The emergence of verse templates through iterated learning

Title The emergence of verse templates through iterated learning

Authors deCastro-Arrazola, Varuṇ
Kirby, Simon

Address Varuṇ deCastro-Arrazola
Meertens Instituut
Oudezijds Achterburgwal 185
1012 DK Amsterdam

E-mail varunasarman@gmail.com

Abstract

Every language produces some type of verse in the form of songs, poems or nursery rhymes, which can be analysed as a layer of words set to a template (e.g. a tune, a poetic metre). Verse templates typically consist of hierarchically organised sections: songs are made up of stanzas, divided into lines, containing bars, etc. We hypothesise that this kind of patterns may emerge in the process of cultural transmission; unstructured sound sequences impose a challenge to short-term memory, but chunking the input makes it easier to parse and reproduce the sequences accurately.

In order to test this hypothesis, we have run an iterated learning experiment where random sequences of syllables are evolved across four transmission chains with ten generations of subjects each (all native Dutch speakers). The initial random sequences are generated by concatenating twelve tokens of the set {*ban, bi, ta, tin*}, as a way to materialise the abstract verse templates without using content-words. More precisely, the experiment aims to model the sequences of nonsense syllables used in many traditions to communicate the rhythmic patterns underlying songs (e.g. *bols* in Hindustani music, *lalay* patterns in Berber verse). Participants listened to the sequences of syllables, and tried to reproduce them using four computer keys, each mapped to one of the four syllables used in the input sequences. The relative timing of the participants' responses were normalised so that the input always consisted of completely isochronous sequences.

Overall, the results show that sequences become shorter, easier to recall and more structured in the transmission process. Some regularities can be related to a global tendency to chunk the input and increase the popularity of a handful of ngrams. Besides, sequences increasingly tend to be opened by a heavy syllable (e.g. *ban*) and closed by a light syllable (e.g. *ta*), which can derive from a Dutch-specific bias.

1 Introduction

Language is present in every human society, in a variety of forms and contexts. Beyond everyday speech, language is also widely found in verse form, for instance in songs, poetry, chant or nursery rhymes. All these phenomena set words to some kind of non-linguistic template, such as a musical tune, a beat pattern, or a poetic metre.

Songs can also exist without the presence of words, by using sounds with no semantic content (e.g. vocables, scat singing, humming), or by using exclusively musical instruments. Nevertheless, songs prototypically include a text with meaningful words and a musical tune. Indeed, whereas some societies do not produce purely instrumental songs (e.g. the Pintupi of Central Australia; Moyle 1979), vocal music is regarded as a cultural universal, since there is no evidence of any society lacking it (Brown & Jordania 2011). Hence, it is most parsimonious to investigate the origins of verse templates in the context of vocal productions.

One feature of verse templates is that they are typically composed of chunks. The template underlying a nursery rhyme like *Eeny meeny miny moe* could be described as consisting of one big entity with 28 syllable-holders (Example 1). However, it can be argued that the template contains further sub-groupings. Those 28 syllables can be divided into 4 lines of 7 syllables each. The 7 syllables in each sequence are also structured: they follow an alternating strong-weak pattern (strong syllables are marked with bold and italics in the example). Knowing this, one can produce a new line by looking for words which match the template.

The relative prominence of syllables is one of the strategies used to divide verse templates into groups and sub-groups. The lines of this nursery rhyme can be analysed as having four constituents: (+ -)(+ -)(+ -)(+ -), where each parenthesis represents a pair of strong-weak beats in musical terms. Given the pattern of the first three constituents, a final eighth syllable could fill the last weak beat, but we see that this is not done in the case of *Eeny meeny miny moe*. Leaving this empty gap leads to the grouping of the series of 16 strong-weak pairs of the song into 4 lines. Besides, a parallelism between the last sounds of each line (*moe, toe, go, moe* all share the rhyme -/əʊ/) further strengthens the segmentation into lines. The regular alternation

Example 1: The English nursery rhyme *Eeny meeny miny moe.*

1	2	3	4	5	6	7
ee	ny	***mee***	ny	***mi***	ny	***moe***
8	9	10	11	12	13	14
catch	a	***ti***	ger	***by***	the	***toe***
15	16	17	18	19	20	21
if	he	***ho***	llers	***let***	him	***go***
22	23	24	25	26	27	28
ee	ny	***mee***	ny	***mi***	ny	***moe***

of prominence, structural parallelism and systematic use of gaps are recurrent chunking cues used in verse systems; for an overview of these and other cues, see Fabb (1997).

