Efficient coding in speech sounds: Cultural evolution and the emergence of structure in artificial languages
Verhoef, T.

Citation for published version (APA):
Verhoef, T. (2013). Efficient coding in speech sounds: Cultural evolution and the emergence of structure in artificial languages

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Discussion

The studies described in this thesis have been conducted with the aim to contribute to an understanding of the evolution of speech. This research fits within a framework in which language is considered to be a complex adaptive system (Brighton and Kirby, 2001; Kirby, 2002; Steels, 1997b) which is shaped through complex interactions between different levels of organisation including that of the individual speaker and of the language system as a cultural phenomenon. The focus of the studies presented is on the emergence of combinatorial structure in speech sounds. Two different methods were used in the current research: computer simulations and laboratory experiments with human participants. In chapter 3 a first attempt at experimentally investigating the emergence of combinatorial structure in sound systems was described. In that experiment we could not observe the emergence of combinatorial structure due to limitations with the design of the sound-production interface that appeared to be very difficult to use. In order to deal with the issues that characterised this first experiment, the study described in chapter 4 was conducted. Here, participants produced sounds with a more intuitive device, namely a slide whistle. The improved experiment resulted in a clear emergence of combinatorial structure in artificial whistled languages, showing that this type of structure can emerge through cultural transmission. In chapter 5 a more elaborate analysis of the combinatorial structure in the emerged whistled languages from chapter 4 was presented. By means of additional experiments it was shown that human participants are aware of the regularities in the emerged systems and they are able to use it to distinguish between two different languages. Chapter 6 provided a description of a follow-up experiment involving artificial whistled languages in which the words referred to meanings. Also in this case, combinatorial structure at the level of the signal emerged reliably. Finally, complementing all these studies about emergence, adaptation and change, chapter 7 described a

This chapter contains parts that also appear in the following articles:
computational study in which it was shown how cultural phenomena can influence the preservation of structure over generations.

In chapter 1 the sharp contrast between different ideas that relate to the evolution of language was outlined. There are theories that assume language is innate in the form of a language-specific module unique to humans (Chomsky, 1976; Piattelli-Palmarini, 1989; Pinker and Bloom, 1990), which sometimes is assumed to have evolved in human biology through natural selection (Pinker and Bloom, 1990). Pinker and Bloom (1990) mention two criteria from evolutionary theory in support of their idea of language as a biological adaptation that has undergone natural selection. These criteria define when natural selection can be used as a valid explanation for a phenomenon: “complex design for some function, and the absence of alternative processes capable of explaining such complexity”. As we have seen, we currently are in fact aware of alternative processes that can account for the emergence of complex linguistic structures. The process of cultural evolution, in which languages are transmitted from one generation of naïve learners to the next, has been proven to be able to account for “the appearance of design without a designer” as well as biological evolution (Kirby et al., 2008). It has been proposed that language change needs an ‘invisible hand explanation’ (Keller, 1994), a term that originated from the field of economics and was introduced by Adam Smith. This is based on the idea that sometimes individuals, who are not directed to work towards a central goal, behave in a certain way ‘as if they were guided by an invisible hand’, which results in unintended structure at the level of the community. Language can be seen as one such structure as both Keller (1994) and Fitch (2007) pointed out. This view nicely illustrates the importance of considering both the individual micro level and the population macro level in language evolution. The view of cultural transmission as a key process in the way languages are shaped complies with this.

8.1 Main findings

The results presented in this thesis provide additional support for the idea that the process of cultural evolution is important in shaping linguistic structure, by showing that aspects of phonological organisation can emerge as the result of cultural transmission, as opposed to the assumption that humans are born with a set of innate phonetic features as Chomsky (1976) proposed: “languages have a partially determinate structure as a matter of biological necessity, much as the general character of bodily organs is fixed for the species. The theory of distinctive features is perhaps the most familiar case. It has been proposed that a certain set of features is available in principle for phonetic representation; each language must make its selection from among these. (...) it seems to me not unreasonable to approach
8.1. Main findings

