spanned by a number of *quality dimensions*. Each quality dimension relates to some quality (or feature) that can be measured by an agent's sensors. For instance, the qualities Red, Green and Blue are quality dimensions of a conceptual space for colour in artificial agents. For holistic languages, I assume that there is one conceptual space that is spanned by all possible quality dimensions – I call this space the *holistic conceptual space*. In compositional languages, conceptual spaces are of lower dimension and relate to certain qualities such as colour or shape. According to Gärdenfors, a conceptual space can form the semantic representation of linguistic categories [6]. For the current study, I hypothesise that in language, holistic utterances (represented in holistic conceptual spaces) evolved first, and compositional structures emerged from these holistic utterances, cf. [8, 17]. Note that I do not claim that holistic utterances initially referred to holistic

[1] Current version THSim v3.2 can be downloaded from www.ling.ed.ac.uk/~paulv/thsim.html. This version is still based on holistic signalling only, future releases will include the compositional model described in this paper as well.

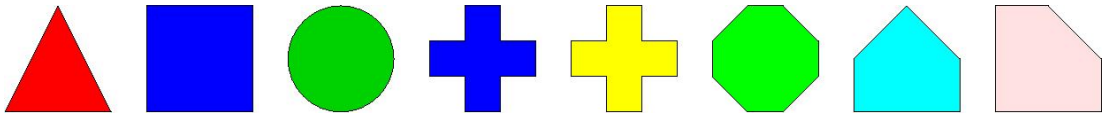


Figure 1: Some of the figures that can occur during a language game.

conceptual spaces, I merely adopt the hypothesis for practical reasons. Given this assumption, to find linguistic categories in a holistic conceptual space, it suffices to discover conceptual spaces of a lower dimension.

Discovering conceptual spaces is guided by a two way process in language development: On one hand, semantic structures are induced from regularities in the interaction between agents and their ecological niche, though constrained by the syntactic structures of their language. On the other hand, syntactic structures are induced from regularities in culturally transmitted linguistic utterances, though constrained by the semantic structures. This principle is based on the findings that language and meanings co-develop [3]. The next section will describe the implemented model in detail. The induction steps in the model are adapted from Kirby's [8] model to integrate symbol grounding.

3 Discovering conceptual spaces

As mentioned, the model is implemented as an extension of the THSim toolkit [16]. In this tool a population of agents can play a series of various language games to develop a language that allows them to communicate about coloured geometrical figures that are displayed on the screen. Whenever the agents fail to communicate successfully, they adapt their ontology (i.e., the set of meanings) and linguistic knowledge (lexicon and grammar) to increase their performance in future games. In the current paper, the population contains only one adult and one learner; the adult takes up the role of speaker, while the learner acts as hearer. After playing a number of language games, the learner replaces the adult and a new learner without any linguistic knowledge enters the population. The period in which the population remains constant is called an *iteration*.

3.1 Sensing the environment

At the start of a language game, a context C of geometrical coloured figures (or objects o_i) is generated by the environment. Each figure is randomly selected from 10 different shapes such as rectangles, circles, triangles, crosses and 6 other regular and irregular polygons (Fig. 1). In addition, each figure is given a colour selected arbitrarily from a set of 12 different colours. So the environment contains a total of 120 objects. In the current presentation, each language game concerns a different context of 8 objects.

The agents 'look' at the context and obtain for each object $o_i \in C$ a feature vector $\mathbf{f}_i$. A feature vector contains a number of features (or qualities) measured by the agent. Currently, the agents use four features, so $\mathbf{f}_i = (f_\mathbf{r}, f_\mathbf{g}, f_\mathbf{b}, f_\mathbf{s})$, where $f_\mathbf{r}, f_\mathbf{g}$ and $f_\mathbf{b}$ relate to the $\mathbf{rgb}$ colour space representation of the object, and $f_\mathbf{s}$ is a *shape feature*. The way these features are calculated is not relevant for this paper (see [16] for a description), it suffices to say that they can be measured by a real robot and that each shape has a distinct shape feature. Unlike the real Talking Heads, the features are measured without any noise. This is done to reduce – for the time being – the complexity of the study. After the agents obtained feature vectors for all objects in the context, the context can be described as $C' = \{\mathbf{f}_1, \ldots, \mathbf{f}_N\}$, where $N = 8$ is the context size.

Once the context is set, the speaker of the language game arbitrarily selects one object from the context as the *topic* of the game and informs the hearer which object this is. This strategy of informing the hearer about the reference of the game is based on establishing joint attention and is called the *observational game* [13]. It differs from the *guessing game*, which was originally played in the Talking Heads experiments [11], where the speaker provides corrective feedback after the hearer guessed what the topic was.

3.2 Meaning formation: the discrimination game

Given the context C' and the topic o_t (described by $\mathbf{f}_t$), the agents try to form a meaning to represent the topic. One way to form meanings is to use the discrimination game model, e.g., [11, 13], which is played by an individual agent. The aim of the discrimination game is to find one or more *semantic hypotheses* for a topic that distinguishes the topic from all other objects in the context. Semantic hypotheses are (compositions of) categories that are defined as regions in a conceptual space, represented by a prototype. A prototype is a point in the conceptual space and its category is that region of which all points are nearest to the prototype.

Definition: A conceptual space is said *to cover* certain quality dimensions. The holistic conceptual space covers all quality dimensions, while a non-holistic conceptual space (or *conceptual space* for short) covers only a subset of all quality dimensions.

The way a (holistic) conceptual space is covered is indicated by feature letters. For instance, in the current study there are 4 quality dimensions **r, g, b** and **s**. The holistic conceptual space covers $\mathbf{rgbs}$, the conceptual space for colour covers $\mathbf{rgb}$ and the 'shape space' covers $\mathbf{s}$. If an agent has a holistic category represented by $\mathbf{c} = (1.0, 0.0, 0.0, 0.1)$, this category can be decomposed into two categories covering $\mathbf{rgb}$ and $\mathbf{s}$, represented by $\mathbf{c}' = (1.0, 0.0, 0.0, ?)$ and $\mathbf{c}'' = (?, ?, ?, 0.1)$, where the $?$s are wild cards. If an object is observed by an agent, it could categorise its feature vector holistically (yielding the category set $\{\mathbf{c}\}$) or compositionally (yielding $\{\mathbf{c}', \mathbf{c}''\}$). All categories $\mathbf{c}_i$ of agent a are stored in its ontology $\mathcal{O}_a = \{\mathbf{c}_1, \ldots, \mathbf{c}_p\}$, which is initially empty.

At the start of a discrimination game, the agent categorises all $\mathbf{f}_i \in C'$ by searching those categories in each different conceptual space for which the feature vector is nearest to the category that covers the conceptual space. From these categories, category sets are constructed such that the compositions cover all dimensions of the holistic space. This yields for each feature vector $\mathbf{f}_i \in C'$ a *set of category sets* $\mathcal{C}_{a,i}$.

If all sets are constructed, the agent removes all category sets $H_n \in \mathcal{C}_{a,t}$ for which for some $i \neq t$: $H_n \in \mathcal{C}_{a,i}$, yielding a semantic hypothesis set $\mathcal{H}_{a,t} = \{H_k\}$ for topic f_t. In other words: the semantic hypothesis set contains those category sets for the topic that are not category sets for any other object in the context, thus distinguishing the topic.

If $\mathcal{H}_{a,t} = \emptyset$, the agent has no hypothesis that allows it to distinguish the topic from the rest of the context, and the discrimination game fails. In this case, the agent will create a new holistic category by taking the topic's feature vector $\mathbf{f}_t$ as an exemplar. If $\mathcal{H}_{a,t} \neq \emptyset$, the discrimination game succeeds and the semantic hypothesis set $\mathcal{H}_{a,t}$ is forwarded to the production or interpretation phase of the language game.

3.3 Production

After the speaker has successfully played a discrimination game, it tries to produce an expression to convey the reference to the topic. Production is done in three stages. First, the speaker searches

Grammar	$m \backslash F$	$bluesquare$	red	$triangle$	rue
$r1 = S \rightarrow bluesquare/\mathbf{rgbs}$	$m1 = (0,0,1,1)$	0.6	0.0	0.0	0.0
$r2 = S \rightarrow A/\mathbf{rgb}\ B/\mathbf{s}$	$m2 = (1,0,0,0)$	0.1	0.0	0.0	0.0
$r3 = A \rightarrow red/\mathbf{rgb}$	$m3 = (1,0,0,?)$	0.0	0.5	0.0	0.2
$r4 = B \rightarrow triangle/\mathbf{s}$	$m4 = (?,?,?,0)$	0.0	0.0	0.7	0.0
$r5 = A \rightarrow rue/\mathbf{rgb}$	$m5 = (1,0,?,?)$	0.0	0.0	0.0	0.0
	$m6 = (?,?,0,0)$	0.0	0.0	0.0	0.0

Table 1: An example grammar and lexicon. The left column presents the grammar. The right part of the table shows the lexicon where the forms F are presented in the columns and the meanings m in the rows. The values in the cells represent the association scores $\sigma_{F,m}$.

grammatical rules that fit the semantic compositions of the semantic hypothesis set. Second, the speaker tries to lexicalise the composed categories that fit a grammatical rule. Third, the speaker selects that lexicalisation that has been most effective in the past.

During the agents' lifetimes, each agent a constructs a private grammar $\mathcal{G}_a = \{r_1, \ldots, r_q\}$ with rewrite rules r_i like the ones presented on the left-hand side of Table 1. In this table, the symbol S is the start symbol of a sentence, other upper case letters, such as A and B, are arbitrarily named terminals, the italic lower case strings are word-forms and the bold face strings indicate the covering of the terminals. One might expect that, rather than indicating which conceptual space is covered by the word-forms (as in rules 1, 3, 4 and 5), one could indicate the forms' meanings. However, a form may be associated with more than one meaning (and vice-versa), as shown in the lexicon at the right hand side of Table 1.

Each agent a additionally has a lexicon $\mathcal{L}_a$, defined as an associative memory that associates forms F_i with meaning m_i mediated by an association score σ_{F_i,m_i}. An association score indicates the effectiveness of an element in previous language games. Lexical elements $l_i \in \mathcal{L}_a$ are notated by $l_i = \langle F_i, m_i, \sigma_{F_i,m_i} \rangle$. Initially, both $\mathcal{L}_a = \emptyset$ and $\mathcal{G}_a = \emptyset$.

When searching rules that match the semantic compositions, the speaker searches for a way to parse each semantic hypothesis with the grammar by matching the covers of the composition. Suppose the speaker has obtained the semantic hypothesis set $\mathcal{H}_{s,t} = \{\{m2\}, \{m4, m3\}, \{m5, m6\}\}$. In this case only the first two sets are parseable with respect to the grammar presented in Table 1. The first set $\{m2\}$ fits a rule like $r1$, because it covers **rgbs**. Likewise, the second set $\{m4, m3\}$ fits rule $r2$ (note that the order of categories is discarded in the semantic hypotheses, the grammatical rules represent the order). The final set $\{m5, m6\}$ does not fit any rule, because the composition covers **rg** and **bs**, which do not combine to form a rule in the grammar.

Given these compositions, the speaker tries to find forms that match the categories of the compositions in the same way as done previously for holistic communication, e.g., [13, 16]. The speaker searches its lexicon for elements of which the meaning matches one of the categories. Continuing our example, composition $\{m2\}$ can be lexicalised with $bluesquare$ and $\{m4, m3\}$ with $triangle$ and red. Thus the speaker has two ways to express the two hypotheses: $bluesquare$ and $redtriangle$, which are derived from compositions $r1$ and $r2 \circ r3 \circ r4$ respectively. Note that the composition $r2 \circ r3$ indicates that that $r3$ is applied to the leftmost free terminal of $r2$. Further note that when an expression is composed of more than one form, the forms are concatenated such that the hearer cannot explicitly detect word-boundaries. Now the speaker will select the expression that was most effectively in the past based on the average association scores of the lexical elements. In the example, the association $\langle m2, bluesquare, 0.1 \rangle$ has an average score of 0.1, while the associations $\langle m3, red, 0.5 \rangle$

and $\langle m4, triangle, 0.7\rangle$ have an average score of 0.6. As the latter is higher, the speaker will express $redtriangle$.

If the speaker fails to produce an utterance, which is the case when it has no grammatical rule to cover the semantics or when it (partially) has no matching association in its lexicon, the speaker expands its grammar and lexicon. There are two possibilities:

1. *The speaker has a rule of more than one constituent that covers the semantics, but there is no matching (or partially matching) association in its lexicon.* In this case the speaker invents one or more new word-forms to associate with the meaning parts for each unassociated category.
2. *The speaker has no rule to cover any of the semantic hypotheses.* In this case the speaker invents a new word-form that is associated with a holistic hypothesis. If no such hypothesis exists, the categories of a compositional hypothesis are merged into a holistic category.[2]

The first case occurs, for example, when the speaker has the semantic hypothesis set $\mathcal{H}_{s,t} = \{\{(1, 0, 0, ?), (?, ?, ?, 1)\}\}$ and the above grammar and lexicon. In that case, it can select rule $r2$, together with rule $r3$ to form a partial expression $red...$. The speaker will then invent a new form, for instance *square*, and adds the association $\langle square, (?, ?, ?, 1), 0.01\rangle$ to its lexicon. In addition, it will add the rule $B \rightarrow square/\mathbf{s}$ to its grammar. (Note that in the simulations forms are invented as sequences of consonant-vowel pairs randomly selected from a finite alphabet.)