Despite lacking a comprehensive survey of chunking in the verse systems of the world, ethnomusicological overviews suggest it is a pervasive feature of song (see e.g. phrase structure and rhythmic subdivision as reported by Lomax 1976 and Savage et al. 2015). Besides, it is considered central to music cognition more generally: "grouping can be viewed as the most basic component of musical understanding" (Lerdahl & Jackendoff 1983:13). It is thus appropriate to ask why segmented verse templates are so widespread. Our proposal can be summarised as follows: (1) verse templates are cognitive objects acquired by imitation; (2) humans readily chunk external stimuli in order to facilitate its storage in working memory; (3) the process of cultural transmission amplifies this chunking bias. Before presenting the experimental evidence supporting the proposal, we describe the theoretical motivation backing these three points.

Verse systems are transmitted by social learning, similarly to other aspects of human culture (Boyd & Richerson 1985). Specifically, imitation plays a central role, as seen in the fact that e.g. nursery rhymes remain stable across generations (Morin 2016, Opie & Opie 1959). Young children learn these instances of verse from adults or older peers by imitation, but small changes may be introduced in the process. Verse systems, hence, evolve through time via social learning.

In addition to verbatim imitation, the underlying structures of songs and poems are also used productively in order to create new instances of verse (this is manifest whenever established forms such as the blues or the sonnet are used as the basis to compose original texts). Interestingly, many traditions offer ways to dissociate the templates from the actual song instances,

Example 2: Sample sequences from two different musical traditions where nonsense syllables are used to realise metrical templates. Vertical lines indicate grouping of syllables within the sequence.

Berber tradition:	*a-lay-da* | *la-la-lay* | *da-lay-la* | *lal*
Hindustani tradition:	*dhin-nā* | *dhin-dhin-nā* | *tin-nā* | *dhin-dhin-nā*

namely through nonsense syllables which are used to explicitly represent, communicate or realise verse templates and their internal divisions. English typically represents a weak-strong pattern with the syllables *da-dum* (Fabb & Halle 2008); in Tashlhiyt Berber, verse lines can be composed by using a set of syllables for strong positions (*ay, lay(l), day(l)*), and a different one for weak positions (*a, la(l), da(l)*) (Jouad & Lortat-Jacob 1982, Dell & Elmedlaoui 2008). Similar systems of mnemonic syllables are used in Hindustani and Karnatic music (Clayton 2000, Reina 2013), and West-African drumming traditions (Knight 1984, Stone 1985, Euba 1990). In Example 2 we can see how these kinds of syllables are combined into small chunks to produce metrical templates or cycles used to create new songs.

Following these observations, our experimental design employs nonsense syllables as a proxy for the building blocks of verse templates. By doing so, we abstract away from real-world verse texts, and avoid potential syntactic and semantic confounds.

How does a human subject address the task of imitating a sequence of nonsense syllables? A strategy readily used by humans is to chunk the input, as seen e.g. in the way we memorise telephone numbers by grouping them in twos or threes. Chunking is a general cognitive strategy used spontaneously and in a wide range of contexts (Gobet et al. 2001). It appears early in ontogeny (Rosenberg & Feigenson 2013), which argues for its basic status in cognition. Crucial to working memory, hierarchical chunking can expand its limits to allocate tens of items at a time (Ericsson, Chase & Faloon 1980). Hence, we hypothesise that, when learning a verse template from a previous generation, individuals will show a chunking bias, effectively introducing regularities in the input and facilitating the task.

The third point of the proposal emphasises the fact that individual chunking biases can be amplified as their effect accumulates across generations of learners. This kind of cross-generational amplification has been successfully modelled in laboratory settings, by teaching participants

some stimuli and taking their output as the learning input for a subsequent participant (see Tamariz 2017 for a review). By *iterating* the transmission process we are effectively able to capture weak biases present in the learning of individuals. Alternatively, individual biases may be strong enough to create a fully-segmented template, without the need for a long-term iterated process of learning and reproducing (Morin 2018).