the study of language as we would the study of some organ of the body” (Chomsky, 1976, p.46). The results of chapter 4 and chapter 6 demonstrate that cultural evolution can cause a system of random continuous signals to become organised in a way that is very similar to how speech is organised: a small number of basic elements is combined into a larger number of signals, resulting in systems that are more constrained and show transmission-chain specific ‘traditions’. The emergence of combinatorial structure in the presented experiments seems to be due to general properties of the way humans learn and generalise signals, which drives the systems to become more predictable and more learnable. The structure in the artificial whistled languages cumulatively developed in adaption to the process of being transmitted from participant to participant, without any influence of communication and without the invention of structure by individual participants. It therefore seems unnecessary to assume that phonetic features are innate.

With respect to current theories in evolutionary phonology, it has been proposed that the emergence of combinatorial structure was driven by vocabulary expansion and dispersion: the limits of the signal space were reached at some point and no more distinguishable holistic signals could be added. Similar ideas have been proposed by Abler (1989); Studdert-Kennedy and Goldstein (2003). However, as mentioned in chapter 2, there are reasons to believe this is not the complete picture, following for instance from the example of ABSL with its high functionality but still emerging combinatorial structure (Sandler et al., 2011). Also, other mechanisms have been proposed in reaction to the fact that dispersion theories do not explain consonant inventories well. Ohala (1980) for instance suggested that the organisation in speech sounds seems to follow a principle of “Maximal use of available distinctive features”. This theory, as well as others that are based on similar ideas (Clements, 2003; Maddieson, 1995), focuses on principles of economy and the efficient reuse of basic elements. The experimental data presented in this thesis seems to be more in line with economy principles, showing that combinatorial structure emerged in sets of whistles that were culturally transmitted in absence of pressures from vocabulary growth. The vocabularies in the experiments contained only twelve whistles and in all conditions combinatorial structure emerged long before the signal space had been fully covered. Apparently, a good reason to have combinatorial structure, even for a very simple system, is that a system with such structure is easier to learn and reproduce.

In line with earlier findings on the dynamics of iterated learning (Kirby et al., 2008; Kirby and Hurford, 2002), the whistles that fit the structure and conform to people’s cognitive biases are more likely to be preserved from generation to generation in cultural evolution. Combinatorial structure therefore potentially emerged within a gradual cultural evolutionary process. As follows from the results presented in chapter 5, different parallel chains result in whistle languages that are recognisably different in
terms of the specific rules, building blocks and constraints. This further supports the view that the emerging structure in the artificial languages is the result of conventionalisation and emerges through cultural transmission. In the following sections some more general points of discussion are presented involving implications, limitations and future plans for extensions of the work described in this thesis.

8.2 The protolanguage debate

As mentioned in chapter 2, there is an ongoing debate on the nature of a possible ancestral protolanguage. Did human protolanguage consist of holistic utterances (in the form of produced sounds or gestures) that were segmented into words (Arbib, 2005; Wray, 1998) or did it start with simple words that were combined into more complex structures (Bickerton, 1992; Tallerman, 2007)?

Arguing against holistic protolanguage, Tallerman (2007) suggested why holistic protolanguage would be unlikely to lead to compositional syntax. First of all, holistic protolanguage would by definition be irregular. In addition, if it is learned, it is expected to change rapidly over generations, just as modern language does. Tallerman therefore argues that this would not provide a sufficiently stable target for analysis into components. In response to this argument, it has been pointed out that learners often overgeneralise and that inconsistencies in the input do not necessarily prevent the discovery of regularities (Smith, 2008b; Fitch, 2010). This argument is backed up by computer simulations showing that regularisation can indeed happen in cultural transmission (Brighton and Kirby, 2001; Kirby and Hurford, 2002; Kirby et al., 2004), as well as by cultural learning experiments with humans that show emergence of compositional structure from initially holistic sets of utterances (Kirby et al., 2008).

Tallerman's (2007) response to these arguments and to the computational models is that they already assume that there are building blocks: the target words are built up out of a set of discrete segments. She assumes that the task of segmenting signals into relevant elements is impossible if there are no predefined segments, because in that case even the tiniest distinctions between signals should be considered potentially significant. She therefore does not find demonstrations that rely on pre-existing segments convincing.

