ONOMÁZEIN
Journal of linguistics, philology and translation

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE LETRAS

Modality effects in the cultural evolution of language: An experimental iterated learning approach

Fernanda Weinstein Perelman
Universidad del Desarrollo
Chile

ONOMÁZEIN 45 (September 2019): 103-125
DOI: 10.7764/onomazein.45.05
ISSN: 0718-5758

Fernanda Weinstein Perelman: Centro de Investigación en Complejidad Social, Universidad del Desarrollo, Chile. | E-mail: rfweinsteinp@udd.cl

Received: May 2017
Accepted: May 2018

Abstract

Cultural evolution has been proposed as the mechanism by which human languages' distinct features emerge. One of such features is structure, which is regarded as an optimal solution to the competing pressures for simplicity and expressivity in language learning and use. A recent experimental iterated learning study (Kirby et al., 2015) shows that structure can emerge from an unstructured language under these competing pressures, by implementing both a learning and a communication task in a transmission chain setup. However, as most iterated learning experiments, it was run on a written modality, which might be problematic if the aim is to drive conclusions about language in general—writing is not language's default modality and was not present in early stages of language evolution.

The present study carried out a partial replication of the aforementioned experiment, contrasting a written condition (analogous to the original) with a spoken condition, in order to test for possible modality effects. Results for the written condition did not replicate those in Kirby et al. (2015) in any of the measures, suggesting that motivational factors could have played a crucial role in the previous findings. This hinders the interpretations of modality effects and suggests the need of further work.

Keywords: cultural evolution; iterated learning; modality effects.

1. Introduction

1.1. Background

Human languages exhibit a series of features which distinguish them from other animal communication systems, and which appear to be well designed (Hockett, 1960). One of such is structure: on the one hand, our languages have *combinatorial structure*, which consists in the reuse and recombination of meaningless sounds to form meaningful units which in turn are reused and recombined (Kirby et al., 2015). On the other hand, they display *compositional structure*, which means that the meaning of complex signals is derived from the meaning of their parts (Kirby et al., 2015). Both kinds of structure make possible to encode languages in a way that is shorter than a simple list of every possible utterance and also to convey unlimited meanings using only a limited set of signals. This encodings are usually called grammars.

But, where does structure come from? If we discard the hypothesis of a "conscious designer", there must be a mechanism by which language's distinct features emerge along the process of evolution (Kirby, 1999). One possible explanation is cultural evolution (Mesoudi, 2016). In an influential paper, Christiansen and Chater (2008) propose that language, rather than the product of Darwinian natural selection (as suggested by Pinker and Bloom, 1990) or a byproduct of biological evolution (e.g. Chomsky, 1972, cited by Pinker and Bloom, 1990), is an evolutionary system itself, and its distinctive features emerge in response to constraints imposed by the human brain. Regarding structure, it is suggested that these constraints correspond to two competing biases: *simplicity*, or the tendency of cognitive systems to organise data in the shortest possible way (Chater and Vitányi, 2003); and *expressivity*, or what enables to discriminate between an intended meaning and alternative meanings in a given context (Kirby et al., 2015).

Can we prove this? We have no access to our (pre-literate) ancestors' communications systems, so we cannot possibly track a hypothetical moment of phylogeny in which its distinct features' appeared. However, language evolutionists do test their hypotheses using laboratory experiments and computer simulations[1]. One experimental and modelling paradigm which recreates the process of cultural evolution is iterated learning. Iterated learning is defined as the process by which individuals learn behaviours by observing other individuals who learned them in the same way (Kirby et al., 2008). The setup for experimental iterated learning implements a transmission chain method (Mesoudi and Whiten, 2008), where participants are

1 For the case of the trade-off between simplicity and expressivity, there is also cross-linguistic evidence on category systems (cf. Kemp and Regier, 2012).

organised into generations and information is passed from person to person like in the "telephone" game. The first participant (generation) in the chain is exposed to a certain behaviour, which they are asked to learn and recall, and their output is used as training behaviour for the next generation, and so on (for modelling approaches, see Kirby, 2000; Smith and Kirby, 2008).

In a recent experimental and modelling study, Kirby et al. (2015) showed that compositional structure could emerge through iterated learning from an initial *holistic language*[2]: that is, a language that does not allow a simpler representation than a list of every possible utterance (Kirby et al., 2015). In the experimental part, they did so by passing holistic languages through transmission chains of human participants and putting those participants to play a communication game, where the goal was to produce signals that could be associated with a particular meaning. In their setup, each generation consisted of a pair of participants (rather than an individual, as in Kirby et al., 2008) who were trained in the same input language and were asked to recall it when playing the game. Supported by their results, the authors propose that the mechanism by which the aforementioned pressures for simplicity (or compressibility) and expressivity operate: the former arises from learning, the latter does so from communication, and the interplay between both processes is involved in the emergence of structure. Furthermore, they show that learning without communication produces *degenerate languages*: that is, languages containing only one word to refer every possible meaning (as defined by Kirby et al., 2008). Conversely, communication without learning produces holistic languages (closed-group condition in Kirby et al., 2015).

1.2. Written language bias

Kirby et al. (2015) and other iterated learning studies rely on a written modality[3]. This may be problematic if the aim is to derive conclusions about language in general. After all, writing is not language's default modality, it appears much later in history, it is not present in all modern languages, it requires formal instruction, it is used only by a subset of speakers, etc. Several authors have pointed out the existence of a written language bias in linguistics (Olson, 1996; Coulmas, 2003; Linell, 2004), which arises from a conception of language guided by alphabetic writing (Linell, 2004). For instance, the fact that many speakers who are illiter-

2 This idea builds upon the holistic protolanguage hypothesis. For instance, Wray (1998) proposes that in early stages of human evolution, our ancestors would have used holistic, arbitrary signals to refer complex meanings. These signals would have shared sequences of sounds just by chance, and over time humans would have developed the capacity to segment these sequences and associate them with the shared meanings, constituting the foundations of grammar (Wray, 1998).

3 There are other iterated learning experiments that consider other modalities: for example, Verhoef et al. (2014) used slide whistles, while Theisen-White et al. (2011) used graphical communication.

ate in alphabetic systems cannot separate words into phonological segments might suggest that segment-based phonology is not a good representation of language as a mental system (Faber, 1992, cited by Coulmas, 2003).

It can be argued that the experiment in Kirby et al. (2015) exhibits a written language bias in at least two dimensions. Firstly, participants were all literate in an alphabetic system, and this matters in the extent to which results are generalizable to a larger sample of speakers and languages. Evidence points out that phonological awareness differs from literates to illiterates (Morais et al., 1979; Reis and Castro-Caldas, 1997) and from alphabetic to ideographic literates (Read et al., 1986). This implies that phonological awareness, a skill that according to the holistic protolanguage hypothesis was developed in early stages of language evolution (Wray, 1998: 55-58), may be a consequence of literacy rather than a prerequisite. Either way, knowing which skills are particular of alphabetic literates might provide an insight on which skills are unlikely to be responsible for the emergence of language's distinct features, common to all languages (written and non-written). These kinds of differences should also be considered in evaluating the plausibility of the holistic protolanguage hypothesis in the first place (see e.g. Tallerman, 2007).