The experiment reported below investigates how patterns of syllables evolve in an iterated learning context. Previous studies have used such a paradigm to test how cultural transmission can make systematic structure emerge out of initially unstructured stimuli (Kirby, Cornish & Smith 2008). Studies of this kind typically show some material for a participant to learn (e.g. random associations between graphical objects and pseudo-words), and then ask the participant to use or reproduce the newly-learnt material. Subsequent subjects are given the output of the preceding participant as their input, so that small changes introduced by individuals can be transmitted from participant to participant, and the overall shape of the initial material evolves as a consequence of the accumulation of these changes. Within this paradigm, each experimental subject serves as a model for a generation in the process of cultural evolution.

Our experiment builds particularly on a previous study (Cornish, Smith & Kirby 2013) where random sequences of colour signals become more learnable and more structured in the transmission process between participants, and a study using an electronic drum pad to transmit musical rhythms (Ravignani, Delgado & Kirby 2016). We follow a similar procedure but employ sequences of syllables as stimuli, which resemble more closely the building blocks of verse templates. The main question we address is whether random sequences of syllables can evolve into structured patterns; for instance, by using some kind of chunking such as the one found in verse templates (Table 1). Specifically, we test a pair of related hypotheses, namely that the syllable-sequences produced by later generations will be more structured, and that they will become easier to recall as a result. Moreover, we expect the increase in structure to be gradual, reflecting the accumulation of small innovations within each generation, as shown in the studies cited above.

2 Method

The experiment follows an iterated learning approach (Mesoudi & Whiten 2008, Cornish, Smith & Kirby 2013, Kirby, Cornish & Smith 2008), where the set of stimuli presented to a participant (i.e. the input) is the set of responses given by the previous participant (i.e. the previous output). Each participant imitates whatever the predecessor has produced, not unlike the routine of Chinese whispers or the Telephone game. Four participants are given an initial set of stimuli created by the experimenter, and each of these four sets further develops independently through transmission chains. Each transmission chain involves ten human participants; the first participant in each chain listens to (and imitates) 30 randomly-generated sequences of 12 syllables; subsequent participants listen to (and imitate) the 30 sequences which the preceding participant in the chain has produced. Hence, every participant is asked to imitate (using a computer keyboard) 30 different sequences of syllables, one sequence at a time. Each of the ten subjects who take part in a transmission chain is referred to as a *generation*. Note that the sequences have a fixed length of 12 syllables only when presented to the first participant of a chain; the length of the sequences presented to subsequent participants will be variable, as it depends on how the preceding participants have imitated their input.

2.1 Participants

In total, 40 participants took part in the experiment (mean age = 24.3 years; 22 females, 18 males; left-handed = 7). All of them were native speakers of Dutch, and nine spoke an additional language natively. Participants were recruited at Leiden University and Radboud University (The Netherlands) to take part in a *Syllable Imitation Game*; all signed an informed consent before performing the task (in accordance to Leiden University's LUCL procedure). Each participant was assigned randomly to one of the four transmission chains, but at no point during the experiment were they informed that their input and output stimuli belonged to a chain connecting several subjects.

2.2 Stimuli

The first player of each of the four transmission chains received a different collection of semi-random sequences of syllables. A set of four syllables was used to generate all the sequences: {*ban, bi, ta, tin*}. Each of these syllables can be defined as a concatenation of three phonological units: an onset (i.e. the initial consonant), a nucleus (i.e. the vowel), and a coda (i.e. the final consonant or lack thereof); this is summarised in Table 1. In our case, each of these features takes one of two possible values: the onset can be [b] or [t]; the nucleus can be [a] or [i]; the coda can be present ([n]) or absent (-). These four syllables were recorded by a female native speaker of Dutch, and were normalised for pitch and intensity. Length was not kept constant because the two items lacking a coda (*bi, ta*) were meant to be shorter.

One important property of the set is that each syllable shares one and only one feature with each of the other three syllables in the set. This means that one can group the syllables in pairs according to their onset, nucleus or coda, resulting in three distinct similarity configurations to which subjects can be sensitive. The choice of onset and nucleus contrasts ([b] vs [t], and [a] vs [i]) seeks to maximise the perceptual distance between syllables.

The third dimension, namely introducing syllables with and without the [n] coda, intends to provide the participants with some cue for prominence. Syllables ending in a consonant tend to attract stress in the world's languages (Gordon 2006), and this is also holds for the Dutch lexicon (Van der Hulst 1984, van Heuven & Hagman 1988).[1] Making available syllables with potentially different degrees of perceived prominence is relevant given that Dutch poetry uses regular alternations of syllabic prominence (de Groot 1936), and Dutch songs place prominent and non-prominent syllables in a systematic way with respect to melodies (deCastro-Arrazola, van Kranenburg & Janssen 2015).