The second case occurs, for example, when the speaker has the semantic hypothesis set $\mathcal{H}_{s,t} = \{\{(0, 1, 0, \frac{1}{2})\}\}$. In that case a new form is invented, say $greenpentagon$, and the association $\langle green\ pentagon, (0, 1, 0, \frac{1}{2}), 0.01\rangle$ is added to the lexicon. In addition, the rule $S \rightarrow greenpentagon/\mathbf{rgbs}$ is added to the grammar. If the hypothesis set would have been $\mathcal{H}_{s,t} = \{\{(0, 1, ?, ?), (?, ?, 0, \frac{1}{2})\}\}$, then the two categories are merged into $(0, 1, 0, \frac{1}{2})$ and the above mentioned adaptations are made. Note that the speaker is not able to invent new compositional structures; it can only exploit existing ones.

3.4 Interpretation and induction

Interpretation Upon receiving the expression, the hearer (or learner) tries to interpret the expression. If it fails, the hearer will try to induce new linguistic knowledge. Interpretation is processed in two stages: parsing the expression and checking the semantics. Parsing is done at the syntactic level, i.e. the expression is parsed relative to the grammar while the semantics is ignored. I will not go into the details of the parser, as this is relatively straightforward. The only complication is that the word-boundaries are not visible. For the time being, the parser results only in one possible parse. In practise, however, there may emerge situations where more than one parse could be possible, but such situations are currently disregarded for practical reasons. The parser results in a list of forms that are interpreted, together with the interpreted composition.

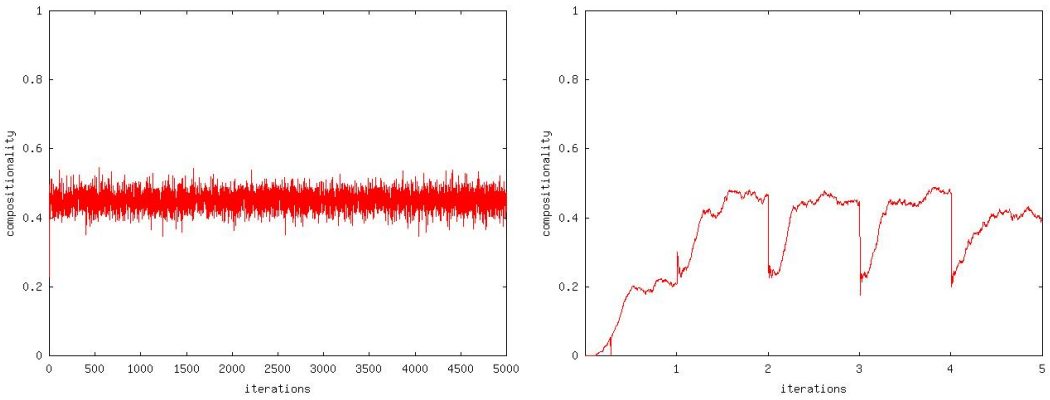
When a parse is found, the resulting list of forms is evaluated relating to the hearer's lexicon and its semantic hypothesis set $\mathcal{H}_{h,t}$ for the topic. So, if the parser returns the result $E = \{e_1, \ldots, e_n\}$, where each e_i is a part of the expression, the hearer searches for each element $e_i \in E$ a lexical element $l_j = \langle e_i, m_j, \sigma_{e_i,m_j}\rangle \in \mathcal{L}_h$ for which the association score $\sigma_{e_i,m_j} > 0$ and $m \in H$, where $H \in \mathcal{H}_{h,t}$ is a hypothesis. A semantic interpretation is complete if the entire expression E can be fully interpreted by a $H \in \mathcal{H}_{h,t}$. If more such interpretations exist, the hearer selects that interpretation H for which the average association score is highest. The language game is successful if the entire expression is completely interpreted by a semantic hypothesis of the topic.

[2] This latter procedure is not ideal, but was implemented to solve impasses occurring when it was not done.

Following our example, suppose the hearer received the expression $redtriangle$. Parsing this expression to the grammar presented above, yields the following expression $E = \{red, triangle\}$ and composition $r2 \circ r3 \circ r4$. If the hearer's hypothesis set $\mathcal{H}_{h,t}$ includes $H = \{m3, m4\}$ or $\{m4, m3\}$, then, given the lexicon of Table 1, this H is the interpretation of $redtriangle$. In this case the language game is a success and the association scores between the used elements are increased by $\sigma = \eta \cdot \sigma + 1 - \eta$, while competing associations are laterally inhibited by $\sigma = \eta \cdot \sigma$. An association is competing if the form matches (part of) the expression but not its meaning or vice versa. Given this scheme, the association scores $\sigma_{red,m3}$ and $\sigma_{red,m4}$ are increased, while $\sigma_{rue,m3}$ is inhibited. The speaker also receives feedback on the outcome and adapts its association scores in a similar way. If the hearer fails to interpret the expression, both agents lower the score of any of the used associations.

Induction When the learner fails to parse the expression syntactically and/or semantically, it will try to induce new linguistic knowledge from the expression with respect to previously learnt knowledge. There are basically three reasons why parsing can fail.

1. *The hearer is able to parse the expression syntactically, but not semantically.* This occurs when the hearer obtained a non-empty expression list E, but failed to find an interpretation. In this case the hearer associates the words of the expression with a $H \in \mathcal{H}_{h,t}$ of equal size and of which the meaning parts cover the conceptual spaces of the terminals of the parsed rule. Going back to our example, suppose the learner received the expression $ruetriangle$, which it can parse using composition $r2 \circ r5 \circ r4$. Further suppose that the only H that covers this composition is $\{(0, 0, 1, ?), (?, ?, ?, \frac{1}{4})\}$. The learner will then add the associations $\langle rue, (0, 0, 1, ?), 0.01\rangle$ and $\langle triangle, (?, ?, ?, \frac{1}{4}), 0.01\rangle$ to its lexicon.
2. *The hearer is able to parse the expression partially on both the syntactic and semantic level.* This happens when the learner finds a composition by which only a part of the expression is interpreted. In this case, the learner associates the not interpreted part of the expression with the remaining elements of the H that is partially interpreted, constrained by the grammar. If there are more ways to interpret the expression partially, the hearer prefers to adapt the remaining part of the expression with the minimum number meaning parts. If there are more than one such partial matches, the one with the highest average association score is selected. For example, if the hearer receives the expression $bluetriangle$ while having a $H = \{(0, 0, 1, ?), (?, ?, ?, 0)\} \in \mathcal{H}_{h,t}$, then, using composition $r2 \circ r? \circ r4$, it is able to match the expression partially with the association $\langle triangle, m4, 0.7\rangle$. The remaining part of the expression $blue$ should then relate to the '$r?$' part of the composition. To achieve this, the association $\langle blue, (0, 0, 1, ?), 0.01\rangle$ is added to the lexicon and the new rule $A \rightarrow blue/\mathbf{rgb}$ is added to the grammar, which then relates to $r?$.
3. *The hearer cannot parse the expression at all.* This occurs when there are no rules in the grammar that match the expression. In this case, the learner tries to split the expression such that it partially matches a split in an existing rule, both syntactically and semantically. Although in principle splits could be applied to every type of rule, they are currently only applied to holistic rules. For example, suppose the learner receives $yellowsquare$ that could relate to $H = \{(1, 1, 0, 1)\}$. A split can then be made in rule $r1$ with the shared form $square$. Additionally, a split can be made in the semantics (if this is not the case, the split is not pursued further), yielding a shared category $(?, ?, ?, 1)$. Rule $r1$ is now rewritten as $S \rightarrow A/\mathbf{rgb}\ square/\mathbf{s}$, and the rules $A \rightarrow blue/\mathbf{rgb}$ and $A \rightarrow yellow/\mathbf{rgb}$ are added to the grammar. In addition the element $\langle bluesquare, m1, 0.6\rangle$ is replaced by the elements $\langle blue, (0, 0, 1, ?), 0.6\rangle$ and $\langle square, (?, ?, ?, 1), 0.6\rangle$. Also the association $\langle (1, 1, 0, ?), yellow, 0.01\rangle$ is added to the lexi-



(a) (b)

Figure 2: (a) The compositionality during the final 50 games of each iteration. (b) How compositionality evolved during the first 5 iterations.

con.

If no split can be made, the expression is added holistically. This would occur, for example, when the hearer received $greencircle$ or $yellowsquare$ with $H = \{(1, 1, 0, 0)\}$. In this case, a rule with start node S and cover $\mathbf{rgbs}$ is added to the grammar, and the association of the form with a category covering $\mathbf{rgbs}$ is added. If no such category exists, a H composed of two (or more) categories is merged such that the resulting category is holistic again.

3.5 Generalise and merge

When the speaker or hearer has changed a rule, the agent will make sure the grammar contains no redundancies by *generalising* and/or *merging* rules. If two non-holistic rules contain constituents relating to the same linguistic category, these rules can be generalised. For example, if the grammar contains the rules $S \rightarrow A/\mathbf{rgb}\ triangle/\mathbf{s}$ and $S \rightarrow A/\mathbf{rgb}\ circle/\mathbf{s}$, then both rules are removed from the grammar and replaced by the rules $S \rightarrow A/\mathbf{rgb}\ X/\mathbf{s}$, $X \rightarrow triangle/\mathbf{s}$ and $X \rightarrow circle/\mathbf{s}$, where the terminal X can be any yet unused upper case letter.

If there are two rules that have constituents with different terminals that are acting on the same conceptual space, these rules are merged. For instance, rules $S \rightarrow A/\mathbf{rgb}\ B/\mathbf{s}$ and $S \rightarrow C/\mathbf{rgb}\ B/\mathbf{s}$ will be merged into $S \rightarrow A/\mathbf{rgb}\ B/\mathbf{s}$ and all terminals C are replaced by A throughout the grammar.

4 Results

This section presents the results of a representative simulation. In this simulation, the population contained one adult/speaker and one learner/hearer. The simulation was run for 5000 iterations of 350 language games each. Splitting of utterances was only processed on holistic signals with the consequence that only compositions of two constituents could emerge. The simulation was repeated 10 times with different random seeds.

Figure 2 shows the averaged results of the 10 simulation runs. Graph (a) shows the *compositionality* at the end of each iteration. Compositionality is the average number of compositional expressions

R	iteration 1	iteration 2500	iteration 4999
A	$S \rightarrow x/\mathbf{rgbs}(279)$	$S \rightarrow x/\mathbf{rgbs}\ (110)$ $S \rightarrow A/\mathbf{s}\ B/\mathbf{rgb}(170)$ $S \rightarrow D/\mathbf{gbs}\ gi/\mathbf{r}(48)$ $S \rightarrow B/\mathbf{rgb}\ bi/\mathbf{s}(0)$	$S \rightarrow x/\mathbf{rgbs}(75)$ $S \rightarrow A/\mathbf{s}\ B/\mathbf{rgb}\ (258)$ $S \rightarrow C/\mathbf{s}\ D/\mathbf{rgb}(0)$ $S \rightarrow de/\mathbf{r}\ A/\mathbf{gbs}\ (1)$ $S \rightarrow D/\mathbf{rgb}\ hi/\mathbf{s}(0)$
L	$S \rightarrow x/\mathbf{rgbs}(21)$ $S \rightarrow A/\mathbf{r}\ B/\mathbf{gbs}(5)$ $S \rightarrow C/\mathbf{s}\ D/\mathbf{rgb}(0)$ $S \rightarrow E/\mathbf{rgs}\ ica/\mathbf{b}(0)$ $S \rightarrow F/\mathbf{rbs}\ da/\mathbf{g}(1)$	$S \rightarrow x/\mathbf{rgbs}(1)$ $S \rightarrow A/\mathbf{s}\ B/\mathbf{rgb}\ (45)$ $S \rightarrow C/\mathbf{gbs}\ D/\mathbf{r}\ (1)$	$S \rightarrow x/\mathbf{rgbs}(7)$ $S \rightarrow A/\mathbf{gbs}\ B/\mathbf{r}(3)$ $S \rightarrow C/\mathbf{s}\ D/\mathbf{rgb}(0)$ $S \rightarrow de/\mathbf{r}\ A/\mathbf{gbs}(1)$
A	$S \rightarrow x/\mathbf{rgbs}(156)$ $S \rightarrow A/\mathbf{r}\ B/\mathbf{gbs}(39)$ $S \rightarrow C/\mathbf{s}\ D/\mathbf{rgb}(66)$ $S \rightarrow E/\mathbf{rgs}\ ica/\mathbf{b}(8)$ $S \rightarrow F/\mathbf{rbs}\ da/\mathbf{g}(21)$	$S \rightarrow x/\mathbf{rgbs}(140)$ $S \rightarrow A/\mathbf{s}\ B/\mathbf{rgb}(132)$ $S \rightarrow C/\mathbf{gbs}\ D/\mathbf{r}(44)$	$S \rightarrow x/\mathbf{rgbs}(106)$ $S \rightarrow A/\mathbf{gbs}\ B/\mathbf{r}(32)$ $S \rightarrow C/\mathbf{s}\ D/\mathbf{rgb}(183)$ $S \rightarrow de/\mathbf{r}\ D/\mathbf{gbs}(0)$

Table 2: The grammar of some adults (A) and learners (L) that emerged during one simulation run. Note that the symbol x in the holistic rules is a variable that is filled with different forms. See the text for details.

produced or interpreted by the agents during the previous 50 language games. As the figure shows, the compositionality is already established at approximately 50% from the second iteration onward. Figure 2 (b) shows how the compositionality evolves within the first five iterations. It increases rapidly toward a value near 50% after which it stabilises.