The experimental results that were presented in chapter 4 and chapter 6 address the above-mentioned concerns. The results show that modern human learners quickly generate structure in initially structureless (holistic) sets of continuous signals. Because the initial set is too difficult to learn precisely, learners tend to overgeneralise the structure they (think they) observe. This introduces reuse of a small set of building blocks and increases the learnability of the set of signals.
8.2. The protolanguage debate

Apparently, modern humans have no problems finding (apparent) structure in holistic, continuous utterances. Both experiments, with and without meaning, show that participants introduce combinatorial structure very rapidly and as a system independent of structure in the meaning space. Subsequently, as Kirby et al. (2008) have shown, signals built up of discrete elements with an associated meaning (but without a systematic form-meaning mapping) can transform into systems with a systematic form-meaning mapping.

As for the argument about rapid change: rapid change does happen in the experiments, but it leads to structure and better learnability. This is an example of how repeated introduction of naïve learners with acquisition limitations drives the development of a linguistic system towards being learnable (Zuidema, 2003). Therefore the argument that the changeability of a culturally transmitted holistic system would prevent emergence of structure is not supported by empirical evidence from modern human behaviour.

Of course, these observations do not necessarily generalise to ancestral hominids, who may have had very different cognitive adaptations. However, as Smith (2008b) has pointed out, research with cotton-top tamarins (Hauser et al., 2001) has shown that at least some non-human primates already have simple abilities for segmenting streams of speech. It is therefore possible that the ability to find regularities in speech is much older than the split between humans and the other apes. Given a pre-existing ability to analyse, we can expect that re-use of regularities could have been possible at the earliest stages of protolanguage.

Although the current findings refute arguments against holistic protolanguage, they may also imply that the idea of an extended holistic protolanguage phase is unlikely. It appears that such a system would perhaps not be stable for a very long time, because combinatorial structure at the signal level would emerge rapidly. I would also not exclude the possibility that the holophrases were concatenated into larger constructs simultaneously, following a synthetic scenario in parallel. In fact, it would be possible to have aspects of both holistic and synthetic protolanguage in one system: holistic phrases break up into smaller units, while at the same time words could be combined into short utterances. As Smith (2008a) suggested, there is no fundamental contradiction between these two points of view.

The main aim of this thesis was not to unravel the nature of protolanguage and the above mentioned ideas are still speculative and preliminary. However, it is exciting to note that iterated learning experiments can be used to empirically investigate some issues that are alive in the debate on holistic and synthetic protolanguage. It is therefore no longer necessary to base protolanguage theories solely on conjectures, because relevant data can be obtained even with modern humans.
8. Discussion

8.3 Cultural transmission and efficient coding

As reviewed in chapter 2, the influence of cultural evolution on the way linguistic structure emerges is increasingly being studied with the use of laboratory experiments. This method has generated a vast amount of data in the past few years, including the results presented in this thesis. The emergence of compressible and predictable systems appears to be a prevalent result of cultural transmission experiments and the results in this thesis are no exception. The emergence of combinatorial structure in the sets of whistles forms an additional example. In these systems, whether they include meanings or not, discrete sets of basic building blocks could be identified in the sounds and these were reused and combined in a predictable way. Quantitatively, a cumulative decrease of entropy over the reuse of basic elements could be measured in the languages, indicating that equally large languages could be described using fewer basic elements. The whistled systems therefore became more constrained and more compressible.