The initial stimuli for each chain consisted of 30 sequences of 12 syllables each. The sequences were generated by randomly permuting a pool containing 3 tokens of each of the 4 syllable types. The time interval between the onset of a syllable and the onset of the following syllable was kept

[1] Other acoustic features such as pitch, duration or spectral balance do provide more unambiguous cues for stress (Heuven & Jonge 2011), but we have not introduced these variables in order to avoid a strong bias in the stimuli.

	nucleus = [a]	nucleus = [i]
onset = [b]	*ban*	*bi-*
onset = [t]	*ta-*	*tin*

Table 1: Set of syllables used to create the initial sequences. Here displayed according to their defining units; shading represents the presence of the coda [n].

constant at 600 ms. Figure 11 of the Supplementary Information shows all 30 sequences produced as the initial generation of chain 1 and presented to the first participant of the chain.

2.3 Procedure

Participants are instructed that they will listen to sequences of syllables, and they are asked to reproduce them using a keyboard. The sequences are presented in auditory form through headphones (Beyerdynamic DT 880), and, after each of them, an on-screen microphone symbol indicates that it is their turn to reproduce the sequence. This is done using four keys positioned in a row, which correspond to the four syllable types. Each participant is assigned a random key-to-syllable mapping, kept constant throughout the task; if the same mapping was given to every participant, a motor bias or preference for particular keys could be transmitted and amplified over generations.

Before the task starts, participants are given the chance to try the keys in order to adjust the volume and familiarise themselves with the way of typing in syllables. Then, a first training round is presented. The aim of this round is to ensure that they are competent in the key-to-syllable mapping; this mapping has to be memorised, as no visual cues are provided. This round presents (in random order) all the 16 two-syllable combinations which can be generated with the four syllable types.

Subjects are asked to reproduce each two-syllable pattern immediately after it has been played. If the subject presses an incorrect key, a written message appears in the screen requesting to try again. Once the pattern is reproduced correctly, the following pattern gets played.

After the training round is finished, the 30 experimental sequences are presented one by one, and participants try to reproduce each sequence immediately after having listened to it. The 30