Table 2 shows parts of the private grammars that emerged during one of the simulation runs in iterations 1, 2500 and 4999 of the adult (1st row) and learner (2nd row), and in iterations 2, 2501 and 5000 of the adult (bottom row). The numbers behind the brackets indicate how many times the rules were produced or interpreted during one iteration of 350 language games. At the end of their lifetimes, the agents have around 100 rules, of which about 40 are holistic rules. The table shows all rules that have more than one constituent. Most interesting are the grammars of the adults from iteration ≥ 2. There you see that most of the produced expressions are compositional rather than holistic, although the holistic expressions still make up a great deal of the utterances.

It is also interesting to see that of the compositions made, those that cover the composition **rgb** and **s** (representing the conceptual spaces for colour and shape) occur most frequently. These reflect the regularities that can be found in the world, where there are a given number of objects (shapes) that are combined with a given number of colours.

Although the table indicates a rather stable grammar (the grammars look very similar), further analysis revealed that the word order flips very frequently (in about 40% of the iterations for rules covering the composition **rgb** and **s**). In addition, this combination is not always the most dominant composition. It competes strongly with other compositions, most notably those covering **r** and **gbs**, which reflect regularities of the **rgb** colours used. The colour/shape compositions are inexistent in less than 1% of the iterations, while they are most dominant in about 65% of the iterations.

5 Discussion and conclusion

This paper presents a new computational model to study the origins and evolution of compositional languages of which the semantics are grounded in the population's interaction with their world. The

model combines previous work on the emergence of syntax [8] and lexicon grounding [11, 13] with the idea of discovering conceptual spaces [6].

The first results do not show the expected transition from holistic protolanguages to compositional languages, as was obtained by Brighton and Kirby [4, 8]. Their results were obtained by imposing a bottleneck on the transmission of the language from one generation to the next, similar to the poverty of the stimulus. In the current simulation no such bottleneck was imposed (more language games were played than there are objects), because when it was used, compositional languages emerged very infrequently within iterations and died away immediately after. However, looking at the number of meanings that were formed, one could argue that there was a bottleneck, because there emerged roughly an average of 105 meanings, including those covering non-holistic conceptual spaces. Nevertheless, no sudden transition toward compositional languages, after a large number of iterations with holistic languages was observed.

Whether this finding is fundamental or not remains to be seen. It might be caused by a wrong parameter setting, such as the size of the alphabet. This size controls the probability of finding regularities in the initially unstructured expressions, which guides the formation of compositional rules in the splitting part of the inducer. In the simulation presented, the alphabet contained only 6 consonants and 3 vowels, which may be rather small. Future work should investigate the effect of this parameter more carefully.

Another aspect that requires more attention is the selection of rules during production and interpretation. At the moment, the parser stops when it has found a possible parse, which is the first parse that occurs in its grammar. However, it is well possible that an agent has more ways to parse a sentence or semantic hypothesis. If the agent can select a rule from different possible parses, for instance, based on the effectiveness of the rule in previous language games, the transmission of the language may become more stable. Candidates for such selection-based learning algorithms are, e.g., *data-oriented parsing* [2] and *alignment-based learning* [12].

In addition, the discrimination game is likely to be a source for the instability of the emerging language. As the development of the agents is asynchronous, they have a different trajectory of constructing meanings. Furthermore, a discrimination game can succeed even if the semantic hypothesis are no direct representations of the topic's feature vector. This is because the semantic hypotheses are formed from categories that are *nearest* to the topic's feature vector and that distinguish the topic from other objects in the context. Hence, because the context is of limited size and does not include all objects in the world, the categories need not have a prototype that is in a (very) close proximity of the feature vector. An alternative method would be what I have called the *identification game* [14], where a feature vector is categorised with the nearest prototype that is within a certain distance. If no such category exists, a new one is constructed by taking the feature vector as an exemplar. If the threshold distance is sufficiently low, the emerging ontology would correspond more closely to the world. In the current study, the threshold could be set to a value asymptotically approaching 0 and the ontology would resemble the observed objects exactly, because the sensing is not subject to noise. However, this would not be interesting as it makes the grounding trivial, which in reality is not the case.

Nevertheless, the results show that a compositional language which reflects the structure of the world can emerge. The rules most frequently used were composed from colour and shape conceptual spaces, whereas the second most frequently used rules were composed from conceptual spaces that contained a major regularity of the colour space (the **r** component) and the remaining quality dimensions. The mentioned improvement on rule selection could help to favour the most regular aspects of the world. In addition, an incremental statistical analysis of the holistic conceptual space and how it is used could be used to decide how new conceptual spaces should be constructed more reliably.

To conclude, the proposed model for evolving compositional languages grounded in the populations' ecological niche is a promising model to further investigate this problem, although more research is required to improve the model. We are still far from understanding how human language evolved and models such as the one presented here can help to increase our understanding of language evolution. One aspect this model has shown is that languages may be shaped, at least to some extent, by the way in which language users interact with the world.

Acknowledgements

The work in this paper was done as part of a Visiting Research Fellowship at the Language Evolution and Computation unit of the University of Edinburgh, sponsored by the Royal Society of Edinburgh and the Caledonian Science Foundation. The author is very grateful to the members of LEC for their hospitality and support. Special thanks to Henry Brighton and Simon Kirby for their comments on earlier versions of this paper.

References

[1] J. Batali. Computational simulations of the emergence of grammar. In J. R. Hurford, M. Studdert-Kennedy, and C. Knight, editors, *Approaches to the Evolution of Language*, Cambridge, UK, 1998. Cambridge University Press.

[2] R. Bod. *Beyond grammar – An experience-based theory of language*. CSLI Publications, Stanford, CA, 1998.

[3] M. Bowerman and S. C. Levinson, editors. *Language acquisition and conceptual development*. Cambrige University Press, Cambridge, 2001.

[4] H. Brighton. Compositional syntax from cultural transmission. *Artificial Life*, 8(1):25–54, 2002.

[5] A. Cangelosi and D. Parisi, editors. *Simulating the Evolution of Language*. Springer, London, 2002.

[6] P. Gärdenfors. *Conceptual Spaces*. Bradford Books, MIT Press, 2000.

[7] S. Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.

[8] S. Kirby. Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2):102–110, 2001.

[9] S. Kirby. Natural language from artificial life. *Artificial Life*, 8(3), 2002.

[10] L. Steels. The emergence of grammar in communicating autonomous robotic agents. In W. Horn, editor, *Proceedings of ECAI-2000*, Amsterdam, 2000. IOS Press.

[11] L. Steels, F. Kaplan, A. McIntyre, and J. Van Looveren. Crucial factors in the origins of word-meaning. In A. Wray, editor, *The Transition to Language*, Oxford, UK, 2002. Oxford University Press.

[12] M. van Zaanen. ABL: Alignment-based learning. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, 2000.

[13] P. Vogt. Bootstrapping grounded symbols by minimal autonomous robots. *Evolution of Communication*, 4(1):89–118, 2000.

[14] P. Vogt. Grounding language about actions: Mobile robots playing follow me games. In Meyer, Bertholz, Floreano, Roitblat, and Wilson, editors, *SAB2000 Proceedings Supplement Book*, Honolulu, 2000. International Society for Adaptive Behavior.

[15] P. Vogt. Grounded lexicon formation without explicit meaning transfer: who's talking to who? In *Proceedings of ECAL*. Springer-Verlag, 2003.

[16] P. Vogt. THSim v3.2: The Talking Heads simulation tool. In *Proceedings of ECAL 2003*. Springer-Verlag, 2003.

[17] A. Wray. Protolanguage as a holistic system for social interaction. *Language and Communication*, 18:47–67, 1998.

Grounding As Learning

Gregory M. Kobele, Jason Riggle, Travis Collier, Yoosook Lee,
Ying Lin, Yuan Yao, Charles Taylor, Edward P. Stabler
University of California, Los Angeles

http://taylor0.biology.ucla.edu/al/

1 Grounding

Communication among agents requires (among many other things) that each agent be able to identify the semantic values of the generators of the language. This is the ”grounding” problem: how do agents with different cognitive and perceptual experiences successfully converge on common (or at least sufficiently similar) meanings for the language? There are many linguistic studies of how human learners do this, and also studies of how this could be achieved in robotic contexts (e.g., (Steels, 1996; Kirby, 1999)). These studies provide insight, but few of them characterize the problem precisely. In what range of environments can which range of languages be properly grounded by distributed agents? This paper takes a first step toward bringing the tools of formal language theory to bear on this problem. In the first place, these tools easily reveal a number of grounding problems which are simply unsolvable with reasonable assumptions about the evidence available, and some problems that can be solved. In the second place, these tools provide a framework for exploring more sophisticated grounding strategies (Stabler et al., 2003). We explore here some preliminary ideas about how hypotheses about syntactic structure can interact with hypotheses about grounding in a fruitful way to provide a new perspective on the emergence of recursion in language. Simpler grounding methods look for some kind of correlation between the mere occurrence of particular basic generators and semantic elements, but richer hypotheses about relations among the generators themselves can provide valuable additional constraints on the problem.

2 Learning Grounding

A first useful perspective on learning can be gained from the “identification in the limit” paradigm (Gold, 1967), a framework that is useful for identifying learning problems that are solvable (perfectly) when one makes very generous assumptions about the data potentially available to the learner. In this framework, the learner is successively presented with positive examples of a language, making a (possibly new) hypothesis after each example. Each possible order of presentation of every sentence of the language (repetitions allowed) is called a text (for that language). (Formally, a text is an infinite sequence $t \in L^{\infty}$ such that for every $s \in L$, there is some i such that $t_i = s$.) The learner learns the language if on each text there is a point after which the learner’s hypothesis never changes, and the hypothesis is correct.

We capitalize on the insight in (Siskind, 1996) that keeping track of the cooccurrance of morphemes and meaning atoms (sememes) can allow us to learn the morpheme-sememe association by a process of elimination. Extending this approach to morpheme to morpheme cooccurrence will allow the learner to extract coherent hypotheses from incomplete information and will allow us a foothold into bootstrapping syntax.

We take a language to be a set of sentence-meaning pairs. For our purposes, a sentence is a finite sequence of morphemes (i.e. $s \in \Sigma^*$), and a meaning is a multi-set of sememes ($m \subseteq M^{\mathbb{N}}$). We say a meaning map $\mu^* : \Sigma^* \rightarrow M^{\mathbb{N}}$ is compositional iff there is a map $\mu : \Sigma \rightarrow M^{\mathbb{N}}$ such that for $w \in \Sigma$, $\mu^*(w) = \mu(w)$, and for $s \in \Sigma^*$, $\mu^*(s) = \biguplus_{1 \leq i \leq |s|} \mu(s_i)$ (i.e. μ^* is the homomorphic extension of some μ). If such a map exists it is unique, and we will identify μ and μ^* when no confusion will arise. For $S \subseteq \Sigma^*$, maps μ^*, ν^* are S-equivalent ($\mu \approx_S \nu$) iff for all $s \in S, \mu^*(s) = \nu^*(s)$. We define $\mathcal{L}_S$ to be the set of all pairs $\langle S, \mu \rangle$, where μ is a compositional meaning map. We write $\mu \in \mathcal{L}_S$ for $\langle S, \mu \rangle \in \mathcal{L}_S$. Note that there might be many S-equivalent maps in $\mathcal{L}_S$. We say that $\mathcal{L}_S$ is exactly identifiable in the limit iff there is an algorithm A such that for any $\mu \in \mathcal{L}_S$, A converges to μ on any text from $L = \{\langle s, \mu^*(s) \rangle | s \in S\}$.

A straightforward adaptation of Siskind's cross-situational grounding algorithm to our setting is as follows. To each morpheme w in our lexicon is associated a set $P(w)$ (the *possible* meanings of w). Upon hearing a sentence meaning pair $\langle s, m_s \rangle$, for each morpheme $w = s_i$, if w is not already in the lexicon, it is added and its possible meanings are bounded only by m_s. Otherwise, if w is already in the lexicon, then its possible meanings are reduced to those which occur also in m_s.

Algorithm 1 On input $\langle s, m_s \rangle$

for each $w \in s$

if $w \in Lex$

$P(w) \leftarrow P(w) \cap m_s$

else

$P(w) \leftarrow m_s$

The first question we investigate is this: *if the criterion of learning is exact identification of a particular word to meaning mapping, under which circumstances is successful learning possible?*

Algorithm 1 works by eliminating a sememe from the possible meaning of a morpheme whenever a datum is presented that contains the morpheme without the sememe. This allows for a simple characterization of the language classes it identifies: they are those in which the meaning of each morpheme is exactly the set of sememes that constantly cooccur with the morpheme in the text.

Theorem 1 Algorithm 1 exactly identifies a class of languages $\mathcal{L}_S$ iff for all $\mu \in \mathcal{L}_S$, every $w \in \Sigma$ is such that $\mu(w) = \bigcap \{\mu^*(s) | s \in S \wedge w \in s\}$.