The second dimension is the use of the written modality. Writing was not present in early stages of language evolution, particularly in the transition from protolanguage to language, so it may not be the best modality to test evolutionary claims. However, it is also true that it is ubiquitous nowadays, and it can be hypothesised that the fact that technology allows us to use written modality in situations that used to be exclusive to spoken (or gestural) modality might be changing evolutionary pressures acting on language. Hypothetically, if writing becomes the default modality of language, word inventories will no longer be limited by human articulatory physiology. As well, if writing is the default modality, then visual, rather than auditory language processing skills will be the more required in a daily use, and therefore the ones acting as pressures from language learning. So, seeking for modality effects may allow us to make some predictions on how languages might look like in this hypothetical future.

1.3. Modality effects

Evidence suggests that the skills for processing language in auditory and written modality are independent to some extent (for a review, see Chafe and Tannen, 1987). Caplan et al. (2016) tested basic psycholinguistic abilities for understanding simple words, word structure and sentence structure in both modalities, comparing developing readers to advanced readers. Results showed that written and spoken language rely on processing systems that are different, but related to each other; where the ability for low level language processing (e.g. word recognition) in the spoken modality predicts the correspondent abilities in the written modal-

ity to a greater extent than those for high level language processing (e.g. sentence processing) (Caplan et al., 2016). This independence is also reflected at neural level: fMRI studies reveal that verbal memory tasks using auditory or visual stimuli activate different neural circuits (Crottaz-Herbette et al., 2004).

Regarding language learning, Nelson et al. (2005) tested how fast adults learned new words when presented in either spoken or written modality, and how accurately they were able to recognise them. Results revealed that participants were both faster at learning and better at recalling those presented in a written modality. This was interpreted in terms of episodic memory (Baddeley et al., 2009): because participants know the letter-phoneme correspondences, they can generate a phonological representation of new words presented orthographically, and hence the new memory trace and posterior lexical entry includes both modalities. The opposite process, generating orthographic representations from phonology, is not as easy or automatic, which produces less specified, lower quality lexical entries (Nelson et al., 2005: 39).

Evidence from classroom-based second language acquisition studies reveals that writing is the preferred modality among literate students (Harklau, 2002; Weissberg, 2000). Because in school contexts children often begin to learn how to read and write before they are introduced to a second language, and because formal education in general emphasises the written modality (textbooks, notebooks, written quizzes and exams, etc.) it is natural for second language learners to rely on it (Harklau, 2002). As well, written modality can eliminate pressures of face-to-face communication in the sense that enables the students to interact in a self-paced, editable way (Harklau, 2002: 337). Writing can also be a good aide-memoire, which is in line with what is proposed in Nelson et al. (2005). And last but not least, it is somewhat evident that writing allows more time for memorising and understanding new words, since it is a slow fading medium compared to speech.

1.4. This study

The aim of the present study is to examine whether modality, spoken or written, has an effect in the cultural evolution of artificial languages within an iterated learning experiment run with literate, English-speaking participants. To test for this modality effect, I run a partial replication of the experiment in Kirby et al. (2015), replicating their chain condition, which was written (CW), and contrasting it with a spoken (also chain) condition (CS). Methods for CW attempted to reproduce those in Kirby et al. (2015) as closely as possible.

General predictions, related to the replication of the original experiment and to the consequences of cultural evolution, are the following: a) results in CW will replicate those of the chain condition in Kirby et al. (2015); b) communicative success (correct matches in the com-

munication game) will increase along generations in both conditions; c) transmission error (change from the training signal to the produced signal) will decrease along generations in both conditions; d) structure will increase along generations in both conditions, and final languages will be structured, as opposed to holistic or degenerate; e) alignment (similarity between the signals produced by both participants for every given meaning) will increase along generations in both conditions; and f) languages will maintain the number of unique words along generations in both conditions (they will not exhibit homonymy).

Regarding modality effects, from the evidence provided in the previous section I hypothesise that CS will have more pressure on the learnability side: languages will be harder to learn, because learning is slower and recall is poorer when new words are presented in a spoken modality. Additionally, phonological representations could be less precise due to differences in pronunciation (there was a wide range of dialects amongst the participants I tested), which also puts more pressure for learnability. Furthermore, I think that communication in CS will also be more difficult, for the same reasons (poorer recall and pronunciation differences), but, because the communication task is to produce signals that the other participant can relate to the correct meaning, pressure for expressivity is constant across conditions.

Hence, condition predictions are the following: a) communicative success will be lower and will increase faster in CS, since pressure for learnability is higher; b) transmission error will be higher and will decrease faster in CS, again because pressure for learnability is higher; c) structure will increase faster in CS, but final languages will be equally structured, because pressure for expressivity remains constant (while pressure for learnability is higher); d) alignment will be lower in CS (due to pronunciation differences); e) there will be no condition effect on the length of the signals, since, while the spoken condition has a noisier channel (cf. Shannon, 1948) and part of the signal could be lost, typing is more effortful than speaking (so there is also pressure for written signals to be shorter); and f) the number of unique words will be constant across conditions, since both have the same pressure for expressivity.

2. Methods

2.1. Laboratory and participants

Data collection took place at the laboratory of the Centre for Language Evolution, which is based at the School Psychology and Language Sciences, University of Edinburgh. The code for conducting the experiment was written in Python, based on previous codes developed at the Centre for Language Evolution. For this experiment, data were collected in June 2016 (teaching vacation period). For Kirby et al. (2015), according to the corresponding author, data collection took place in October and November 2010 (mid-semester period), and in August 2011 (teaching summer vacation period).

84 native-English speaking participants with no auditory problems and normal or corrected to normal vision were recruited through the University of Edinburgh careers service and were paid £7 for their participation (which lasted 40-60 minutes). The data produced by 2 participants was excluded due to an experimenter mistake, while 2 other participants could not complete the task due to technical problems. Amongst those 80 participants whose data was used (46 female, 34 male) the mean age was 24.93 years. They were organised into 8 chains (4 per condition), each consisting of 5 pairs. The Ethics Committee of LEL, University of Edinburgh, granted ethical clearance.

The subject population was similar to the one of the experiments in Kirby et al. (2015), which consisted of 60 students (41 female, 19 male) recruited through the University of Edinburgh Student and Graduate Employment service (mean age not reported) and paid at an hourly rate of £6.

2.2. Stimuli and initial languages

The participants were asked to learn a miniature "alien" language and then to use it to play a communication game with a partner. The language consisted in a set of 12 strings (signals) paired with a set of 12 pictures (meanings). The set of meanings was the same one used by Kirby et al. (2015), where each picture is a unique combination of 1 out of 3 possible shapes, 1 out of 4 possible textures and also one unique feature.