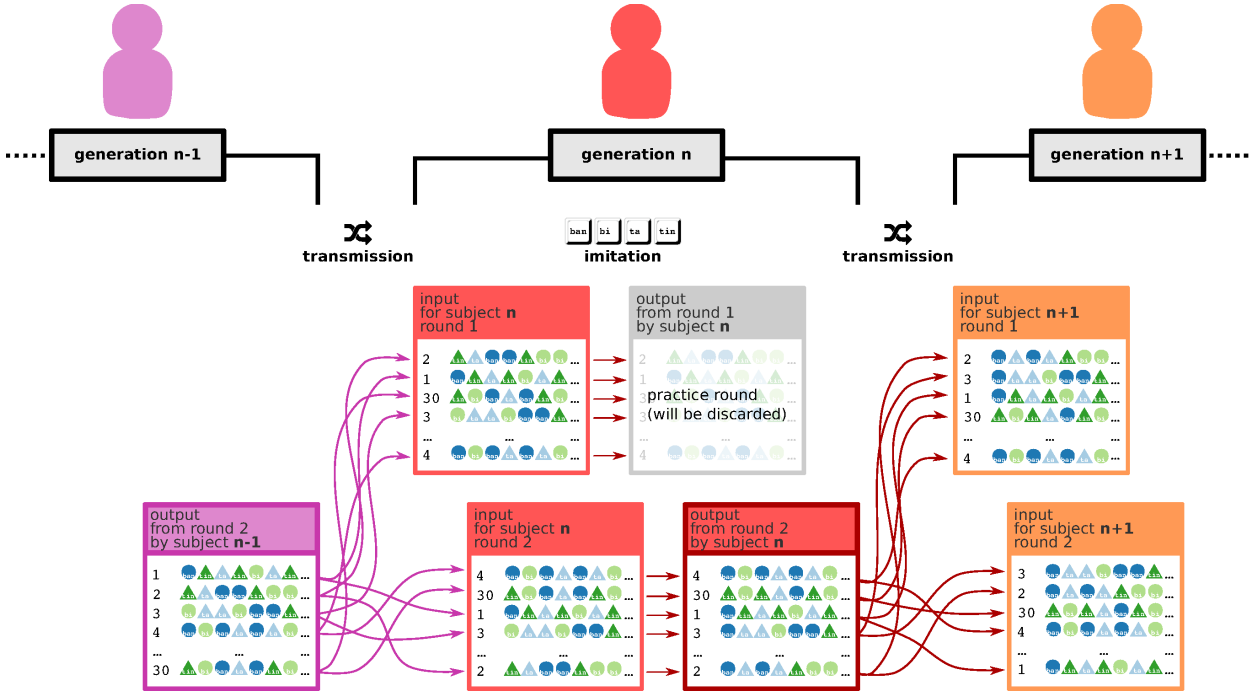


Figure 1: Schematic representation of the two experimental rounds for one subject (n), partially including the preceding subject (n-1), and the following subject (n+1). The box with a dark pink frame represents the 30 sequences produced by subject n-1, which get randomised twice to serve as input for rounds 1 and 2 of subject n. The output of the latter (dark red frame) will continue the iteration by serving as input for subject n+1.

experimental sequences correspond to the sequences produced by the preceding participant, or, in the case of the first-generation participants, to the 30 sequences generated by the experimenter. In order to ensure additional familiarisation with the task, participants are presented with this experimental round twice, with a pause-screen in between. Each of these two experimental rounds contain exactly the same 30 sequences of syllables as an input, although randomised in two different ways. Figure 1 illustrates this section of the experiment; note that only the second of the two experimental rounds is kept and passed on for the following subject to imitate.

The routine for these two experimental rounds is similar to the training one: for each of the 30 experimental sequences (1) the computer plays the pattern of syllables, (2) a microphone sign asks the player to reproduce the pattern, (3) the score for that individual trial is shown in the screen.[2] Thus, unlike in the training round, a score is shown and, even if mistakes are made, the

[2]The score is a value between 0 and 100 reflecting how similar the sequence typed by the participant is with respect to the target sequence. This value corresponds to the *normalised divergence* measure described in Section 2.4.1.

following sequence is played and the task continues.

After the player has typed in some syllables, if no key is pressed for a period of four seconds, the sequence gets recorded and the following sequence begins. As a way of filtering out obvious slips of memory, sequences of six or less syllables are not registered as a legal response and the same sequence is presented again at a random, later point of the round.

Finally, only the second round of experimental trials is kept for the analyses below; also, these are the trials which are given as an input to the following participant. The first experimental round, hence, serves as a practice phase for our purposes, but subjects are not told so. Regarding the transmission of sequences from one generation to the following one, note that the relative timing or rhythm used by a participant when pressing the keys is discarded. That is, subjects always listen to isochronous sequences, where the time interval between the onset of a syllable and the onset of the following syllable is kept constant at 600 ms, regardless of how the preceding subject had typed in the sequence of syllables.

2.4 Descriptive measures and statistical analyses

For reproducibility purposes, all the responses (i.e. syllable sequences), as well as the R scripts used to perform the analyses, are included at the online Supplementary Information.[3]

2.4.1 Similarity and divergence

Several of the analyses require measuring similarity between sequences, and its counterpart, divergence. For instance, to assess how accurately subjects reproduce a sequence of syllables they have heard, the input and output sequences have to be compared, and a similarity score computed. We do this using a normalised Levenshtein distance metric (Levenshtein 1966). First, the minimum number of insertions, deletions and substitutions to get from sequence A to sequence B is computed. Then, this value is divided by the length (i.e. number of syllables) of the longest sequence (i.e. A or B). The value obtained ranges from 0 (i.e. the sequences are identical), to 1

[3] https://github.com/vdca/hch2.

(i.e. the sequences are maximally divergent). We call this measure the *normalised divergence* or *ndiv*. Its counterpart ($1 - ndiv$) represents the *normalised similarity* measure (*nsim*).

Here, the main use of the normalised similarity is to assess the accuracy with which a subject reproduces an input sequence. The average accuracy for a set of sequences can be interpreted as a measure of *learnability* of the set under consideration; the higher the similarity between input and output, the higher the learnability of the input.

2.4.2 Measures of structure

Following previous studies (Mathy & Feldman 2012, Cornish, Smith & Kirby 2013), we hypothesise that higher accuracy is partly a result of a more structured, less random input. A number of metrics are used to quantify the amount of structure in a set of sequences.