Proof: The **only if** direction follows immediately from the definition of the algorithm above. For the **if** direction, assume that for each w, $\mu(w) = \bigcap \{\mu^*(s) | s \in S \wedge w \in s\}$. Then as there are at most finitely many elements in any $\mu^*(s)$, there is a finite subset

$S_w \subseteq \{\mu^*(s) | s \in S \wedge w \in s\}$ such that $\bigcap S_w = \mu(w)$, and thus a finite point after which all elements of S_w have appeared in the text. As there are only finitely many w, there is a point in the text after which all elements of every S_w have been seen, and thus at this point Algorithm 1 will have converged on μ. □

Now that we have an exact characterization of the languages learnable by this algorithm, we can ask what kinds of languages these are. The next theorem provides a necessary syntactic condition on languages learnable by Algorithm 1; no morpheme may constantly co-occur with another. That is, there could be no morpheme *-ing* whose presence in a sentence entails the presence of another morpheme *be*.

Theorem 2 If every $\mu \in \mathcal{L}_S$ is such that for each $w \in \Sigma$, $\mu(w) = \bigcap\{\mu^*(s) | s \in S \wedge w \in s\}$, then for all w, $\bigcap_{s \in S}\{w' \in s | w \in s\} \subseteq \{w\}$.

Proof: Let $\mu \in \mathcal{L}_S$ be as in the statement of the theorem. Toward a contradiction, let w be such that $\bigcap_{s \in S}\{w' \in s | w \in s\} \supset \{w\}$. Then there is some z such that for any $s \in S$, $w \in s$ entails $z \in s$. Then by assumption, $\mu(w) \supseteq \mu(w) \uplus \mu(z)$, and so $\mu(z) = \emptyset$. However, as μ was arbitrary, we have shown that all $\nu \in \mathcal{L}_S$ map z to the empty set, which is a contradiction. □

Example 1 The following set of sentences gives rise to a set $\mathcal{L}_S$ of languages which are not exactly learnable using Algorithm 1 (by Theorem 2, as $\bigcap_{s \in S}\{w' \in s | w_2 \in s\} = \{w_1, w_2\}$):

$$S = \{w_1w_2, w_1w_3, w_1w_4, w_3w_4\}$$

Note that no two meaning maps in $\mathcal{L}_S$ are S-equivalent - because each of w_1, w_3 and w_4 occurs with the other, S-equivalent maps will agree on their meanings. But then there is no choice for the meaning of w_2.

Theorem 1 gives a precise characterization of the classes of languages which are exactly identifiable by Algorithm 1. However, Example 1 exhibited a simple, and not obviously unreasonable language which was unlearnable by Algorithm 1. Because Algorithm 1 does not keep track of when certain morphemes cooccur, it cannot use the successful resolution of the meaning of one morpheme to assist in resolving the meaning of another. We present below an algorithm which does exactly this. This will allow us to exactly identify any class $\mathcal{L}_S$ of languages with the property that no two meaning maps are S-equivalent.

A (partial) hypothesis is a (partial) function $h : \Sigma \rightarrow M^{\mathbb{N}}$. Given partial functions h, g, they are consistent (hRg) iff whenever both are defined, they agree (for all w, if $\downarrow h(w)$ and $\downarrow g(w)$, then $h(w) = g(w)$). We define a partial operation $\vee$ ('join') over partial functions such that $h \vee g$ is defined just in case hRg and, if defined, is their set theoretic union.

Given a multi-set M, $\Pi(M)$ is the set of partitions of M.

Algorithm 2 first constructs the set of partial hypotheses (defined only on morphemes present in the current sentence) which are consistent with the presented datum. Then each hypothesis already in the lexicon[1]is successively paired with each hypothesis in the newly constructed set. If this pairing of hypotheses is consistent, then their join is added to the lexicon.

[1]We are using the word 'lexicon' here to denote our set of working hypotheses. Once the learner converges on a language the lexicon will contain all and only S-equivalent hypotheses (Theorem 3). Of course, if no two maps are S-equivalent, then the learning will be exact.

Algorithm 2 On input $\langle s, m_s \rangle$

$T \leftarrow \emptyset$

$H \leftarrow \bigcup_{\pi \in \Pi(m_s)} \{h : \{s_1, \ldots, s_{|s|}\} \rightarrow \pi | m_s = h^*(s)\}$

if $Lex = \emptyset$

$Lex \leftarrow H$

else

for each $h \in Lex$

for each $g \in H$

if hRg

$T \leftarrow \{h \vee g\} \cup T$

$Lex \leftarrow T$

The following examples illustrate the behaviour of Algorithm 2.

Example 2 Imagine that the first piece of data a learner saw was $\langle w_1 w_1 w_2, \{0, 0, 2, 2\} \rangle$. Then every partial hypothesis which is consistent with this datum (there are exactly four) is in $H = \bigcup_{\pi \in \Pi(m_s)} \{h : \{s_1, \ldots, s_{|s|}\} \rightarrow \pi | m_s = \biguplus_{1 \leq i \leq |s|} h(s_i)\} = \{h_1, h_2, h_3, h_4\}$, where

$$\begin{array}{rcccl} h_1 & : & w_1 & \rightarrow & \{0, 2\} \\ & & w_2 & \rightarrow & \emptyset \end{array}$$

$$\begin{array}{rcccl} h_2 & : & w_1 & \rightarrow & \{0\} \\ & & w_2 & \rightarrow & \{2, 2\} \end{array}$$

$$\begin{array}{rcccl} h_3 & : & w_1 & \rightarrow & \{2\} \\ & & w_2 & \rightarrow & \{0, 0\} \end{array}$$

$$\begin{array}{rcccl} h_4 & : & w_1 & \rightarrow & \emptyset \\ & & w_2 & \rightarrow & \{0, 0, 2, 2\} \end{array}$$

As this is the first datum presented to the learner, $Lex = H = \{h_1, h_2, h_3, h_4\}$.

Example 3 Continuing from Example 2, imagine that the next datum presented to our learner was $\langle w_2 w_2 w_3, \{0, 0, 0, 0, 1\} \rangle$. Computing H, we find the only consistent maps to be h_5 and h_6:

$$\begin{array}{rcccl} h_5 & : & w_2 & \rightarrow & \{0, 0\} \\ & & w_3 & \rightarrow & \{1\} \end{array}$$

$$\begin{array}{rcccl} h_6 & : & w_2 & \rightarrow & \emptyset \\ & & w_3 & \rightarrow & \{0, 0, 0, 0, 1\} \end{array}$$

Now, as $Lex \neq \emptyset$, we proceed into the **for each** loop in Algorithm 2.

We begin by evaluating h_1 and h_5. As $\neg(h_1 R h_5)$ (because $h_1(w_2) = \emptyset \neq \{0,0\} = h_5(w_2)$), we go on to the next map, h_6. Since $h_1 R h_6$, we set $T = \{(h_1 \vee h_6)\}$, where

$$\begin{array}{rcccl} (h_1 \vee h_6) & : & w_1 & \rightarrow & \{0,2\} \\ & & w_2 & \rightarrow & \emptyset \\ & & w_3 & \rightarrow & \{0,0,0,0,1\} \end{array}$$

Continuing on, we find that the only other pair of maps to bear R to one other are h_3 and h_5. We set $T = \{(h_3 \vee h_5), (h_1 \vee h_6)\}$, where

$$\begin{array}{rcccl} (h_3 \vee h_5) & : & w_1 & \rightarrow & \{2\} \\ & & w_2 & \rightarrow & \{0,0\} \\ & & w_3 & \rightarrow & \{1\} \end{array}$$

Exiting the **for each** loops, we set $Lex = T = \{(h_3 \vee h_5), (h_1 \vee h_6)\}$.

Note that the size of Lex decreases monotonically throughout the course of grounding. Once Lex is initialized (upon seeing the first datum), each successive iteration reduces the number of distinct hypotheses stored. Thus the main computational cost of Algorithm 2 lies in computing H.

We are now ready to prove the main result of this paper:

Theorem 3 For any $S \subseteq \Sigma^*$, $\mathcal{L}_S$ is identifiable in the limit.

Proof: Let $S \subseteq \Sigma^*$, and $\mu \in \mathcal{L}_S$ be arbitrary. Let t be a text for $L = \{\langle s, \mu^*(s)\rangle | s \in S\}$. We show that at some point in the text t_n, every $h \in Lex_n$ is such that $h \approx_S \mu$, and that at every subsequent point, $Lex_n = Lex_{n+i}$.

Note that for every datum $\langle s, \mu^*(s)\rangle \in L$, there is some hypothesis $h \in H$ such that $\mu R h$. Note also that if $\mu R h$ and $\mu R g$, then $h R g$. This is due to the fact that the domain of μ includes the domains of h and of g. From this we conclude that at every step there is a hypothesis $h \in Lex$ such that $\mu R h$.

Now, after seeing a datum $\langle s, \mu^*(s)\rangle$, there will be a hypothesis $h \in Lex$ such that for every $1 \leq i \leq |s|$, $h(s_i) = \mu(s_i)$. Thus, after seeing a finite number of sentences (at most one for each $w \in \Sigma$)[2], $\mu \in Lex$. Similar reasoning tells us that at that point $\{\nu | \mu \approx_S \nu\} \subseteq Lex$.

Now, let $h \in Lex$ be such that $h \not\approx_S \mu$. Then there is some $s \in S$ such that $h^*(s) \neq \mu^*(s)$. $h \vee g$ will not be defined on any g which is such that $g^*(s) = \mu^*(s)$, and so once $\langle s, \mu^*(s)\rangle$ is seen, Lex will contain only hypotheses consistent with it (as H will, and at the next iteration Lex contains only those hypotheses which were the join of some hypothesis in H with some hypothesis already in Lex), and thus at no later point will inconsistent hypotheses enter into Lex. As there are finitely many meaning maps (each is uniquely defined by its behaviour on the finite set Σ), there are thus a finite number of sentences that need to be seen to eliminate them. At that point, $Lex = \{\nu | \mu \approx_S \nu\}$.

After $Lex = \{\nu | \mu \approx_S \nu\}$, there will be no sentence which will change Lex, as for any hypothesis $h \in H$, for any $\nu \in Lex$ such that $h R \nu$, $\nu \vee h = \nu$. □

Returning to the question we began with, namely, when it is possible to exactly identify a class of languages, Theorem 3 provides us immediately with the following answer:

Corollary 1 For any $S \subseteq \Sigma^*$, $\mathcal{L}_S$ is exactly identifiable in the limit iff for any $\mu, \mu' \in \mathcal{L}_S$, $\mu \approx_S \mu' \rightarrow \mu = \mu'$.

[2]This is assuming that for every $w \in \Sigma$, there is some $s \in S$ such that $w \in s$.

3 Grounding Syntax

The semantic representations used above are almost completely disconnected from the syntax of the language - they are attuned only to which morphemes occur and how many times they do so. This notion of the syntax-semantics interface does not allow semantic representations to provide any information about the syntax of the language beyond simple numerical relations between elements in sentences. Our definition of compositionality is a naïve approximation of the more common usage, whereby the mode of combination of the meainings of the parts of expressions is related to the abstract syntactic structure of those expressions.

Luckily for language learners, the situation in natural language is (perhaps) less difficult. If we were to assume (see e.g. (Fulop, 1999) and related work) that the semantics of expressions very closely mirrors their syntactic structure, then the syntactic learner would have an easier job of things once the grounding had taken place. Note, however, that this relies on an aspect of the grounding problem that we have been able to ignore up to this point due to our overly simple semantics. Namely, that the more structured the meanings, the more work the grounding algorithm need do. (Kanazawa, 2001) is an investigation of this problem in a type-logical setting. We will continue to ignore it here.

In minimalist grammars (Stabler, 1997), there are two basic dependencies between morphemes given by the two structure building operations, MERGE and MOVE. In the Chomskyian tradition (Chomsky, 1965; Chomsky, 1986), the dependencies given by merger are those which correspond most closely to predicate argument structures (movement is usually taken to have scopal effects). Enriching our semantic representations to reflect the predicate argument structures corresponding to merger dependencies[3] would result in a much more robust syntax-semantics interface. This in turn will allow the learner to use the results of grounding to tightly constrain its initial hypotheses about the syntactic structure of its language. If in addition we require that movement dependencies are evidenced by string displacement from merged positions, this would enable the learner to reconstruct the dependency structures from which syntactic learning can be shown to successfully take place (Kanazawa, 1998; Stabler, 2001; Kobele et al., 2002).

4 Extending Grounding

The previous sections detailed a very idealized perspective on the grounding problem. One assumption we made was that the learner was able to determine exactly the intended meaning of the utterance. We can relax this assumption by redefining a text for a language. A referentially uncertain text is one in which each sentence of the language is paired not only with its meaning, but also with other possible meanings. The algorithms of §2 may be extended to referentially uncertain texts with varying degrees of success, depending in part upon how we choose the incorrect meanings. Note that every text in the sense of last section can be viewed as a (degenerate) referentially uncertain text. For a trivial case, if every possible meaning always accompanies each sentence of the language, there is no way to determine the 'correct' (even up to S-equivalence) meaning map for the language. On the other hand, if the incorrect meanings are chosen in such a manner so as to preserve the fact that the only sememes that constantly cooccur with a morpheme are exactly the meaning of

[3] In minimalist grammars this is simply the yield of the derivation tree for a sentence (Hale and Stabler, 2001).

that morpheme, even Algorithm 1 will (exactly) identify this class of languages. Another assumption was that the level of analysis of the sentence was abstract enough so as to filter out any ambiguity in the language (i.e. instead of the ambiguous *bank*, there are $bank_1$ and $bank_2$). Of course, this does not seem to be the case in (the early stages of) natural language learning.