The set of signals for CW was generated in the same way as in Kirby et al. (2015) by randomly selecting combinations of 2, 3 or 4 syllables composed of randomly selected combinations of 8 consonants (g, h, k, l, m, n, p, w) and 5 vowels (a, e, i, o, u). A total of 4 languages were generated (see figure 1 for an example).

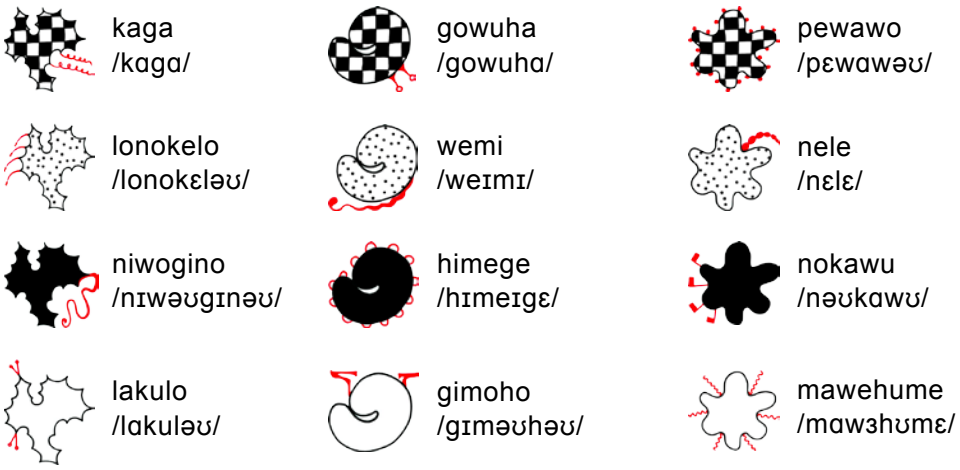
For CS, a native Standard Southern British English speaker recorded the strings of the 4 initial languages from CW in the same laboratory and using the same microphone as participants in the experiment. The recording process increased the vowel inventory from 5 to 11 (9 vowels and 2 diphthongs: /ɑ/, /ɜ/, /ɛ/, /i/, /ɪ/, /o/, /ʊ/, /u/, /ə/, /eɪ/ and /əʊ/). A transcription of a recorded initial language is provided in figure 1 (labels below).

2.3. Procedure

Two participants corresponding to one generation of a chain went into the laboratory at the same time. Each participant had a separate cubicle with a networked computer, headphones and a microphone. All pairs experienced a training phase in which they were instructed to learn the language and an interaction phase in which they played the communication game through the computer.

FIGURE 1

Example of initial language. Labels above are those randomly generated (used in CW); labels below are transcriptions of the corresponding recorded labels (used in CS).

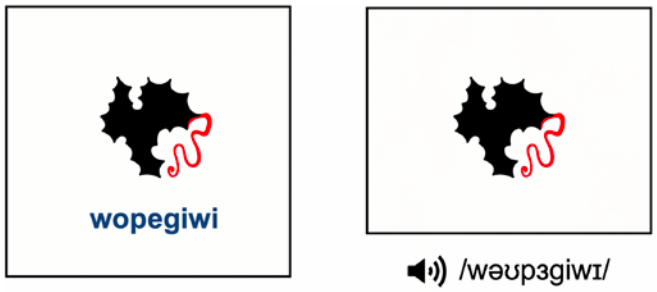


2.3.1. Training phase

Both participants were exposed to the same language (the same pairing of meanings and signals) presented in the same random order. In CW, each meaning appeared on the screen for 1 second and then the corresponding signal (as text) appeared below, both remaining on the screen for 5 seconds (see figure 2). In CS, each meaning was displayed on the screen for 6 seconds, and during that interval, the recording of the corresponding signal (which had a duration of 2 seconds) was played over the headphones 2 times with a pause of 1 second prior to the first playback and between repetitions (see figure 2). As in Kirby et al. (2015) the training phase consisted of 6 blocks, each block containing the whole language (independently randomised for each block), so the participants were exposed to each meaning-signal pairing 6 times in total. This was the same for both conditions.

FIGURE 2

Example of training screen on CW (left) and CS (right). The latter shows only the meaning while the signal is played over the headphones.



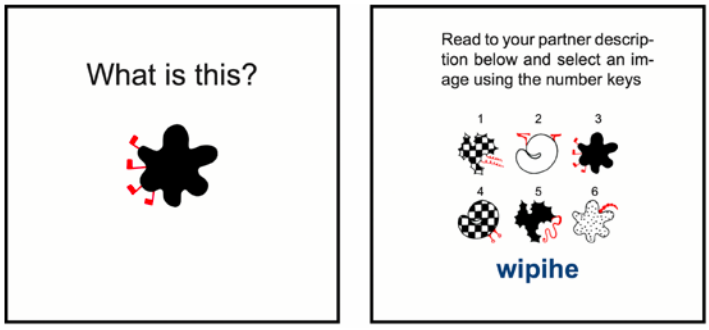
2.3.2. Interaction phase

Participants took turns as Director and Matcher in a series of interaction trials.

2.3.2.1. Written condition

In each trial, the target meaning appeared on the Director's screen, who was instructed to type the corresponding signal and send it to the Matcher by pressing the enter key (see figure 3). The Director's typed signal then appeared on the Matcher's screen together with 6 meanings (the target and 5 others chosen at random), from which the Matcher had to choose the one corresponding to the Director's signal by pressing the number keys 1-6. After the Matcher had made their selection both participants got feedback: a success/failure message appeared on both screens followed by the Director's signal paired with correct meaning on the Matcher's screen and by the Matcher's choice paired with the Director's signal on the Director's screen. Note that they only got to see their partner's output/choice but not the training signal. A point was added to their score if the interaction was successful (i.e. the Matcher chose the correct meaning).

FIGURE 3
Example of interaction (written). Director (left) has to type the signal, Matcher (right) has to choose the correct image between 6 possibilities.