The normalised divergence is used to evaluate the dispersion of a set of sequences, also referred to as *within-set dispersion*. All the sequences in a set are compared in a pairwise manner, and the mean normalised divergence is calculated. In a minimally disperse set, all the sequences would be identical, yielding a normalised divergence score of 0.

We hypothesise that within-set dispersion decreases over generations, i.e. sequences within a set look more alike in later generations, and that this results in a higher overall accuracy. The basis for the dispersion-advantage is that repeated exposure to similar sequences facilitates their recollection. Following this reasoning, if certain syllable patterns (sub-sequences) occur frequently, subjects will identify and recall them with less effort.

We further test this advantage at the level of the individual sequence by analysing the sequence-internal dispersion, also referred to as *within-sequence dispersion*. We slice each sequence at its midpoint and compute the normalised divergence between the two sections. Sequences with lower internal dispersion indicate a higher degree of repeated material and are expected to develop in later generations.

Less disperse sets of sequences are also more compressible from an information theory point of view, which is related to a lower Kolmogorov complexity (Kolmogorov 1963). File compression algorithms rely on chunking in order to represent the same information in a more efficient way.

If the pattern {112233} repeats itself very often in a set of sequences, it can be stored once using a less verbose symbol (e.g. *a*), and then be referred back to every time it is encountered.

In order to obtain a working compressibility measure we use a computer file-compression method. First, we write all the sequences produced by a subject into a file. Then, we compress the file using the Zlib algorithm (Gailly & Adler 2016). Finally, we divide the size of the compressed file by the size of the original file to obtain a compression ratio. Lower values indicate that a file is more compressible because more structure (i.e. more repeated chunks) has been detected by the algorithm.

The emergent regularities in the chains can be a consequence of two general processes: (1) a global bias common to all the participants (due to e.g. general cognition or linguistic bias), and (2) a random bias amplified in a chain-specific way. If the second process is producing at least some of the regularities, the chains should be seen to diverge over time. A way of assessing this is to calculate the evolution of sequence-identifiability. A sequence is identifiable as belonging to its set if the similarity with the sequences in the set (*within-group-nsim*) is higher than the similarity with sequences from the other three chains in the same generation (*across-group-nsim*). This measure of sequence-identifiability, also known as *lineage divergence* (Matthews, Roberts & Caldwell 2012), is formalised as a proportion:

$$\textit{within-group-nsim}/(\textit{within-group-nsim} + \textit{across-group-nsim}) \tag{1}$$

2.4.3 Mixed effects models

So far, we have discussed a number of measures which describe some aspect of the (sets of) sequences produced by each subject. The main hypothesis of the experiment is that some of the variation in these measures can be explained by the subject's generation, i.e. by the position of the subject within its chain of transmission. Specifically, we compare our data to a pair of related null hypotheses: viz. that reproduction accuracy cannot be predicted by (1) participants belonging to later generations, and (2) sequences being more structured. Instead, we expect more structured sequences to be reproduced more accurately, and these to occur more frequently in

later generations. We build mixed effects models to assess the amount of variation in the data due to the effect of generation while controlling for variability across chains (Winter & Wieling 2016), using the statistical package lme4 (Bates et al. 2015).

All the tests follow the same general structure. The outcome of the model is the descriptive measure (e.g. the normalised similarity of an output sequence); the fixed predictor is the generation the measure belongs to; the random predictor is an intercept and slope specific to each of the four chains of transmission.[4] Each of these models is compared to a null model where the generation has been removed but the rest of the predictors are kept unchanged. The fit of each model to the data is assessed through a likelihood ratio test to determine whether the full model bears greater explanatory power, hence showing support for the predictor under consideration.

2.4.4 Interesting patterns

The main analyses involve testing whether structure increases within sets of sequences produced by a participant, or within individual sequences, and we tackle this issue by employing a number of proxies for structure (Section 2.4.2). These analyses only attempt to explain whether the initial randomness of the computer-generated sets is somehow reduced by the transmission process; however, we also want to inspect the concrete regularities in a principled way by searching for emerging syllable patterns.