5 Grounding and Language Change

In the previous sections we were unconcerned with the efficiency of the learning algorithms. In particular, Algorithm 2 might require huge amounts of computational resources to compute all partial hypotheses consistent with a particular datum - this cost will only increase with referential uncertainty. We can bound the computational resources required by Algorithm 2 by, for example, limiting the size of H - data for which there are more than a certain number of consistent hypotheses might be ignored.[4] This restriction, while possibly comprimising the learnability theoretic properties of the system, introduces a new possibility for language change. This doesn't represent a selectional pressure, but does introduce a perturbation in the linguistic system which other selectional pressures might interact with to give rise to changes over time.

However, even without modifying the learning algorithms themselves, the bare fact that the learner is *not* given arbitrarily much time to identify the text it is faced with introduces the possibility of imperfect (incomplete) learning. This is exploited in the system for linguistic emergence and transmission described in (Stabler et al., 2003).

References

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, Massachusetts.

Chomsky, Noam. 1986. *Knowledge of Language*. Praeger, NY.

Fulop, Sean. 1999. *On the Logic and Learning of Language*. Ph.D. thesis, University of California, Los Angeles.

Gold, E. Mark. 1967. Language identification in the limit. *Information and Control*, 10:447–474.

Hale, John and Edward P. Stabler. 2001. Notes on unique readability. ms. UCLA.

Kanazawa, Makoto. 1998. *Learnable Classes of Categorial Grammars*. CSLI Publications, Stanford University.

Kanazawa, Makoto. 2001. Learning word-to-meaning mappings in logical semantics. In R. van Rooy and M. Stokhof, editors, *Proceedings of the Thirteenth Amsterdam Colloquium*, pages 126–131. University of Amsterdam.

Kirby, Simon. 1999. Syntax of learning: The cultural evolution of structured communication in a population of induction algorithms. In D. Floreano, J-D. Nicoud, and F. Mondada, editors, *Advances in Artificial Life: 5th European Conference, ECAL'99*, Berlin. Springer-Verlag.

Kobele, Gregory M., Travis Collier, Charles Taylor, and Edward P. Stabler. 2002. Learning mirror theory. In *Proceedings of the Sixth International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6)*, Venezia.

[4] There are other means of bounding resources - perhaps only the first n hypotheses formed for a sentence are used. Or perhaps if there are more than n unknown morphemes in a data point it is ignored, etc.

Siskind, Jeffrey M. 1996. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61:39–91.

Stabler, Edward P. 1997. Derivational minimalism. In Christian Retoré, editor, *Logical Aspects of Computational Linguistics*. Springer-Verlag (Lecture Notes in Computer Science 1328), NY, pages 68–95.

Stabler, Edward P. 2001. Recognizing head movement. In Philippe de Groote, Glyn Morrill, and Christian Retoré, editors, *Logical Aspects of Computational Linguistics*, Lecture Notes in Artificial Intelligence, No. 2099. Springer, NY, pages 254–260.

Stabler, Edward P., Travis Collier, Gregory M. Kobele, Yoosook Lee, Ying Lin, Jason Riggle, Yuan Yao, and Charles Taylor. 2003. The learning and evolution of mildly context sensitive languages. In *European Conference on Artificial Life, ECAL'03*.

Steels, Luc. 1996. Synthesizing the origins of language and meaning using co-evolution, self-organisation and level formation. In J. Hurford, C. Knight, and M. Studdert-Kennedy, editors, *Evolution of Human Language*. Edinburgh University Press, Edinburgh, pages 161–165.

Creole Viewed from Population Dynamics

Makoto Nakamura*[†], Takashi Hashimoto[‡] and Satoshi Tojo[†]
Graduate School of {[†]Information, [‡]Knowledge} Science,
Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Tatsunokuchi-machi, Nomi-gun, Ishikawa, 923-1292, Japan

Abstract

Creole is one of the main topics in various fields concerning the language origin and the language change, such as sociolinguistics, the developmental psychology of language, paleoanthropology and so on. Our purpose in this paper is to develop an evolutionary theory of language to study the emergence of creole. We discuss how the emergence of creole is dealt with in the perspective of population dynamics. The proposal of evolutionary equations is a modification of the language dynamics equations by Komarova et al. We show experimental results, in which we could observe the emergence of creole. Furthermore, we analyze the condition of creolization in terms of similarity among languages. We conclude that a creole becomes dominant when pre-existing languages are not similar to each other and rather similar to the newly appeared language (would-be-creole); however the new language must not be too similar, in which case pre-existing languages remain and coexist.

Keywords: Population Dynamics, Creole, Similarity among Languages, Language Dynamics Equations

1 Introduction

Generally, all human beings can learn any human language in the first language acquisition. One of the main purposes of language use is to communicate with others. Therefore, it is easy to consider that the language learners come to obtain the language which they hear most in the community, i.e., in most cases, children will develop their parental languages correctly. When people do not have a common language to communicate with each other, such as plantation economies, slave trade, and so on, they come to use a simplified language called *pidgin* to bridge communication gaps between speakers of mutually unintelligible languages. After that, the children of the pidgin speakers may obtain a full-fledged new language called *creole* as their native language [6]. Thus, children have an ability to learn and create the most communicative language in the community.

In the stream of simulation studies of language evolution [4], the emergence of creole is also studied [10]. Briscoe [3] has reported sophisticated models of human language acquisition by means of a multi-agent model. However, because the number of agents was finite, the results were often hard to be generalized to explain general phenomena in the real world, from which the most multi-agent models had suffered.

To overcome this drawback of multi-agent models from a different viewpoint, Nowak et al. developed mathematical theory of the evolutionary dynamics of language [13]. By defining similarity and payoff between languages, based on the assumption of the universal grammar, Komarova et al. [8]

proposed *language dynamics equations* in which the transition of population among finite number of languages described by differential equations. However, in the framework of evolutionary dynamics of language, the emergence of creole was not discussed yet.

Our purpose in this paper is to develop the evolutionary theory of language in order to investigate the emergence of creole. We have already seen creolization by introducing the assumption that language acquisition of children is affected both by the distribution of population and by the exposure rate to other languages than their parental one [9]. In this paper, we analyze the condition of relationship among languages for creole to emerge and to be dominant.

In Section 2, we discuss how we consider creolization in the context of population dynamics. In Section 3, we describe the language dynamics equations and our modification of the equations. Section 4 reports our experiments. We present a discussion and a conclusion in the last two sections.

2 Creolization in Population Dynamics of Language

In this section, we describe creole from the viewpoint of population dynamics. We showed the emergence of creole in population dynamics of language [9], which is caused by transition of population among grammars and the exposure probability of children. Here, we discuss how the emergence of creole is considered in population dynamics.

2.1 Creole and Population Dynamics

We presuppose that the emergence of creole strictly depends on the population distribution, as opposed to traditional linguistic explanations [2, 6]. From the viewpoint, a creole is considered as such a grammar G_c that; A) $x_c(0) = 0$, $x_c(t) > \theta_c$ or B) $x_c(0) = 0$, $x_c(t) > \theta_d$, where $x_c(t)$ denotes the distribution of the population of G_c at time t, and θ_c and θ_d denote certain thresholds to be regarded as *coexistent* and *dominant*, respectively. These definitions represent that some individuals come to speak a language that no one spoke at the initial state, and consequently, A) a fixed number of individuals keeps the grammar, and B) the distribution of the language speaker occupies the most in the community.

2.2 Similarity among Languages

The S matrix in population dynamics denotes the similarity between grammars, which is determined by the probability s_{ij} that a speaker who uses a grammar G_i will say a sentence that is understandable by speakers of another language G_j; thus, each of $S = \{s_{ij}\}$ is a constant diachronically. Generally, the S matrix is uniquely calculated when the grammars and the probability for each sentence are given. Suppose that an individual who uses G_i utters a sentence in $L(G_i)$ with the uniformly same probability, then s_{ij} is the number of common sentences between $L(G_i)$ and $L(G_j)$ divided by the number of sentences in $L(G_i)$. Therefore, diagonal elements of the S matrix are always 1. Under the assumption, Fig. 1 shows the relationship among languages in the S matrix. The shaded part in the figure denotes that $L(G_1)$ and $L(G_2)$ share common sentences. In this case, s_{12} is greater than s_{21}, because the common part is rather small in $L(G_1)$ than in $L(G_2)$. The size of $L(G)$ concerns the generative capability of the grammar. Because the power of expressiveness is considered to be similar among languages, we should regard that the size of $L(G_i)$'s are same and that s_{ij} is nearly equal to s_{ji}. Thus, the S matrix should be an approximately symmetrical matrix.

In the above discussion, it is assumed that the member of conceivable grammars is finite and predefined. In this sense, creole is also included in them and has the similarity with the other languages

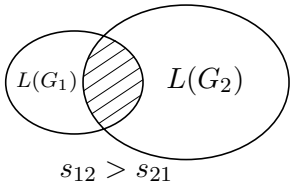


Fig. 1 The relationship among languages in the S matrix

in the S matrix. This is justified by the perspective of the universal grammar. We presuppose that creole may occur according to the similarity to the other languages, and thus we study the conditions of the similarity for creolization.

3 Population Dynamics of Grammar Acquisition

We introduce modified language dynamics equations after reviewing Komarova et al. [8]'s original ones.

3.1 Komarova et al.'s Language Dynamic Equations

Komarova et al. [8, 13] proposed a mathematical theory for the evolutionary and population dynamics of grammar acquisition. In their model, given the principles in the universal grammar, the search space for candidate grammars is assumed to be finite, that is $\{G_1, \dots, G_n\}$. Let $x_j(t)$ be the ratio of the population of G_j speakers, where $\sum_{j=1}^{n} x_j(t) = 1$. Thus, the model is defined in population dynamics in which individuals change their own grammar from generation to generation. The language dynamics equations are mainly composed by (i) the similarity between languages as the matrix $S = \{s_{ij}\}$ and (ii) the probability that children fail to acquire their parental language as the matrix $Q = \{q_{ij}\}$. Individuals reproduce children, the number of which is determined by the *fitness* such as: $f_i(t) = \sum_{j=1}^{n}(s_{ij} + s_{ji})x_j(t)/2$. The language dynamics equations are given by the following differential equations:

$$\frac{dx_j(t)}{dt} = \sum_{i=1}^{n} q_{ij} f_i(t) x_i(t) - \phi(t) x_j(t) \qquad (j = 1, \dots, n), \tag{1}$$

where $\phi(t) = \sum_{i=1}^{n} f_i(t)x_i(t)$ and the term '$-\phi(t)x_j(t)$' makes the total population size keep constant.

In those equations, the fitness f_i for each grammar is regarded as its communicability, which represents a probability that a sentence uttered by an individual is recognized in the community. Total distribution of children of G_i speakers becomes $f_i x_i$. By the definition of the Q matrix, children are allowed to make mistakes during language acquisition. It is possible for a child to learn grammar G_i from her parents and to end up speaking grammar G_j. The probability of such transition is defined as $Q = \{q_{ij}\}$. In their work, it is also assumed that only adult individuals talk to the other language groups, while children communicate with only their parents. In this circumstance, it may be difficult to consider that the children mistake their parental grammar for another one.

3.2 Niyogi's Model

Niyogi [11, 12] gives actual examples of the Q matrix with linguistically well-grounded grammars together with the trigger learning algorithm (TLA) [7]. However, there is an unrealistic Markov structure which implies that some children cannot learn certain kinds of language, as we pointed out in [9].

3.3 Our Modification

Thus far, we have modified the language dynamics equations to include some constraints concerning transition among languages [9]. We have shown by computer simulations [10] that the population could be changed when children are exposed not only to their parental language but also to other languages. It is reasonably supposed that the transitions depend on the distribution of population of languages for children to be exposed. Therefore, the Q matrix should change through generations. Our prime revision is to introduce the probability α that children are affected by the other language speakers than their parents. We call α the *exposure probability*. A child hears not only parental language but also other languages in proportion both to the rate of the exposure α and to the distribution of population of grammars (See Fig. 2(a)). The probability which the children learn a language from their parents comes to $(1-\alpha)$. Note that α does not exclude children's parental language; it is also included in α in proportion to the distribution of population as well as the other languages.

Since the distribution of population changes in time, the Q matrix should include the time parameter t, that is, Q is redefined as $\overline{Q}(X(t)) = \{\overline{q}_{ij}(t)\}$, where $X(t) = (x_1(t), x_2(t), \ldots, x_n(t))$. We call $\overline{Q}(X(t))$ the *modified accuracy matrix*. Together with the S matrix and a given α, a learning algorithm determines $\overline{Q}(X(t))$. Thus, the new language dynamics equation is as follows:

$$\frac{dx_j(t)}{dt} = \sum_{i=1}^{n} \overline{q}_{ij}(t) f_i(t) x_i(t) - \phi(t) x_j(t) \qquad (j = 1, \ldots, n). \tag{2}$$

3.4 The Learning Algorithm

We introduce a simple learning algorithm which resolves Niyogi [11]'s problem mentioned above. The learning algorithm becomes as follows (See also Fig. 2(b)):

1) In a child's memory, there supposed to be a score table of grammars.

2) The child receives a sentence uttered by an adult.