2.3.2.2. Spoken condition

The target meaning appeared on the screen of the Director, who was instructed to press the space bar to start recording and say the signal into the microphone. The recording stopped automatically after 2 seconds and was played over the Director's headphones, and they had the option of sending the signal to the Matcher or make a new recording, repeating the process as many times as needed. This record-and-approve process was meant to be analogous to CW, in the sense that the Director gets to see what they are typing and make as many changes

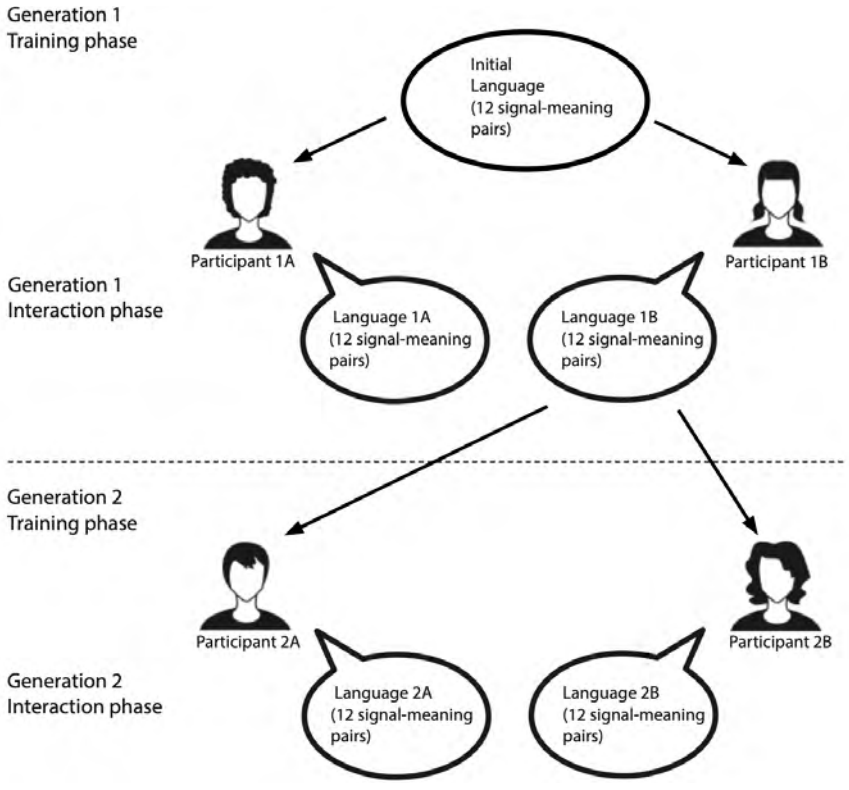
as needed but the Matcher only gets to see the final version. Once the Director's signal was sent, 6 meanings appeared on the Matcher's screen (the target and 5 others chosen at random) and after 1 second the signal was played 2 times over the Matcher's headphones with an interval of 1 second in between. After the Matcher's choice and the failure/success message, the Director's signal was played once over both participant's headphones while the correct/chosen meaning was shown on the screen. As in CW, participants got a point for success.

2.3.3. Iteration process

In both conditions, each participant acted as Director and Matcher 2 times for each meaning during the interaction phase, organised in 2 blocks of 12 independently randomised trials per participant (48 trials in total). Participants got their score on the screen at the end of each block. As in Kirby et al. (2015), the data produced by one of the participants (chosen at random) in the second interaction block formed the language to be passed on as the training language to the next generation (see figure 4 for a schematic explanation).

FIGURE 4

Schema of Iteration Procedure. In Generation 1, both participants are trained in the same randomly generated language. Each participant produces a whole language during the interaction phase. The language produced by one of the participants, chosen at random, is the training language for the next generation.



2.3.4. Orthographic questionnaire

As a post-experiment questionnaire, participants in CS were exposed to the whole set of meanings again. When each meaning appeared on the screen, the signal they produced on the second interaction block was played over the headphones twice and they were asked to type what they thought was the appropriate spelling of that signal. The main reason for this additional step was to see whether participants were thinking of the signals as words or phrases.

3. Results

Prior to the analysis, the recordings were transcribed using Speech Assessment Methods Phonetic Alphabet SAMPA (Wells et al., 1992). Although there was a diversity of dialects among the participants, in order to avoid excess of noise given by individual differences, transcriptions were performed attending to phonemes of Received Pronunciation, considered the best-known manifestation of Standard Southern British English (Giegerich, 1992), and a variant to which the participants, most of whom study or work in a British university environment, are exposed to. An exception was made for rhotic consonants, where two variants were transcribed (/ɹ/ and /ɾ/).

3.1. Qualitative analysis

A qualitative analysis of the final languages (the ones produced by the participants in the 5th generation during the second interaction block) showed some unexpected results. Examples of final languages are shown in figures 5 and 6.

Most final languages exhibited ambiguity (with at least one label corresponding to more than one meaning), and one of them showed systematic underspecification (see figure 6), distinguishing only between shapes and ignoring textures, with a few exceptions. In all other cases, ambiguity did not follow any obvious rules: see for example languages in figure 5, where some ambiguous signals (or very similar signals that differ in a single character only) are paired with meanings with shared shape and others with meanings with shared texture. Some languages exhibited compositional structure, but not for all meanings: see figure 5, for example, where there is a re-use of “haku-” for spiky shapes, but then no consistent morphemes for the textures nor for the other shapes.

As for the spelling post-experiment questionnaire of CS, only three participants used spaces to separate the signals, yet it seems those participants were separating syllables rather than word boundaries: they separated as many syllables as the signal had, rather than as

many features the meaning had. There was no correlation between the use of this strategy and performance in the interaction phase.

There were other unexpected findings. A few participants in CS made use of tonal and long vowels to distinguish between meanings (see figure 6). As well, two final languages in CW showed an orthographic bouba-kiki effect (cf. Cuskley et al., 2015), where spiky shapes were associated with signals containing one or more "k" and "h" while rounded shapes had a predominance of "m", "n" and "s" (see again figure 6, specifically the left column in contrast with the middle and right columns).

FIGURE 5
Example of final written language

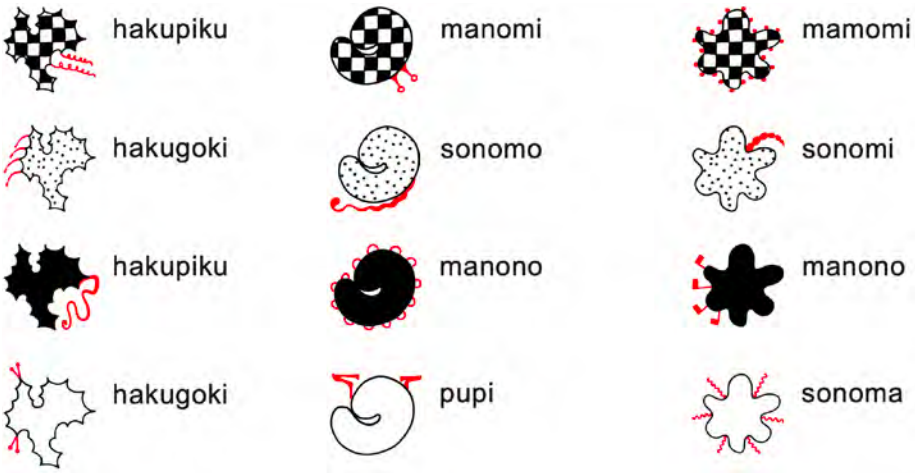
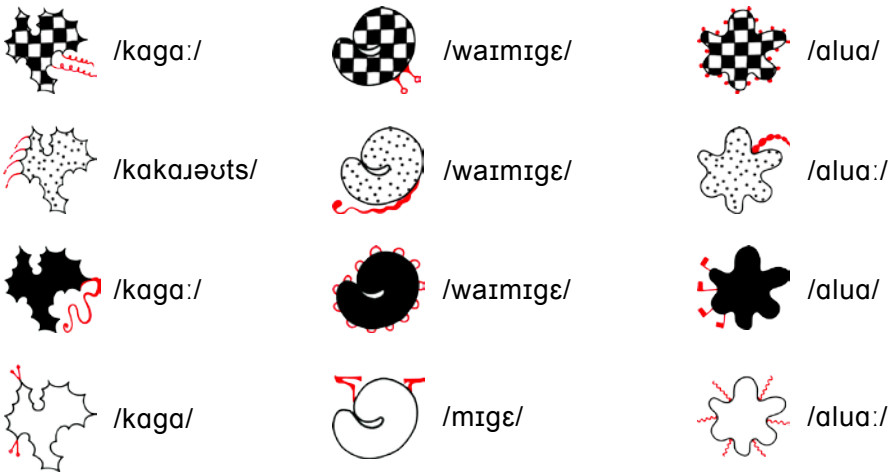