In order to examine the properties of the emerging structures, we analyse the extent to which each possible ngram of size 2 through 4 is over- or under-represented within each subject. We first create a baseline of expected frequencies consisting of one million sequences generated in the same way the sequences for the initial generation of each chain were generated; i.e. a base sequence containing three instances of each of the four kinds of syllables (*ban*×3, *bi*×3, *ta*×3, *tin*×3) is randomly shuffled one million times. We then compare the frequencies observed in each subject to the baseline frequency.

For each possible ngram and for each of the ten sets of sequences produced in a chain, we

[4] In the cases where the metric refers to the production of a subject as a whole, e.g. dispersion, only a random intercept was included, because the available degrees of freedom did not allow for random slopes.

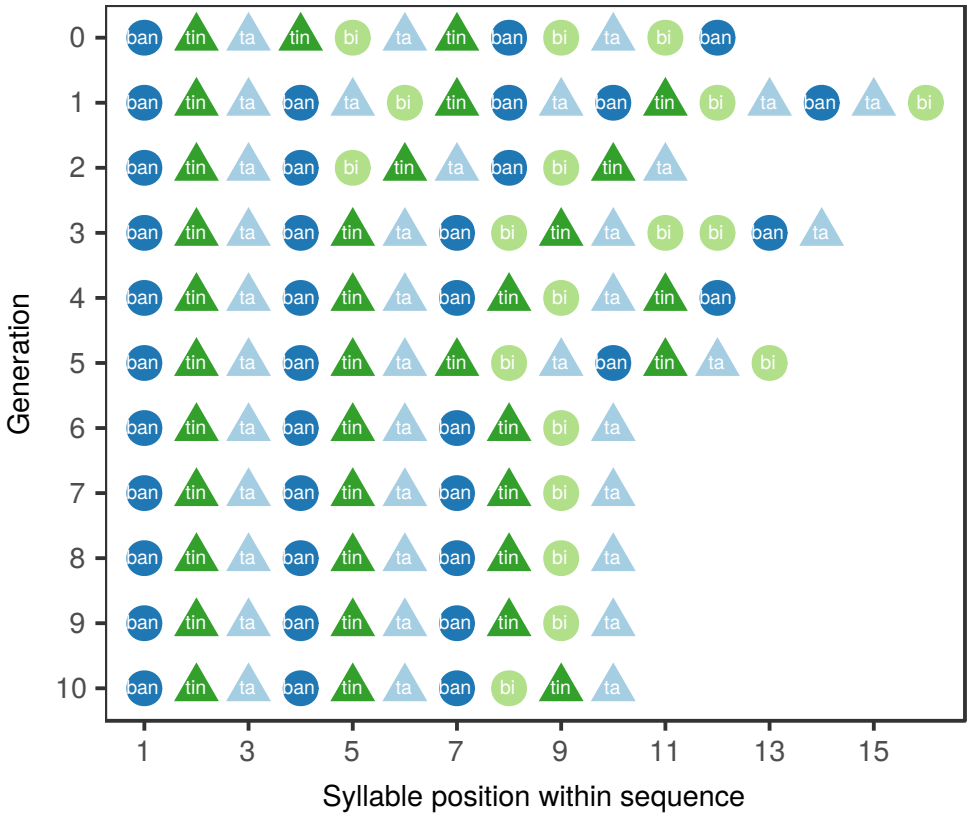


Figure 2: Evolution of sequence lineage 15 in chain 1, from generation 0 (random), to generation 10.

calculate how many sequences contain the ngram.[5] The raw count is divided by the number of sequences in the set: 30 in the experimental subjects, and one million in the random baseline. Additive smoothing is applied in order to avoid zero probabilities (Chen & Goodman 1996). The ngram frequency for the subject is then divided by the baseline frequency to obtain an odds ratio, which we log transform for visualisation purposes. Ngrams with a positive ratio are considered to be over-represented compared to the random baseline.

In order to identify ngrams with a robust increase in popularity across chains, we build a mixed model with generation as a predictor of ngram frequency ratio, and chain as a random effect. To determine ngrams with a *chain-specific* increase, we also run separate linear regression models (one per chain). In both kinds of models, we focus on ngrams where generation is a significant predictor of the ngram becoming over-represented, and the ngram reaches a mean frequency of

[5] We intentionally do not count the total number of instances of the ngram because we want to assess how representative individual patterns are of the set produced by a subject as a whole. This method avoids an ngram's frequency being inflated by its repetition within a single sequence (Conklin 2010).

at least 0.2 by the last generation. We calculate the significance threshold by applying a Bonferroni correction based on the total number of ngrams which can be generated using the available syllables.