In chapter 2 a short overview was presented of advances in the field of computational neuroscience in which principles of compression and simplicity in neural processing are studied. As we have seen, it can be proven that many features of natural stimuli are optimally efficiently encoded in both the visual and auditory cortices. The study by Smith and Lewicki (2006) in particular is interesting in the context of language evolution, because here it was shown that the auditory cortex of a cat optimally encodes the sounds of speech. This provides convincing evidence in favour of the view that the sounds used in language are adapted to the (mammalian) auditory cortex. This is in line with the suggestion that transmitted systems adapt to human biases and constraints (Christiansen and Chater, 2008; Deacon, 1997; Griffiths and Kalish, 2007; Kirby and Hurford, 2002). It is unlikely that cat auditory processing has evolved to efficiently encode human speech, therefore a more plausible assumption would be that the sounds used in speech have adapted to be efficiently coded by the brain. Likewise, it is expected that linguistic structure at other levels of organisation has adapted to general cognitive ‘simplicity’ biases and is shaped in such a way that it is compressible. The study by Smith and Lewicki (2006) provides an exciting example of more direct evidence of adaptation through cultural evolution. Even though this has so far only been shown for very early processing and sound primitives for speech, it is a promising avenue for further research. Following this direction we should try to formulate experiments and create biologically plausible models that can provide this kind of evidence for other levels of organisation in linguistic structure as well.

As Deacon (2009; 1997) argues, researchers have not been able to associate human linguistic behaviour with a unique change or difference in brain anatomy as compared to non-human ancestors. Instead it is likely that a large variety of systems, with perhaps different
functions in our ancestors, contributed to and are involved in modern human linguistic behaviour. The study by Smith and Lewicki (2006) is a brilliant example of how such a homologous system (involving auditory processing in this case) can be linked to efficient coding of speech sounds in non-human species. There may be other aspects of language processing and learning for which it is possible to demonstrate preferences or efficient coding in homologous systems inside non-human mammalian brains. The method of demonstrating such efficiency by predicting properties of measurable brain responses through computational modelling of optimally efficient coding is a path that deserves exploring. Especially in the case that we can show this effect for cognitive processing of (linguistic) compositional and combinatorial structure, this would be compelling evidence against language-specific biological adaptations and must indicate a strong influence of general cognition and cultural evolution. In addition, this may be a direction that can potentially reveal relevant differences between human and non-human processing. Perhaps it is therefore fruitful to consider an integrative framework combining the study of cultural transmission, the systems that emerge from it and the neuroscientific study of efficient coding in the brain.

8.4 Possible concerns

A possible concern with the current results, if we were to consider this experiment as a reconstruction of language evolution, could be that we use modern human participants who obviously have modern cognitive adaptations unlike our ancestors. This fact is shared among all language evolution experiments that make use of human participants, but should not necessarily be viewed as problematic. As Scott-Phillips and Kirby (2010) point out, the results of this type of work should not be interpreted as an attempt to reconstruct the emergence of linguistic structure, or, in this case of structure in speech sounds, but as a method to shed light on what mechanisms may be involved in this emergence. The current work is meant to illustrate how human cognitive biases influence a sound system when it is repeatedly transmitted to new learners and what role these biases play in the maintenance of combinatorial structure. In addition, when experiments such as the ones presented in this thesis are paired with results from computer simulations, a stronger point can be made. As reviewed in chapter 2, the iterated learning model has been studied with a variety of learning mechanisms (Kirby, 2002). In these simulations naïve agents are used that obviously do not have language built in and also do not have any experience with language prior to the model runs. Still, the results are very similar to what has been found with modern humans in the laboratory. The computer simulation that was presented in this thesis is not directly modelled after the specific experiments that were conducted for this research, but this is planned in the future continuation of this work.
Another concern that has been expressed in response to the experiment described in chapter 4 involves the lack of meanings conveyed by the whistled signals. The design of that early experiment abstracted away from full human semantic complexity by not having an explicit meaning connected to the whistles. As was pointed out in chapter 4, the systems do in a way acquire meaning when the experiment progresses because of the fact that participants have to reproduce a complete set of twelve whistles. The languages therefore have some degree of expressivity. As an adaptation to the learning constraint, the whistles evolve in a way that makes them share more and more features or building blocks. This makes it possible to remember the signals as subsets, which makes learning and recall easier. The idea that chunking of information in this way facilitates encoding more information in short-term memory is well established (Miller, 1956). Participants tended to categorise the whistles as subsets, such as ‘the ones that all start with a falling slide’. This first investigation of combinatorial structure in a set of whistles without referents was necessary to be able to control for effects of semantics such as iconicity or compositional structure. With such influences present it would be harder to distinguish whether the emerging structure relates to the structure of the meaning space or whether they are truly meaningless units being recombined. In addition it would be harder to know what drove the emergence of structure. Chapter 6 presents experiments in which meaning was added and the analysis of the emerging whistled languages indeed proved to be non-trivial. However, together the two experiments have already provided interesting new data for studies about the emergence of phonology. The results conform to the idea that phonology is an autonomous system with generative power (Studdert-Kennedy and Goldstein, 2003) and they also show that combinatorial structure is not necessarily linked directly to vocabulary size or driven by signal distinctiveness.