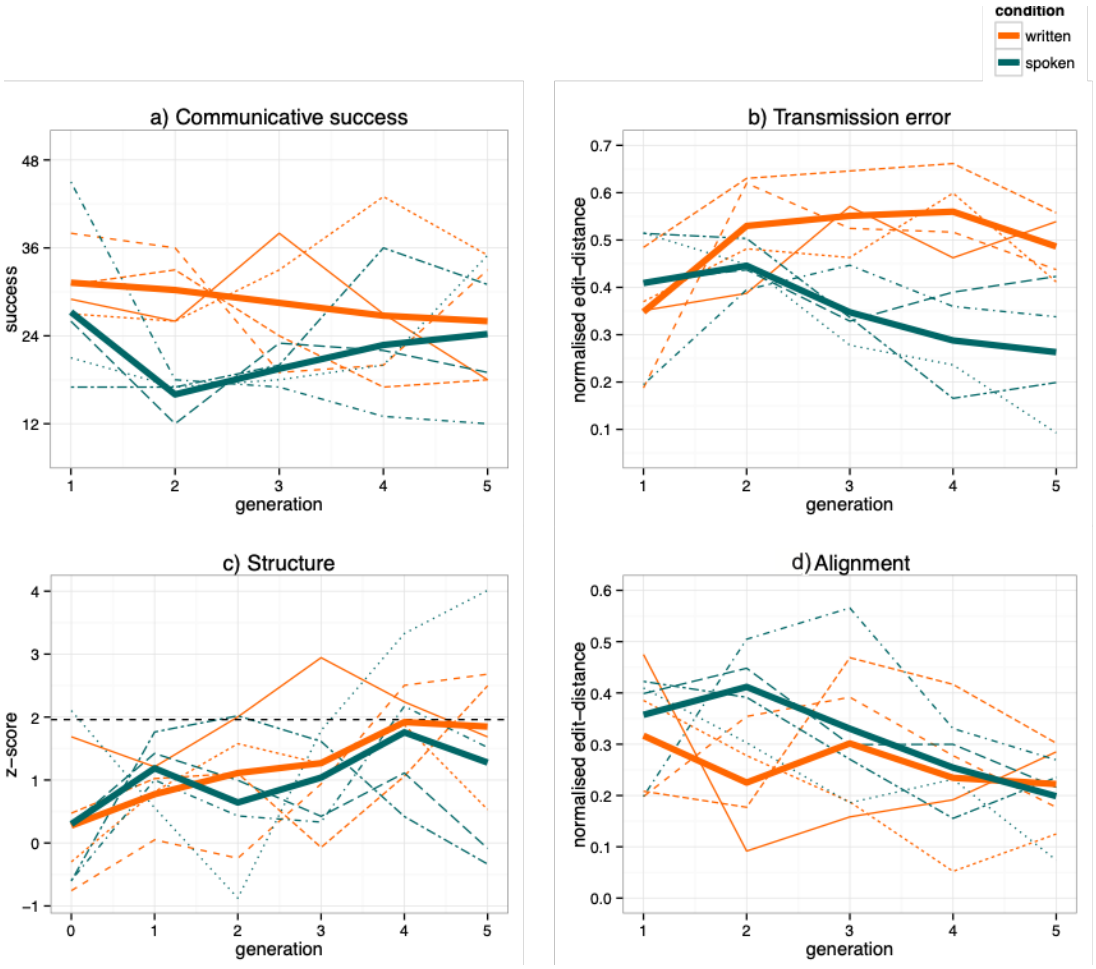
FIGURE 6
Example of final spoken language

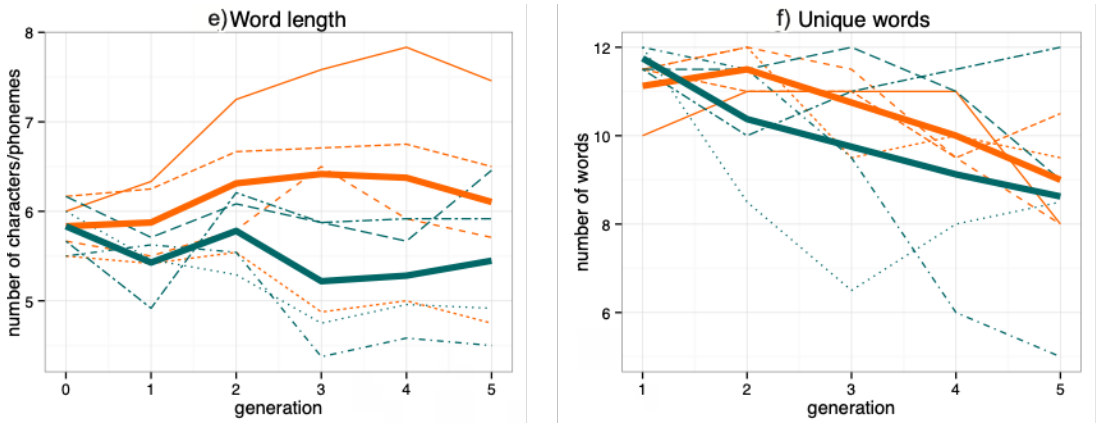


3.2. Quantitative analysis

Some of the measures used were based on Kirby et al. (2015): communicative success, transmission error and structure. The present study also implemented new measures: alignment, average word length and number of unique words. A summary of quantitative measures is provided in figure 7. Mixed-effects linear models were performed for all measures, considering condition (C), generation (G) and their interaction as fixed effects and chain as random effect (random intercept for chain and by-chain random slope for generation). The use of chain as a random effect aimed to account for possible idiosyncratic variation between initial languages. Each model was then compared against a null model for each fixed effect to obtain statistical significance using the Likelihood Ratio Test (as advised by Winter, 2013). All statistical analyses were performed using R (R Core Team, 2015), and for models *lme4* package (Bates et al., 2015) was used.