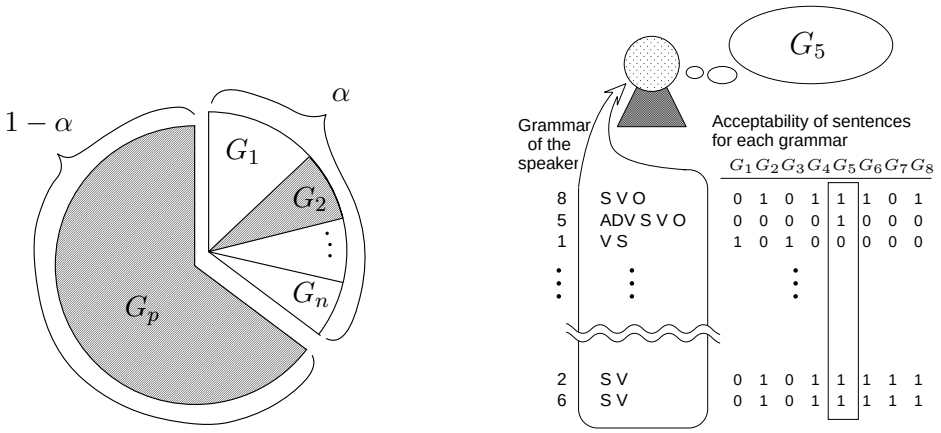
3) For each grammar, if a sentence is acceptable for the child, the grammar scores a point in her memory.

4) 2) and 3) are repeated until the child receives a fixed number of sentences that is regarded as enough for the estimation of the grammar.

5) The child adopts the grammar with the highest score.

Here, we introduced the exposure probability α that prescribes the ratio a child talks to people other than her parents. Thus, the estimated grammar of the child is G_{j^*} such that:

$$j^* \quad = \quad \underset{j}{\mathrm{argmax}} \{\alpha \sum_{k} s_{kj} x_k(t) + (1-\alpha) s_{pj}\}. \tag{3}$$



(a) The exposure probability α ($p = 2$)

(b) The simplest learning algorithm

Fig. 2 Introducing the exposure probability and the learning algorithm

From the learning algorithm, we give the modified accuracy matrix $Q(X(t)) = \{\overline{q}_{ij}(t)\}$ in [9] as follows:

$$\overline{q}_{ij}(X(t)) = \frac{(\alpha \sum_k s_{kj} x_k(t) + (1-\alpha) s_{ij})^{n-1}}{\sum_l (\alpha \sum_k s_{kl} x_k(t) + (1-\alpha) s_{il})^{n-1}}. \tag{4}$$

4 Experiments

In this section, we show the experimental result of the language dynamics equations of population-based transition in Section 3. We examine the conditions that creole appears and comes to be dominant in combinations of the S matrix.

4.1 Settings

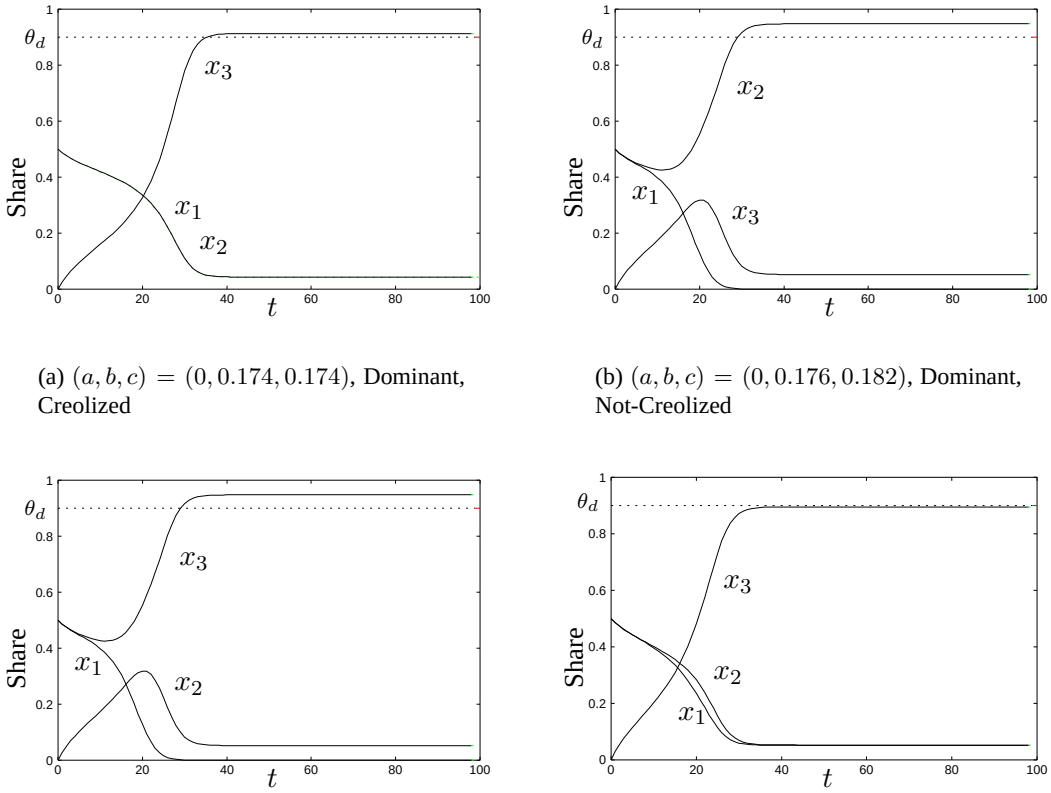
Here, we give parameters for the experiments. Since it is clear that creolization is the most observable in case $\alpha = 1$, we examine this case through the experiments. We analyze the case of three languages with the symmetry in s_{ij} and s_{ji} from the reason explained in Section 2.2, that is, the S matrix is formed as below:

$$S = \begin{pmatrix} 1 & a & b \\ a & 1 & c \\ b & c & 1 \end{pmatrix}. \tag{5}$$

The initial populations are given as $x_1(0) = x_2(0) = 0.5, x_3(0) = 0$. Therefore, we parametrize a, b and c in Eqn (5), and then research the mutual dependency in which G_3 becomes creole.

4.2 Conditions of Creole to be Dominant

The experiment aims at finding boundaries in the parameter space as to which language would be dominant. We refer to this situation as *dominant creolization*. Fig. 3(a) shows that the creole G_3 is



(a) $(a, b, c) = (0, 0.174, 0.174)$, Dominant, Creolized

(b) $(a, b, c) = (0, 0.176, 0.182)$, Dominant, Not-Creolized

(c) $(a, b, c) = (0, 0.182, 0.176)$, Dominant, Not-Creolized

(d) $(a, b, c) = (0, 0.188, 0.189)$, Not-Dominant, Creolized

Fig. 3 The relationship between dominant creole and the S matrix

dominant, in which the threshold for a language to be dominant is defined as $\theta_d = 0.9$. The S matrix is set to $(a, b, c) = (0, 0.174, 0.174)$, that is $b = c$. The value of $a = 0$ denotes that there is no common sentence in G_1 and G_2. Because the languages G_1 and G_2 play same roles, the dynamics of x_1 and x_2 are completely the same. In the figure, the language G_3 which no one spoke at the initial state comes to occupy the population with the rate of more than θ_d, while x_1 and x_2 declined concurrently. Namely, this is the emergence of a dominant creole in population dynamics.

When the values b and c increases slightly, the dominant language changes to another one while the share of creole G_3 is getting smaller. Fig. 3(b) represents the dynamics with the S matrix set to $(a, b, c) = (0, 0.176, 0.182)$. The figure denotes that G_2 becomes dominant, while G_1 eventually disappeared though it had the same population with G_2 at the initial state. When we transposed the value of b and c as $(a, b, c) = (0, 0.182, 0.176)$, the dynamics does not change but the dominant language is replaced (See Fig. 3(c)).

Changing the values of b and c continuously, we observed the sheer boundary of the change of the dominant language between them. Fig. 4 shows that the boundaries for the creole (G_3) to be dominant for several values of a. The crosses ($\times$) in the figure represent the parameter values corresponding to Fig. 3(a)–(d), respectively. Because the parameters b and c work similarly G_1 and G_2, the boundaries are symmetric along the line $b = c$. In the figure, the long curve of the outmost boundary ($a = 0.00$)

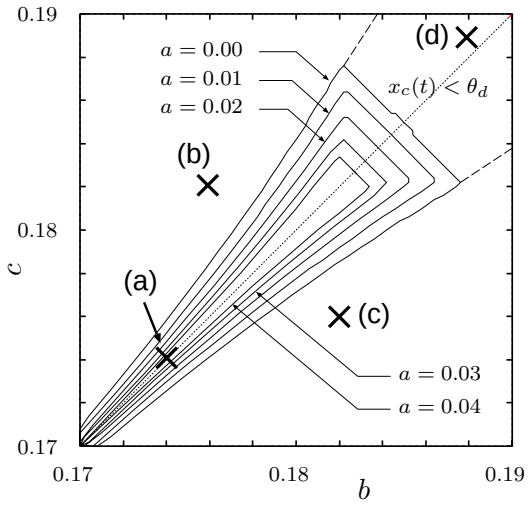


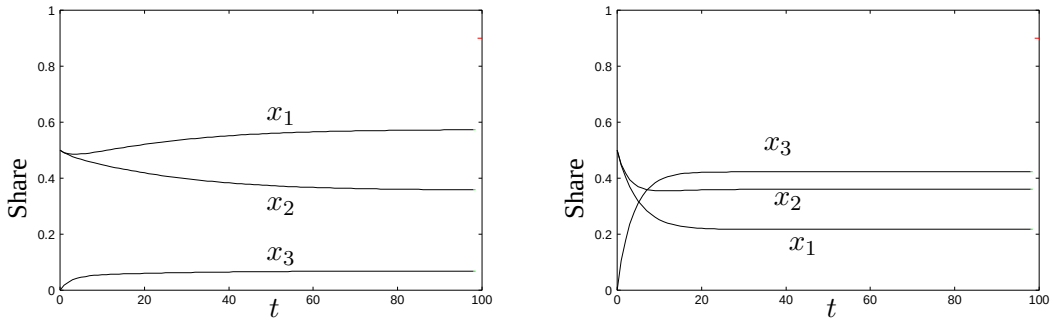
Fig. 4 Conditions for Dominant Creole ($\theta_d = 0.9$)

intersecting between (a) and (b) in Fig. 4 stands for the boundary of the change of the dominant language. Inside of the lines, G_3 is the dominant language (Fig. 3(a)), above of the upper line, G_2 is dominant (Fig. 3(b)) and below of the lower line, G_1 is (Fig. 3(c)). Even if the threshold for dominant language, θ_d, were eased to be lower, this boundary had not changed. It is also the case with the different value of a. Thus, the long side of boundaries among dominant language is independent of θ_d for a given value of a. The broken lines in Fig. 4 are the boundaries of the dominant creolization for smaller values of θ_d at $a = 0.00$.

Next, we consider the short side of the boundaries in Fig. 4, that is the line crossing perpendicularly to the line $b = c$ (dotted line). This also represents critical conditions whether creole occurred or not for several values of a. These boundaries are, however, different from the one mentioned above. In Fig. 3(d) with $(a, b, c) = (0, 0.188, 0.189)$, we observed that G_3 still remained as the most populous language although the rate x_3 was a little less than $\theta_d = 0.9$. If θ_d was eased to lower, G_3 at the parameters of (c) would be regarded as creole. Hence, the position of the short line can shift along the line $b = c$ with the value of θ_d. It is easy for us to recognize that higher θ_d shrinks the area of creolization in the parameter space and vice versa.

As the larger the values b and c, the larger population transfer from G_1 and G_2 to G_3, respectively, and the width between the upper and lower boundaries grows. At the same time, however, $x_3(t)$ converges to smaller values at $t \to \infty$ with larger b and c. At last, $x_3(t \to \infty)$ falls short of θ_d at the short side boundaries in Fig. 4. This is because the more population shifts from G_3 to G_1 and G_2 by the larger values of b and c. Inversely, for the smaller b and c, say $b \approx c \lesssim 0.135$, in spite of the large share of x_3, the time needed for G_3 to dominate the all population comes to be longer that we could not observe further creolization.

To observe further details of the region of creole, we parametrized a in Eqn (5). In Fig. 4, regions of creole come to narrow with increasing a. Since a large value of a promotes communicability between G_1 and G_2 and enlarges the transition between them, no large population shifts from them to G_3. Therefore, the increase of a results in no dominant creolization.



(a) $(a, b, c) = (0.35, 0.25, 0.1)$, Coexistent, Not-Creole

(b) $(a, b, c) = (0.3, 0.4, 0.5)$, Coexistent, Creole

Fig. 5 Coexistent-language set

4.3 Summary of the Results

The conditions of the off-diagonal elements a, b and c in the symmetric similarity matrix S for the emergence of dominant creole is:

$$a \lesssim 0.1 \tag{6}$$

$$0.13 \lesssim b \simeq c \lesssim 0.2 \tag{7}$$

In this range, changes of a, b and c result in the followings:

1) When b and c are large, a must be small; in which case b and c might much differ. In this case, the share rate of G_3 becomes rather small at the time of convergence.

2) On the contrary, when a is small enough, b and c should be small. In this case G_3 dominates and converges in a short period.

5 Discussion

5.1 Conditions of Creolization in Natural Language

We obtained the condition of similarities among languages, in which a creole emerges and is to be dominant. Let us consider what this condition implies in the context of natural language. Suppose two languages, say super-stratum and sub-stratum languages. The condition $a \lesssim 0.1$ (Eqn (6)) indicates that these two languages are not similar. The two languages must be less similar to each other than to creole. If they are similar enough, the communication gap between users of these two languages is not so wide that they can understand each other to some extent. Thus, no pidgin or creole is needed. The condition, Eqn (7), says that the values b and c should not be too small but should be relatively small. If the pre-existing languages are similar enough to the creole, that is, the second in equality of the condition, Eqn (7), does not hold, the creole emerges but the users of the pre-existing languages can communicate with the creole users, then the speakers of the pre-existing languages do not diminish. The former part of the condition, Eqn (7), means that when a newly appeared language

has no similarity to two pre-existing languages, it hardly becomes a creole.[1] Eqn (7) also confines the similarity of the pre-existing languages to the creole within a narrow range ($b \simeq c$). When the similarity of creole to the super-stratum language is enough larger than that of to the sub-stratum language, the super-stratum language comes to be dominant, and vice versa.