It could be argued that the first whistle experiment has little to do with language and should instead perhaps be compared with musical systems because of its lack of meaning. I would like to stress that in my opinion, the observation that the systems emerging in the whistle experiments show characteristics of musical structure (whether this is actually the case or not) should not at all be a reason to conclude it is uninteresting for research on language. In fact, I would argue that the experiment would have been equally successful and equally relevant for theories on the emergence of phonology if the conclusion had been that a discrete set of basic notes or rhythmic primitives had emerged that were combined into different musical styles in the four chains. As Fitch (2006, 2010) argues, music and (spoken) language share many structural characteristics, especially at the level of phonology. Music and phonology are both built from a discrete set of basic meaningless primitives that are combined in a generative way to construct an unlimited number of signals and both are also culturally transmitted (Fitch, 2006; 2010). An important difference is the role of meanings,
which is much more prominent in language and part of a systematic organisation. In music the form has a more holistic and affective relation to meaning and subtle differences in for instance expressive timing (Honing, 2002) may lead to different interpretations. Another point of difference is the nature of the discrete elements. In music, features such as pitch and rhythm or other temporal features are the elements that are discretised, constrained and regularised, while these features tend to be more variable in speech (except for pitch in tonal languages). In speech, the most important elements of recombination tend to be vowels and consonants, which are not normally associated with music. However, in terms of their combinatorial structure and the way this structure forms part of a transmitted cultural tradition, music and phonology seem to be very parallel (Fitch, 2006; 2010). This has also become clear from data on experiments that explicitly compare the two domains with neuroimaging and show there is significant overlap in the processing of language and music (Patel, 2003; 2012).

To further illustrate how subtle the boundaries are between music and phonology, perhaps Pirahã provides an interesting example of a language where phonological contrasts could perhaps be considered to involve music-like as well as language-like features. As Everett described for Pirahã, vowels and consonants seem to play a less important role in this language than patterns of tone, timing and stress (Everett, 1985). This is probably related to the fact that this language is used over several different channels, one of which is hummed speech. Humming is often used there in intimate, close-contact situations such as mother-child interaction. Children also apparently acquire control over the prosodic structure of the language earlier than the specific vowels and consonants (Everett, 1985). In absence of the knowledge that a rich, unrestricted repertoire of meanings can be conveyed with the hummed speech system, perhaps an outside observer may classify it as music based solely on its form.

In summary, sometimes it can be informative to abstract away from the full complexity of language and focus on a specific structural property to learn more about the way it can emerge. The fact that this particular type of structure is shared with other systems such as music (and many other culturally transmitted systems such as dance and art I would say) should not lead to the conclusion that the results are less interesting for language. Instead, as Fitch (2010) also proposed, the parallels between different domains should be used to our advantage in research about the evolution of language and other systems.

8.5 Plans for the future

The studies presented in this thesis together represent an exploratory investigation which involved the adaptation of existing methods that were recently introduced into the field to make them applicable to
questions on the evolution of speech. As reflected in the results that were presented in the different chapters, many issues were resolved and interesting insights obtained. However, many questions remain unanswered or require further investigations. In this section I summarise some of my plans for the future continuation of this work.