FIGURE 7
Summary of Quantitative Measures





Dashed lines represent individual chains; solid lines represent average results of all chains per condition. a) Communicative Success: as predicted, success is lower in CS. b) Transmission error: there is a significant effect of condition and a significant interaction, but in the opposite direction of predictions. c) Structure: there is a significant effect of generation and no difference between conditions, yet only 3 final languages score above 1.94. The language with the highest score for structure is mentioned in the qualitative analysis as the only clear case of systematic underspecification (figure 6). d) Alignment. There is an apparent decrease in distance (on increase in alignment) over generations, although it is not significant according to the model. e) Word length: words in CS are significantly shorter. f) Unique Words: homonymy increases in both conditions.

3.2.1. Communicative success

Communicative success is the score obtained by each pair of participants during the interaction phase, with a minimum of 0 (no successful matching of the target by either of the participants) and a maximum of 48 (all correct matches, considering that each of the 12 meanings appears as a target 4 times in total). Results are shown in figure 7a. In line with predictions, models did reveal a significant effect for C ($\chi^2(2)=9.819$, $p=0.001$): on average, pairs in CS had 7.123 (SE=2.011) less successful interactions than the pairs in CW at generation 1. However, contrary to general predictions, there was no significant effect for G ($\chi^2(2)=0.230$, $p=0.632$) nor a significant interaction ($\chi^2(1)=0.294$, $p=0.588$).

3.2.2. Transmission error

Transmission error was calculated using the normalised edit-distance between each training signal and the corresponding signal (associated with the same meaning) produced by the participants in the second block of the interaction phase. The method used for edit-distance was different between conditions. For the CW, following Kirby et al. (2015), normalised Levenshtein distance was used, which calculates the smallest number of substitutions, insertions and deletions needed to transform one string into another and divides it by the number of characters of the longest string.

For CS a phonological edit-distance was implemented (by adapting an open-source Python function[4]), calculated in a similar way to Levenshtein, but weighting the substitutions according to a (simplified) phonological feature chart (cf. Heeringa, 2004). We considered place, manner and voice as attributes of consonants, and frontness, openness, diphthong and tone/length (only when used for contrasting meanings) as attributes of vowels. This implementation was motivated by the fact that Levenshtein distance does consider the degrees of similarity existing between phonemes from a perceptual point of view: for example, /p/ and /b/ are more easily confusable than /p/ and /f/, and therefore a substitution of the first kind should account as smaller (Heeringa, 2004; Wieling et al., 2014). It also reduces the effects of subjectivity in transcription: inaccuracies in easily confusable phonemes, which are more likely to happen, have a relatively small effect.

Results are shown in figure 7b. Models did reveal a significant effect of C ($\chi^2(1)=15.396$, $p<0.001$), where the difference between training signals and produced signals was on average 0.252 (SE=0.029) normalised edit distances lower in CS than in CW. This means that C had the *opposite* effect to the prediction (and contradicts Nelson et al., 2015): spoken labels were more accurately reproduced. Also contradicting predictions, no effect of G was observed ($\chi^2(1)=0.9$, $p=0.976$), but there was a significant interaction ($\chi^2(1)=6.092$, $p=0.014$), where transmission error decreases at each generation by 0.091 (SE=0.035) more in CS than in CW.

3.2.3. Structure

Structure was defined as the z-scores of the Mantel tests (1000 permutations) between signal similarities (calculated using normalised Levenshtein distance for CW and normalised phonological distances for CS) and meaning similarities. Meaning similarities were calculated using Hamming distance, which counts the number of characters that are different between two strings. For the meaning set in this experiment, two pictures differing only in shape have a distance of 1, and those differing in shape and filling a distance of 2. This measure of structure is meant to capture whether any correspondence between signal distances and meaning distances is greater than what it would be expected by chance: z-scores higher than 1.96 correspond to a p-value lower than 0.05.

Results are shown in figure 7c. As predicted, there was no effect of C ($\chi^2(1)=0.161$, $p=0.688$) and a significant effect of G ($\chi^2(1)=5.089$, $p=0.024$), where the z-scores for structure increased by 0.264 (SE=0.099) points each generation. There was no significant interaction ($\chi^2(1)=0.428$, $p=0.512$): the rate of increase was the same across conditions. As for the final languages, an in-

4 https://en.wikibooks.org/wiki/Algorithm_Implementation/Strings/Levenshtein_distance#Python.

dependent sample t-test showed no significant difference between conditions ($t(4.36)=0.513$, $p=0.633$). This is in line with predictions: an equivalent result for structure was expected as the outcome of cultural evolution. However, as it can be appreciated in figure 7, all but 3 final languages (1 written and 2 spoken) show a structure score below 1.96, and the average structure score for each condition was also below 1.96. That is, contradicting the general prediction, even though there is a significant increase over generations, final languages in general do not show a significant level of structure: hence, they seem to be better characterised as holistic than as compositional.

3.2.4. Alignment

Alignment was defined as the distance between the signals produced by each participant in a pair for each meaning during the second interaction block, and it was measured using normalised Levenshtein distance for CW and phonological distance for CS. Average results are shown in figure 7d. Contradicting predictions, models did not show any significant effect of C ($\chi^2(1)=0.0006$, $p=0.981$) or G ($\chi^2(1)=2.608$, $p=0.106$). No significant interaction was found either ($\chi^2(1)=1.323$, $p=0.25$).

3.2.5. Word length

Word length considers the number of characters of the signals produced by the participants in the second interaction block (transcriptions for CS). Average results are shown in figure 7e, where the spoken chains exhibit a more apparent trend (they tend to maintain or shorten word length) than the written ones. Contrary to prediction, models showed a marginally significant effect for C ($\chi^2(1)=2.755$, $p=0.097$) where signals produced in CS were on average 0.445 ($SE=0.260$) characters shorter than signals produced in CW. No significant effect for G was found ($\chi^2(1)=0.066$, $p=0.797$), nor a significant interaction ($\chi^2(1)=0.7$, $p=0.4$).

3.2.6. Unique words

This measure aimed to detect homonymy, and was calculated by averaging the number of words produced by each participant in a pair. Results are shown in figure 7f. In line with predictions, no effect for C was found ($\chi^2(1)=0.699$, $p=0.403$). However, contradicting predictions, there was a significant effect of G ($\chi^2(1)=6.492$, $p=0.010$): on average, there was a loss of 0.66 words ($SE=0.21$) every generation, and consequently final languages did exhibit homonymy (expressivity is not maintained over generations despite the pressure arising from communication). No interaction was found ($\chi^2(1)=0.175$, $p=0.675$).