5.2 Language and Dialect

In this paper, we thoroughly analyzed the parameter region at which creole is dominant. When we look at the whole parameter space, we found the following four categories about dominance and creolization:

i) Dominant and Creolized; like Fig. 3(a)

ii) Dominant and Not Creolized; like Fig. 3(b)

iii) Coexistent (no dominant language) and Creolized like; Fig. 5(a)

iv) Coexistent and Not Creolized; like Fig. 5(b)

According to our preliminary investigation, the parameter region of the coexistent categories (iii) and iv)) is $a, b, c \gtrsim 0.3$, where the similarities among languages are relatively high and at this rate the language users can communicate with each other to some extent. This situation is better to be regarded as dialects rather than different independent languages. There is, in general, no clear boundary between dialects in a language and different languages from the pure-linguistic viewpoint.[2] From our results, the similarity of 0.3 may be a rough criterion for dividing between them.

6 Conclusion

In this paper, we argued that the emergence of creole in population dynamics of languages and showed that the emergence is affected by the similarity among languages as well as the distribution of population of the languages in the community. We obtained results for the condition of the similarity for dominant creolization as follows.

A) The pre-existent languages are not similar to each other, but to the newly appeared language.

B) The newly appeared language must not be too similar to the pre-existent languages. Otherwise, the pre-existent languages remain and coexist.

C) The pre-existent languages have approximately same distance to the newly appeared language with regard to similarity.

Creolization has not been dealt from the viewpoint of population dynamics and similarity among pre-existing and a creole, although similarity among creoles has been investigated [1]. Our contribution is to address a prediction about similarity among languages for creole to develop. This prediction should be tested empirically by observing grammars of various creoles and their original super- and sub-stratum languages.

[1] Since the similarity here is not the extent of the mixture of grammars in two languages, this implication does not contradict the fact that grammar of a creole is not a blend of those of super- and sub-stratum languages.

[2] The boundary is often settled politically such as Serbian, Croatian and Bosnian in Bosnia and Herzegovina. [5]

We argued the relationship between a language and dialects. Since the difference between language and dialect concerns the grammatical features, it is not possible to distinguish them only with the similarity, much less creole. This is an important problem in the present population dynamics. Therefore, further progress is needed to develop linguistic features into the population dynamics. We need to study in further generalized and actual conditions, to clarify the boundary conditions of creolization.

References

[1] Bickerton, D.: Roots of Language, Karoma Publishers, Ann Arbor, MI (1981)

[2] Bickerton, D.: Language and Species, The University of Chicago Press, Chicago (1990)

[3] Briscoe, E.J.: Grammatical Acquisition and Linguistic Selection, In: Briscoe, T. (ed.): Linguistic Evolution through Language Acquisition, Cambridge University Press, Cambridge (2002) pp.255-300

[4] Cangelosi, A., Parisi, D. (eds.): Simulating the Evolution of Language. Springer, London (2002)

[5] Comrie, B., Matthews, S., Polinsky, M.: The Atlas of Languages Quatro Publishing, London (1996)

[6] DeGraff, M. (ed.): Language Creation and Language Change, The MIT Press, Cambridge, MA (1999)

[7] Gibson, E., Wexler, K.: Triggers, Linguistic Inquiry, **25** (1994) pp.407-454

[8] Komarova, N.L., Niyogi, P., Nowak, M.A.: The Evolutionary Dynamics of Grammar Acquisition, J.Theor.Biol. **209** (2001) pp.43-59

[9] Nakamura, M., Hashimoto, T., Tojo, S.: The Language Dynamics Equations of Population-based Transition – a Scenario for Creolization –, Proceedings of the 2003 International Conference on Artificial Intelligence (IC-AI'03), CSREA Press (2003) (to appear)

[10] Nakamura, M., Tojo, S.: The Emergence of Artificial Creole by the EM Algorithm, Proceedings of the Fifth International Conference on Discovery Science (DS2002), Lecture Notes in Computer Science 2534, Springer (2002) pp.374-381

[11] Niyogi, P.: The Informational Complexity of Learning from Examples, PhD thesis, Massachusetts Institute of Technology, Cambridge, MA (1994)

[12] Niyogi, P., Berwick, R.: The logical problem of language change. Technical Report AI Memo 1516 / CBCL Paper 115, MIT AI Laboratory and Center for Biological and Computational Learning, Department of Brain and Cognitive Sciences (1995)

[13] Nowak, M.A., Komarova, N.L.: Towards an evolutionary theory of language, Trends in Cognitive Sciences **5**(7) (2001) pp.288-295

Modeling Phonological Change

Lee Hartman
Southern Illinois University
lhartman@siu.edu

1. Overview.

In the following pages I present some desiderata discovered, as well as some questions raised, by the experience of developing, over a period of some twenty years, a program, named Phono, for creating and operating models of regular historical sound change. Phono is now in the beta-testing stage of its fourth version (Hartman 2003a and 2003b). The approach is directed toward producing a practical tool for testing given hypotheses about sound-change rules and their order, and (in the case of languages with undocumented ancestor forms) for testing given hypotheses of reconstructed forms—rather than, say, for generating such hypotheses of rules or forms computationally. Given an ordered set of rules, the model performs a single line of derivation, for one word at a time, downstream from etymon to reflex.

The Phono project has evolved through several incarnations. Version 1 was written around 1980 in PL/I to run on an IBM 360 mainframe whose output was limited to the uppercase letters of the Roman alphabet. A model for deriving Spanish words from Latin was hard-coded in the program. (The Spanish model is based mainly on Otero 1971 and Hartman 1974.) Versions 2 and 3 were written in DOS-based Pascal for microcomputers. In these versions the rules of the Spanish model were extracted from the program code and recast as data for the program to read and interpret, thus opening the possibility for Phono to operate models for other languages. The output notation gained access to both upper- and lowercase letters, as well as the rest of the extended ASCII character set, which was used as an ad hoc phonetic alphabet. And finally, Version 4 (in Visual Basic, for Windows, released in 2003) displays output in a phonetic font, essentially the alphabet of the International Phonetic Association (IPA) (©1993 by SIL International, www.sil.org, used with permission), thus solving the main problem cited by Becker (1996) in his review of Version 3.2.

The main practical issues encountered have been those of notation (see Hartman 1993a and 1993b): notation of data words as input, as output, and during the derivation; and notation of sound-change rules. Based on the experience with Phono, I recommend, below, forms of notation (1) for the input of etymon words, (2) for the internal representation of words during derivation, and (3) for displaying the output of derived forms with both readability (the conventional IPA alphabet) and preservation of unexpected details in the results ("feature-based diacritics", explained below). I also present a versatile form of notation for rules as data and consider ways in which this rule notation might be made more user-friendly to the historical phonologist.

Additionally, I recommend mechanisms to handle some deviations from linear rule order (Chafe's concept of the "persistent" rule, and a provision for temporarily

"masking" a rule). I consider some additional complications of rule order and ways in which the program might be modified to handle them.

I recommend a procedure for processing many pairs of words—etymon and known reflex—together in a "batch" mode, in order to test and maintain the integrity of a model during its development.

Finally, I describe a "rule trace" procedure for monitoring the functional load of each rule.

2. Theoretical assumptions.

In order to make Phono potentially useful to a broad variety of researchers on language history, I have attempted to keep the program as nearly theory-neutral as possible. Nevertheless, the project is based on a set of arguable assumptions, including the following: (1) that sound change is regular *in some degree*; (2) that regular sound change can be modeled as a chronologically ordered set of rules, each of which acts on the output of its predecessor; and (3) that the chronological order of the rules is the same for most words in the language. Even while adopting these assumptions, it is important to bear in mind that *regular* change is only one part of language history. Exceptions to regular change do occur, either explicably—through paradigmatic analogy, dialect mixing and other borrowing, "lexical diffusion" (i.e. gradual and sometimes incomplete passage of a change through the vocabulary—see Wang 1969), influence of written language, or other phenomena—or, let's confess, inexplicably.

"Rules" of sound change, formerly called "laws", are of course not *prescriptive* rules: no explicit authority has ever encouraged sound change. But what then *is* the relationship between the rules and the sound-change phenomena? Phono makes no claim as to whether rules "describe", "explain", "cause", or "reflect" the evolution of sounds in people's speech. As a neutral term, we may say that a rule "corresponds to" a set of changes, until such time as other investigators decide what cause-and-effect relations may be at work, or what "psychological reality" a rule may have. And Phono makes no claim about how or why sound changes are propagated through social and geographical space. These questions are not unimportant; they are merely beyond the intended scope of the program.

3. Internal notation: binary features.

From its beginning, the Phono program has been based on the expression of sound-change rules in terms of binary feature values. Binary features were favored by the early proponents of (synchronic) generative phonology for theoretical reasons: certain combinations of feature values efficiently define various "natural classes" of sounds, i.e. groups of sounds that tend to behave similarly (for example, p > b tends to be accompanied by t > d and k > g, all three of which can be expressed collectively as [–voice] > [+voice]). This aspect of features also partially motivates their use by Phono, but the feature mode of notation is chosen here mainly for its versatility and precision.

Phono has adopted, with minimal modification, most of the features defined in *The Sound Pattern of English* (Chomsky and Halle 1968:298-329)—known as "*SPE*"—(20 of them in the present version) because that work represents a unique moment of near-consensus in the development of phonological theory. Since its publication, the role of features in phonological theory has been discussed vigorously—with the case being made variously for replacement of some of the *SPE* features, "feature geometry" (i.e. hierarchical relationships among the features), "unary" features, multi-valued features, underspecified features, feature markedness, etc. But it is not a goal of the Phono project to enter into the debates on synchronic phonological theory. The *SPE* features are fully adequate for Phono's needs of precision and versatility, but their use is not intended to make any further claim about their "psychological reality" or theoretical significance. Perhaps future versions of Phono and computational models of synchronic phonology can share insights with mutual benefit.

4. Encoding, change, and decoding.

While on one hand the sound-change rules are expressed entirely in terms of binary feature values, on the other hand the typical user deals more easily with character strings. As a result, there is a need for translation of character strings into and out of feature notation. In this regard the interactive derivation consists of five phases: input, encoding, diachronic changes, decoding, and display. (1) The user feeds the etymon word to the program as a string of keyboard characters. (2) The program encodes the word, translating the character string to a set of feature values. (3) The program performs the derivation, changing the word's feature values according to the ordered series of sound-change rules. (4-5) After each diachronic change, the program decodes the word's feature values back to a character-string notation for display on the screen.

4.1. Input and encoding.

In practice, the encoding process for etymon input is more complex than a mere lookup procedure in a table of correspondences between characters and feature values. Of the 81 phonetic symbols of the IPA alphabet, only 26 are directly represented on the keyboard (i.e. the lowercase letters of the Roman alphabet). In the likely event that the ancestor language's phonological inventory includes sounds that differ in some respect from those represented in the IPA alphabet by the 26 letters, the user must devise an unambiguous "orthography" for etymon input from the keyboard. This input notation need not be restricted to a one-character-to-one-phoneme correspondence; for example, it may use digraphs (such as <bh> for the bilabial fricative beta), or, conversely, it could be made to allow Latin orthographic <x> to represent the series /ks/. These devices of input notation can be given their phonological interpretation by "etymon input adjustment rules". (The adjustment from <bh> to beta, for example, would be carried out by a rule that makes /b/ fricative before /h/ and then deletes /h/.) In other words, the apparatus for etymon input consists of (1) an alphabet of keyboard characters with feature values fully specified, held in a lookup table, and (2) a series of adjustment rules that can serve to interpret digraphs, assign context-sensitive feature values (such as nasal assimilation or predictable stress), or otherwise fine-tune the feature configuration of the input word before the derivation begins.

4.2. Diachronic change.

The internal representation of the word as a bundle of feature values is subjected to the series of changes specified by the ordered set of rules. Each rule (consisting of "IF-lines" and "THEN-lines", as explained below), in chronological order, performs a scan of the word, from right to left, examining each segment to see if it and its environment fit the pattern of conditions necessary for the change to occur. In the event that the conditions are met—that the IF-lines are collectively "true"—then the set of THEN-lines is executed, making the prescribed changes to the feature values of the segment on which the scan is currently focused. Each time a change takes place, the feature representation of the word is decoded to display that successive stage in the word's evolution.

4.3. Decoding and display.

While an input alphabet must be limited to the characters available on the keyboard, the output display in the present version of the program has access to the 81 characters of the IPA font. And yet, 20 binary features offer the prospect, mathematically, of the 20th power of 2, or more than a million different combinations. Of course not all these combinations are phonetically possible, but even in practical terms, the feature notation for a segment potentially carries more information than an unmodified IPA phonetic symbol can. So the challenge for decoding is not only to translate from features to phonetic symbols, but also to preserve somehow any information that is lost in the process.