8.5.1 Experimental designs

One limitation of the iterated learning experiments described in this thesis is that the generations consisted of only one participant. A first problem with this is that it is obviously not realistic, because real speech communities necessarily consist of more than one speaker. Another problem is that this design ignores the importance of interaction and communication. Participants learn a language, but they do not use it for communication with others. This has been a very important abstraction to be able to demonstrate that linguistic structure can emerge as an adaptation to a transmission bottleneck, independent of communication and without the conscious creation of structure by individuals. It was therefore sensible and necessary to start this way. However, the lack of a pressure for communication is probably the reason why expressivity needs to be maintained in an artificial way in experiments with this design, such as the filtering technique used by Kirby et al. (2008) and my reproduction constraint. Changing the design therefore seems desirable.

Other designs for iterated learning experiments have been explored already, although these have not been applied to the study of structure in continuous auditory signals yet. Tamariz et al. (2012) for instance created transmission chains in which two participants formed each generation. They showed that, when these dyads interacted with each other and could negotiate to arrive at new versions of the artificial language together, the structure increased more than without such interaction. Tan and Fay (2011) also showed that interaction improves faithful transmission in chains where participants had to communicate a description of an event to another person. In addition, the iterated learning paradigm has now been extended by Caldwell and Smith (2012) to involve microsocieties. Here, groups of four participants communicate about meanings with drawings and there is a gradual progression from generation to generation in which the most experienced participant is replaced by a new participant. These designs allow to carefully investigate effects of both conventionalisation and cultural transmission on the way signalling systems are shaped and I expect that these innovations will become more widely used in the future.

The work in this thesis had a strong focus on speech, and therefore on sound systems and structure therein. It would however be fruitful to extend the work to other modalities as well, since combinatorial structure also clearly plays a role in for instance sign languages. del Giudice et al. (2010) used a different modality by studying the emergence of combinatorial structure in transmitted graphical systems.
A study on manual systems seems to be a promising addition to these existing studies. Such an experiment would study with non-signing participants whether and under what conditions a random set of gestures evolves into a system that shows regularities similar to those found in sign languages. This approach would address the role of embodiment in producing signals. Another advantage of this approach would be that it allows a more direct comparison of experimental results with data from Al-Sayyid Bedouin Sign Language, the only known language in which regularised combinatorial structure is still emerging. In addition, given the direct visual-to-visual mapping, sign languages are more conducive to iconic expressions (Perniss et al., 2010). The manual modality may therefore be more suitable and would perhaps allow for a more natural investigation into questions on iconicity.

### 8.5.2 Neuroscience-inspired computer model

Experimental work in language evolution is often modelled after designs and findings that were obtained with the use of computer simulations. Scott-Phillips and Kirby (2010) review examples of this. The two methods nicely complement each other. Human participants in an experiment are unavoidably modern humans who may not have the same cognitive abilities as our ancestors and may be biased by their linguistic experience. Biases of computer learners can be controlled, but computer models have been criticised to be less realistic. Computer models provide a possibility for taking a bottom-up approach to explore the cognitive biases needed to explain behaviours found in the laboratory. However, no simulation exists that is directly comparable to the iterated learning experiments presented in this thesis. The computer model presented in chapter 7 is less suitable for a direct comparison to the experimental findings because of the focus on a different population structure and sound system. The next step would therefore involve the design of a computational model that may be able to explain the observed patterns in iterated learning experiments with continuous signals. To be able to explain the emergence of efficient and compressible representations in languages, perhaps a link should be formed with neuroscience-inspired models on efficient coding strategies in the human brain, as reviewed in chapter 2 and mentioned above.

### 8.5.3 Conclusion

The possible follow-up studies proposed above cover only a small subset of all the possible open questions that deserve further investigation. With the work in this thesis I have merely set the first steps towards developing a paradigm that studies evolutionary phonology empirically. We now know that the same methods that were previously used successfully to study aspects of language that can be represented with discrete symbols such as compositional structure, are useful for studying combinatorial structure in sound systems as well. I expect that many more insights
can be gained in the future with an approach in which computational simulations and laboratory experiments are used hand in hand to unravel the origins of structure in linguistic systems of continuous signals.