4. Discussion

4.1. General predictions

The above results contradict most general predictions of this study. Although the experimental design considers pressures for compressibility and expressivity, final languages are mostly holistic: there is a significant increment in structure along generations, but most final languages exhibit an amount of structure below what it might be expected by chance. There is also homonymy, which according to Kirby et al. (2015) should not develop where there is pressure for expressivity. This is reflected in communicative success: because most languages in the experiment are not developing systematic structure and because they do not consist of one-to-one mappings, evolution is not driving them to be better suited for communication, and participants are failing to produce aligned signals. Results for transmission error are complex: along generations, languages become more learnable only in CS.

Strikingly, results for the written condition do not replicate those in Kirby et al. (2015). Mixed-effect linear models were run for those measures that were based in their study with CW only, to test the main effect of generation. For error, the model showed a marginally significant effect (χ^2 (1) = 6.738, p = 0.009) but opposite to prediction: every generation *increases* error by 0.046 (SE = 0.013). For structure, the model did show a significant effect (χ^2 (1) = 4.175, p = 0.041) in the expected direction, where every generation increased their z score by 0.327 (SE = 0.139). However, as stated earlier, final languages of the written chains show on average an amount of structure that is below what it would be expected by chance. This represents a striking difference with the z-scores obtained in the original experiment, where the average of the chain condition at generation 5 is almost 4 (see figure 4 (c) in Kirby et al., 2015: 97). Finally, for communicative success a model with by chain random slope for generation failed to converge, while a model with random intercept for chain showed no significant effects (χ^2 (1) = 1.466, p = 0.226), suggesting that our results do not replicate Kirby et al. (2015) in this measure either.

It could also be the case that the difference found in transmission error is related to the use of Levenshtein distance for CW (as in Kirby et al., 2015) instead of a phonological distance. Conrad (1964) noted that participants tended to confuse graphemes associated to similar phonemes in visual memory tasks (cited in Baddeley et al., 2009: 22). However, doing the calculations again with phonological distance (a variation of the one used for the spoken condition by mapping roughly graphemes to phonemes) did not make a difference. A new mixed-effect linear model (for CW) showed a significant effect of G (χ^2 (1) = 4.678, p = 0.031), where every generation *increases* error by 0.031 (SE = 0.013). This is less than what we got with Levenshtein distance, but the trend is still opposite to that in Kirby et al (2015).

How can we interpret this difference? As stated earlier, methods for CW attempted to resemble those in Kirby et al. (2015) as closely as possible. I used very similar stimuli: while

images were exactly the same, labels were generated the same way by selecting at random from a set of syllables. As for the software, I adapted a Python code written by one of the co-authors of Kirby et al. (2015), which uses several functions of PsychoPy library (Peirce, 2007). Both experiments were run in networked computers, and the participants were located in separate booths. The participant populations were also very similar, mostly students of the University of Edinburgh, recruited via the University student employment service. The data collection time was not the same (summer vacations in our case, both summer and mid-semester in Kirby et al., 2015) yet, since participation was paid in cash and not academic credit, there is no apparent reason to think that this affected the participants' performance.

The only significant difference in the methods, which in consequence might have affected the results, was that Kirby et al. (2015) were running three experiments at the time and offered a cash prize to the pair obtaining the highest score in the shortest time. This study, in contrast, ran only one pair at the time and all participants were compensated equally, no matter how well they performed. Maybe this prize played a crucial (yet unexpected) role, and it raises an important point: if our aim is to recreate in the laboratory those pressures acting over language "in the wild", these need to be meaningful for the participants in order to act as real pressures. Other than this, I found no other apparent explanation for these puzzling results.

4.2. Condition predictions

Some predictions about modality effect were fulfilled. For instance, the communication game was harder in CS, but it seems unlikely that this can be explained by modality effect in learning. As said before, recall should be poorer in CS (in line with Nelson et al., 2015), whereas our results showed the opposite effects on transmission error. Lower communicative success may therefore reflect other issues. It may be that playing the signal two times was not enough for the matcher to understand and/or recall, especially if their partner spoke a very different dialect or very quietly. However, this shows precisely a main difference between modalities: speech is temporary (while writing is permanent) and its channel is noisy.

As mentioned before, modality effect for transmission error was opposite to prediction. However, error did decrease over generations in CS, suggesting that spoken languages were becoming more learnable. It seems unlikely that the reason is the loss of words or the increase of structure, since these effects were equal across conditions. Instead, it could be a matter of word-length: according to the word-length effect for short-term memory hypothesis, the number of words that someone can remember is inversely proportional to their length (Baddeley et al., 1975).

As for the remaining modality predictions, results for alignment suggest that the effect of pronunciation differences was not greater than the effect of typing mistakes or inaccuracies

(it may also suggest that our phonological distance contributed to lessen the effect of pronunciation). As for length, the difference between conditions suggests that the noisy channel (CS) had a bigger effect than the difference in effort (although it can also be argued that most participants, who use a computer on a daily basis, may find typing quite effortless). However, the fact that the trend for CW is not so clear hinders the interpretation of this measure. Finally, results for unique words again suggest that pressure for expressivity was the same for both conditions, although it was weaker than in Kirby et al. (2015).

5. Conclusion

This study aimed to test modality effects in the cultural evolution of a holistic language through an iterated learning experiment with literate adult participants. To do so, I replicated the experiment in Kirby et al. (2015) contrasting their chain condition, which was written, with a spoken condition. I also implemented a phonological distance method for measuring error and alignment in CS, which takes into account degrees of perceptual similarity between phonemes. Results supported my predictions for modality effects in some of the measures (accuracy, structure, unique words) but contradicted them in others (error, alignment, length). Strikingly, results for CW did not replicate those in Kirby et al. (2015) in any of the measures, even though this set-up was exactly the same as their chain condition, including learning and communication and therefore pressures for both simplicity and expressivity. According to the literature reviewed, compositional languages should emerge under these pressures, which was not the case.

Given the above, it is not easy to draw conclusions about modality effects. The fact that the control condition did not behave as expected might encourage further replications that consider participants' motivation (e.g. a prize for good performance). Furthermore, this work still reflects an alphabetic literacy bias since it was run with alphabetic literate participants. In this sense, in order to generalise about evolutionary pressures acting on all languages (those that use other writing systems, or none), further work might also include non-alphabetic literate participants, either illiterate or literate in ideographic or syllabic systems. As well, my experiment deliberately excluded all visual speech in CS. It may also be interesting to run a replication with and without this input to test its contribution (cf. Sumby and Pollak, 1954). Finally, further work on modality effects may help to understand if the prevalence of writing in everyday communication through computer-mediated conversation (see Soffer, 2010) is having an influence in the cultural evolution of language.

6. Acknowledgments

This paper is largely based on the work done for my MSc Dissertation at the Centre for Language Evolution, University of Edinburgh, under the supervision of Prof. Kenny Smith, who I

thank in the first place. I would also like to thank the CLE in general and Pilar Oplustil for their help, and two anonymous reviewers for their feedback and useful comments. I was funded by the scholarship Magister Becas Chile, 2015, Nº 73160041.