Each phonetic symbol in the output alphabet is "fully specified" with feature values; that is, every feature is assigned either a "plus" or a "minus" value. Additionally, in the alphabet's lookup table, some of these positive and negative values are marked with "double" signs (#, =), while others have "single" signs (+, –). The double signs of a particular phonetic symbol designate the minimal set of features sufficient to differentiate that symbol from any other symbol in the alphabet, and it is only these essential, defining, double signs that are used in pattern-matching to select a phonetic symbol for output. After this selection is made, the values of the nonessential, single-sign features are compared, and if any of these values of the segment differ from the default values in the alphabet, the output word is highlighted in blue in the display. (Unfortunately Visual Basic does not permit individual symbols in the character string to be colored in contrast to others in the same string.) Specification about the discrepancy—which features of which segments differ from those segments' nearest equivalents in the alphabet—can be displayed as "feature-based diacritics" in a box at the top of the screen (see Burton-Hunter 1976 and Hartman 1981).

5. Rule notation.

Most linguists are familiar with a form of rule notation that can be represented schematically as "A > B / C _ D", meaning that element A becomes B in the environment

following C and preceding D, or, in other words, each instance of CAD becomes CBD. CAD is known as the "structural description", and the change of A to B is called the "structural change". Chomsky and Halle (1968) used this template extensively, along with signed feature names, to write (synchronic) phonological rules. Superficially, this notation seems simple, but it must be noted that many of Chomsky and Halle's rules are supplemented by additional devices such as parentheses, curly braces, angled brackets, "diacritic" features, category symbols with subscripts (e.g. $C_1 \ldots C_n$ for different consonants), "truth-functional conditions", etc. (1968:390-399). In order to capture these additional complexities that occur, Phono portrays the rules in a unique notation that is based on *if*-clauses and *then*-clauses that are composed mainly of binary feature values and expressions of locations in the word.

Specifically, Phono's rule notation system is based on four types of "IF-lines" (Branching, COUNT, Constant, and Variable) and five types of "THEN-lines" (Constant, Variable, DELETE, INSERT, and SWAP). (The COUNT, DELETE, INSERT, and SWAP line types are identified in the notation by these respective keywords in all-uppercase letters.) Since the Constant and Variable line types are able to carry out either IF or THEN functions, there are seven types altogether.

The Branching-type IF-line is the basis of the *hierarchy* of IF-lines: it joins the labels of two following IF-lines with a conjunction, either "and" or "or". If the rule has more than one IF-line, the first of them, labeled A, must be a Branching line that states the relationship between lines B and C. Then B and C in turn may be of other line types, or either of them may also branch, into D and E, and so on. Line A, as the top member of the hierarchy, must represent all the IF-lines of the rule combined, and each IF-line below line A must be represented in a Branching line somewhere above itself.

The "Constant" and "Variable" line types, which can do double duty as both IF-lines and THEN-lines, are composed of feature values and references to locations in the word. As IF-lines they *detect*, and as THEN-lines they *alter*, selected feature values in the word. In a Constant-type line the feature names appear with specific values ("+" or "–"), while in a Variable-type line they detect or set the value of one feature equal (or opposite) to that of another — thus corresponding to the "Greek-letter variable" signs (the so-called "alpha" device) introduced in *SPE* (pp. 177-178).

Rules can express the location of a segment in the word either as an "absolute" location (i.e. with reference to the word-initial or -final positions) or as a "relative" location (i.e. with reference to the current focus segment of the scan). A Constant-type IF-line can be used to express conditions such as "If the word-initial segment (absolute location) is [+nasal]", or "If the segment immediately following the 'focus' segment (relative location) is [–continuant]". A Constant-type THEN-line, similarly, can be used to express changes such as "The word-final segment (absolute) becomes [–voice]", or "The focus segment (relative) becomes [+strident]".

A Variable-type IF-line seeks a feature value defined with regard to (i.e., the same as, or the opposite of) another feature value in the word. It can be used, for example, to

express a condition such as "If the [back] and [round] features (of a vowel) agree" — in order to restrict the change to the set of back-rounded and front-unrounded vowels. Used as a THEN-line, the Variable-type line appears in rules of assimilation (as well as in those of dissimilation), where the change of one feature may be to "+" or to "–" according to the value of a neighboring feature. So, for example, in a rule ofvoicing assimilation, it would be a Variable-type line that sets the value of the feature [voice] for one consonant equal to that of the same feature in the following consonant.

Most of the conditions for change (IF-lines) can be expressed by some combination of these Constant- and Variable-type lines based on feature values. But some changes depend additionally on some characteristic of the entire word (for example whether it has one syllable or more than one; or whether it contains a stressed syllable or not), or on a characteristic of some "subscan" of segments within the word (for example whether a certain feature value occurs in a segment located between the focus segment and a potential agent of change; or whether a vowel in focus is the last vowel in the word, regardless of how many consonants may follow it). These conditions are detected by means of the COUNT-type line, which specifies the "head" and the "foot" of the subscan, the feature value being sought, and the number of finds necessary to make the line "true".

Finally, some changes, rather than affecting merely some feature value(s) of the focus segment, may delete the entire segment, change its location in the word (metathesis), or insert an entire new segment. These whole-segment changes are brought about by the DELETE-, SWAP-, and INSERT-type THEN-lines, respectively.

Metathesis is treated as an interchange, a "swap", of positions of two segments (rather than as a movement by the focus segment some number of positions to the left or right) based on real changes such as that of Spanish *miraglo* > *milagro* ('miracle'), in which two non-adjacent segments are relocated simultaneously.

This system of rule notation has proved adequate to express the more-than-130 rules of the Spanish model. It has also supported a model to derive Shawnee from Proto-Algonquian, consisting of 21 rules (bin Muzaffar 1996 and 1997). Its main disadvantage arises from its unconventionality: it constitutes a considerable learning task for the new user. One challenge for future versions of Phono will be to alleviate the unfamiliarity of this rule-notation system. In this regard it remains to be seen whether it will be possible to devise algorithms for translating bi-directionally between the present if/then notation and something more akin to the standard "A > B / C _ D" rule format. If the standard notation proves not to be possible for all rules, then at least the writing of rules in the if/then notation can perhaps be facilitated by providing some automated guidance in the form of a structured questionnaire; and, conversely, the reading of this notation can be facilitated by means of a "prosifier", as was done in Version 3.2, which translates individual lines of rule notation into *if*- or *then*-clauses in English.

6. Deviations from linear rule order.

Although diachronic changes seem archetypally to follow one another in a constant order, the realism of the model can be enhanced by providing the possibility of exceptions to simple linear order of rules. The present version has the capability to label any rule as "persistent" and to temporarily "mask" other rules. Worth considering for incorporation in future versions would be the possibility of marking data words, either for exemption from specified rules, or for reversals of rule order.

6.1. Persistent rules.

In spite of the goal of remaining theory-neutral insofar as possible, I have adopted from the framework of generative phonology the notion of "persistent" rules. Sound-change rules are generally considered "transient", meaning that they act at one specific time in the chronology and never again. But experience with the Spanish model has supported the notion that some rules repeat their changes from the moment they are acquired for the remainder of the derivation, whenever their conditions occur. For this kind of rule, the term "persistent" was proposed and defined by Chafe (1968:131). Specifically, for example, the rule of nasal assimilation (to a following consonant) is apparently active at all moments throughout the history of Spanish, not only in Latin words (as partially reflected in the spellings *-mp-* and *-mb-*), but also centuries later in consonant clusters brought together by the deletion of vowels. For this and other such recurring rules, Phono provides the possibility to mark any diachronic rule as persistent. As the program traverses the diachronic list of rules in the derivation, it records the identities of the persistent rules in a separate list; then, after each subsequent rule that changes the word, it re-traverses the entire list of persistent rules.

6.2. Rule masking.

Phono's newest version provides the capability to temporarily disable, or "mask" any individual rule in a model, in order to observe how the output differs. In the case of the Spanish model, this tactic has helped to reveal that the form of some apparently exceptional words—those traditionally labeled "semi-learnèd", with implied attribution to the influence of written language—is due to their failure to undergo just one regular sound change. Being exempted from a single rule can set a word on a much different path from that of other words that previously were similar to it (see Hartman 1986).

6.3. Marking words for variable order?

Some investigators—beginning with Wang 1969—have suggested that sound changes may spread through the vocabulary gradually (by "lexical diffusion"), and that some changes may lose their momentum and cease to act, leaving part of the vocabulary unaffected. Whether for this reason or for the implied written influence on the semi-learnèd words noted above, it may be useful for sound-change modelers to have the capability, not merely to mask a rule temporarily, but rather to mark individual words permanently for exemption from specific rules.

Additionally, lexical diffusion—by acknowledging that the same change may affect different words in different historical epochs—opens the possibility of rules A and B affecting some words in AB order and others in BA order. If it is found that large numbers of words require variable order of application for certain pairs of rules, then it will be necessary in future versions of sound-change modelers to provide for marking data words individually for exceptional rule orders.

7. Batch testing.

One of the reasons for modeling sound change computationally is to observe it as a system, rather than just one or two rules at a time. And yet, during the development of a model, the experimenter may alter the model to accommodate one set of words only to find—eventually—that the alterations have ruined the accuracy of the model with regard to some other set of words. In order to monitor the integrity of the model throughout its development, it is important to test it at each step against a large vocabulary of pairs of words—etymon and known reflex—in order to insure that the derived results continue to coincide with the known reflexes. For this purpose, Phono is equipped with the capability to run in a "batch" mode, drawing the etymon/reflex pairs from the data file of the model. In the batch mode, the *etymon* words are encoded—translated from character strings to feature values—in the same way as for the interactive mode. Additionally, the input of known *reflexes* is, like etymon input, originally formed as strings of keyboard characters, and needs to be translated to feature values for comparison (the batch mode compares feature values, not character strings). So, analogously with the apparatus for etymon input, the reflex input for the batch mode is expressed in its own alphabet, and it may be adjusted by a set of "reflex adjustment rules" similar in function to the adjustment rules for etymon input.

In summary, the complete model includes three alphabets (keyboard alphabets for etymon input and for reflex input, and the phonetic alphabet for the output display); three rule sequencers (for etymon input adjustment, for reflex input adjustment, and for diachronic change); and one set of rules (a common "supply", upon which each of the three sequencers may draw).

8. Trace procedures.

Version 3 of Phono had the capability to run "rule trace" and "word trace" procedures based on the outcome of the batch test, and these will be incorporated into Version 4 in the near future. The rule trace keeps, for each rule, a record of all the words in which the rule effects a change. Such a record can be useful for measuring the value of each rule, insofar as this is based on its "functional load". Conversely, the word trace summarizes the history of each word as the list of rules that affect the word.

9. Limitations and future development.

The flexibility of Phono's rule-notation system is a double-edged sword. Its advantage is that it probably can portray virtually all sound changes of all natural languages (although it has yet to be tested with tone languages, vowel-harmony phenomena, or languages with discontinuous morphemes such as those of the Semitic family). On the other hand, a potential danger in such versatility is that the notation imposes no limits on the kinds or the complexity of rules that can be written: the investigator must still use human intuition to decide which rules are "natural" or plausible enough to be considered valid, and it is still a human decision how few example words (vs. how many counterexamples) are necessary to justify the formulation of a rule. Hopefully Phono and other programs like it can provide a basis for making these human judgments with more consistency and confidence.

References

Becker, Donald A. 1996. "Historical Linguistics as a Hacker's Paradise: Review of Phono 3.2". *Glot International*, 2:22.

Burton-Hunter, Sarah K. 1976. "Romance Etymology: A Computerized Model". *Computers and the Humanities*, 10:217-220.

Chafe, Wallace. 1968. "The Ordering of Phonological Rules". *International Journal of American Linguistics*, 34:115-136.

Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper & Row.

Hartman, Lee. 2003a. Phono: Historical Sound Change Modeler. Version 4.0-VB. Downloadable from http://mypage.siu.edu/lhartman.

Hartman, Lee. 2003b. "Phono (Version 4.0): Software for Modeling Regular Historical Sound Change". In Leonel Ruiz Miyares, Celia E. Álvarez Moreno, and María Rosa Álvarez Silva (eds.), *Actas: VIII Simposio Internacional de Comunicación Social: Santiago de Cuba, 20-24 de Enero del 2003*, I, 606-609.

Hartman, Steven Lee. 1974. "An Outline of Spanish Historical Phonology". *Papers in Linguistics*, 7:123-191.

_____. 1981. "A Universal Alphabet for Experiments in Comparative Phonology". *Computers and the Humanities*, 15:75-82.

_____. 1986. "Learnèd Words, Popular Words, and 'First Offenders'". In Oswaldo Jaeggli and Carmen Silva-Corvalán (eds.), *Studies in Romance Linguistics* (Dordrecht: Foris), pp. 87-98.

_____. 1993a. "Three Problems of Notation in Modeling Sound Change". Paper presented at Round Table on Computer Applications in Historical Linguistics, Brussels, Belgium, December 8.

_____. 1993b. "Writing Rules for a Computer Model of Sound Change". In *Southern Illinois Working Papers in Linguistics and Language Teaching*, 2:31-39.

Muzaffar, Towhid bin. 1996b. "Computer Simulation of Shawnee Historical Phonology". In *Canadian Linguistic Association Annual Conference Proceedings* (Calgary: Calgary Working Papers in Linguistics), pp. 293-303.

_____. 1997. "Computer Simulation of Shawnee Historical Phonology". M.A. thesis, Memorial University of Newfoundland.

Otero, Carlos-Peregrín. 1971. *Evolución y revolución en romance.* Barcelona: Seix Barral.

Wang, William S-Y. 1969. "Competing Changes as a Cause of Residue". *Language*, 45:9-25.