7. References

Baddeley, A. D., N. Thomson & M. Buchanan, 1975: "Word length and the structure of short-term memory", *Journal of verbal learning and verbal behavior* 14 (6), 575-589.

Baddeley, A., M. Eysenck & M. Anderson, 2009: *Memory*, Psychology Press.

Bates, D., M. Maechler, B. Bolker, S. Walker, 2015: "Fitting Linear Mixed-Effects Models Using lme4", *Journal of Statistical Software* 67 (1), 1-48.

Caplan, D., G. Waters, J. Bertram, A. Ostrowski & J. Michaud, 2016: "Effects of Written and Auditory Language-Processing Skills on Written Passage Comprehension in Middle and High School Students", *Reading Research Quarterly* 51 (1), 67-92.

Chafe, W., & D. Tannen, 1987: "The relation between written and spoken language", *Annual Review of Anthropology* 16, 383-407.

Chater, N., & P. Vitányi, 2003: "Simplicity: a unifying principle in cognitive science?", *Trends in cognitive sciences* 7 (1), 19-22.

Christiansen, M. H., & N. Chater, 2008: "Language as shaped by the brain", *Behavioral and brain sciences* 31 (05), 489-509.

Coulmas, F., 2003: *Writing systems: An introduction to their linguistic analysis*, Cambridge: Cambridge University Press.

Crottaz-Herbette, S., R. T. Anagnoson & V. Menon, 2004: "Modality effects in verbal working memory: differential prefrontal and parietal responses to auditory and visual stimuli", *Neuroimage* 21 (1), 340-351.

Cuskley, C., J. Simner & S. Kirby, 2015: "Phonological and orthographic influences in the bouba-kiki effect", *Psychological Research*, 1-12.

Giegerich, H. J., 1992: *English phonology: An introduction*, Cambridge: Cambridge University Press.

Harklau, L., 2002: "The role of writing in classroom second language acquisition", *Journal of second language writing* 11 (4), 329-350.

Heeringa, W. J., 2004: *Measuring dialect pronunciation differences using Levenshtein distance*. Doctoral dissertation, University of Groningen.

Hockett, C. F., 1960: "The Origin of Speech", *Scientific American* 203, 88-96.

Kemp, C., & T. Regier, 2012: "Kinship categories across languages reflect general communicative principles", *Science* 336 (6084), 1049-1054.

Kirby, S., 1999: "Learning, bottlenecks and infinity: a working model of the evolution of syntactic communication" in *Proceedings of the AISB*, vol. 99, 55-63.

Kirby, S., 2000: "Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners" in C. Knight (ed.): *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, Cambridge: Cambridge University Press, 303-323.

Kirby, S., H. Cornish & K. Smith, 2008: "Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language", *Proceedings of the National Academy of Sciences* 105 (31), 10681-10686.

Kirby, S., M. Tamariz, H. Cornish & K. Smith, 2015: "Compression and communication in the cultural evolution of linguistic structure", *Cognition* 141, 87-102.

Linell, P., 2004: *The written language bias in linguistics: Its nature, origins and transformations*, London: Routledge.

Mesoudi, A., & A. Whiten, 2008: "The multiple roles of cultural transmission experiments in understanding human cultural evolution", *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 363 (1509), 3489-3501.

Mesoudi, A., 2016: "Cultural evolution: Integrating psychology, evolution and culture", *Current Opinion in Psychology* 7, 17-22.

Morais, J., L. Cary, J. Alegria & P. Bertelson, 1979: "Does awareness of speech as a sequence of phones arise spontaneously?", *Cognition* 7 (4), 323-331.

Nelson, J. R., M. Balass & C. A. Perfetti, 2005: "Differences between written and spoken input in learning new words", *Written Language & Literacy* 8 (2), 25-44.

Olson, D. R., 1996: "Towards a psychology of literacy: On the relations between speech and writing", *Cognition* 60 (1), 83-104.

Peirce, J. W., 2007: "PsychoPy - Psychophysics software in Python", *Journal of Neuroscience Methods* 162 (1-2), 8-13.

Pinker, S., & P. Bloom, 1990: "Natural language and natural selection", *Behavioral and Brain Sciences* 13 (1), 707-784.

R Core Team, 2015: "R: A language and environment for statistical computing", Vienna: R Foundation for Statistical Computing, URL https://www.R-project.org/.

Read, C., Z. Yun-Fei, N. Hong-Yin & D. Bao-Qing, 1986: "The ability to manipulate speech sounds depends on knowing alphabetic writing", *Cognition* 24 (1-2), 31-44.

Reis, A., & A. Castro-Caldas, 1997: "Illiteracy: A cause for biased cognitive development", *Journal of the International Neuropsychological Society* 3 (05), 444-450.

Shannon, C. E., 1948: "A mathematical theory of communication", *Bell System Technical Journal* 27, 623-656.

Smith, K., & S. Kirby, 2008: "Cultural evolution: implications for understanding the human language faculty and its evolution", *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 363 (1509), 3591-3603.

Soffer, O., 2010: "Silent orality: Toward a conceptualization of the digital oral features in CMC and SMS texts", *Communication Theory* 20 (4), 387-404.

Sumby, W. H., & I. Pollack, 1954: "Visual contribution to speech intelligibility in noise", *The Journal of the Acoustical Society of America* 26 (2), 212-215.

Tallerman, M., 2007: "Did our ancestors speak a holistic protolanguage?", *Lingua* 117 (3), 579-604.

Theisen-White, C., S. Kirby & J. Oberlander, 2011: "Integrating the horizontal and vertical cultural transmission of novel communication systems" in *Proceedings from CogSci 2011: Cognitive Science Society Conference*, vol. 1, 956-961.

Verhoef, T., S. Kirby & B. de Boer, 2014: "Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals", *Journal of Phonetics* 43, 57-68.

Weissberg, B., 2000: "Developmental relationships in the acquisition of English syntax: Writing vs. Speech", *Learning and Instruction* 10 (1), 37-53.

Wieling, M., J. Nerbonne, J. Bloem, C. Gooskens, W. Heeringa & R. H. Baayen, 2014: "A cognitively grounded measure of pronunciation distance", *PloS one* 9 (1), e75734.

Winter, B., 2013: "Linear models and linear mixed effects models in R with linguistic applications" (arXiv preprint arXiv:1308.5499).

Wray, A., 1998: "Protolanguage as a holistic system for social interaction", *Language & Communication* 18 (1), 47-67.

Wells, J., W. Barry, M. Grice, A. Fourcin & D. Gibbon, 1992: *Standard Computer-Compatible Transcription, Esprit Project 2589 (SAM), Doc. no*, SAM-UCL-037, London: Phonetics and Linguistics Departament, UCL.