Sequence boundaries represent a context with a particular potential of developing fixed or conventional patterns, as illustrated by cadence or rhyme in both music and poetry. We test whether sequence openings become increasingly different from endings by applying the identifiability measure described above. For each subject, we compute an index of how identifiable the first syllable is as belonging to the sequence-opening syllables, as opposed to the syllable-closing syllables.

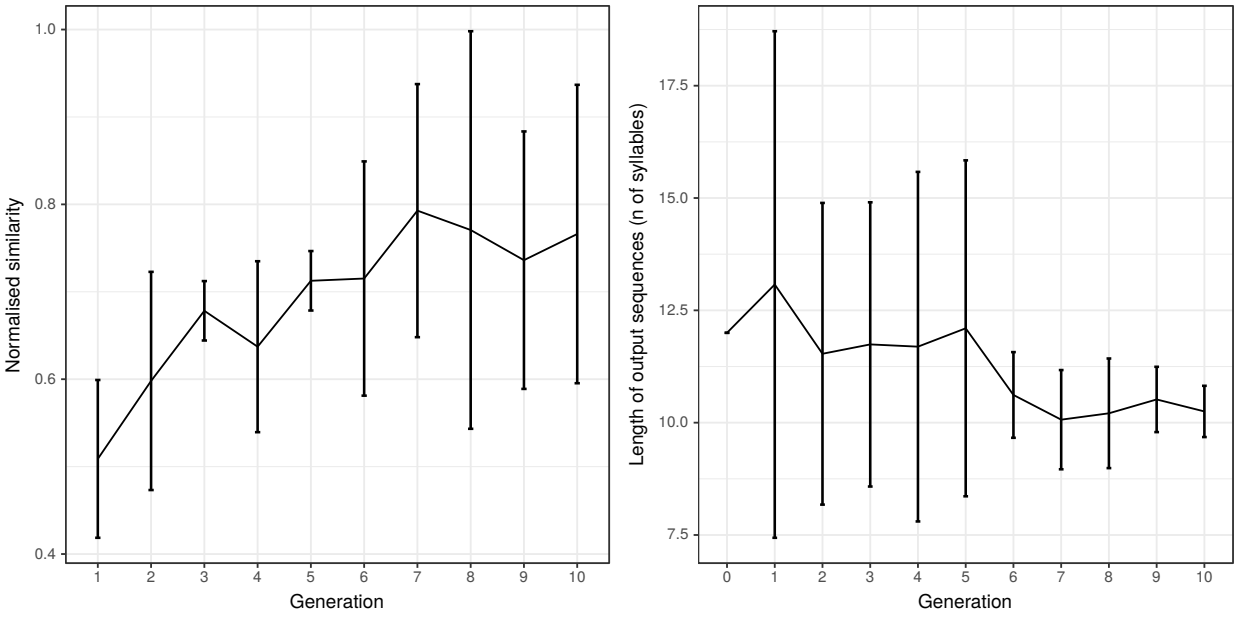
3 Results

3.1 Learnability

Overall, subjects belonging to later generations reproduce their input sequences more accurately (Figure 3a). Hence, we can say that the original random sets of sequences given to the first generation of each chain get more learnable as they get modified by participants. Figure 2 exemplifies this by showing the evolution of a single sequence in the first chain; after the fifth generation, the sequence stabilises and subsequent subjects reproduce it very accurately. Overall, subjects in the initial generation score an average of 0.51, reaching a score of 0.77 by generation 10. When comparing the null model to a model with generation as a fixed predictor (cf. Section 2.4.3), we obtain a statistically significant improvement in prediction (Table 2, Similarity).

3.2 Length

The random sequences given to the first generation are all twelve-syllable long. However, their mean length decreases over time, stabilising at a length of $\sim$10 syllables (Figure 3b). Longer sequences will be typically harder to remember, i.e. length is inversely correlated with learnability ($r = -.28$). Hence, the number of syllables in the input sequences needs to be controlled for



(a) Learnability. **(b)** Length.

Figure 3: Evolution of sequence learnability over generations. Error bars indicate the 95% confidence interval.

in order to assess whether the improvement in accuracy (Figure 3a) is just a function of length (Figure 3b), or depends on some other factor.

A mixed effects model with generation *and* length of the input as predictors of accuracy (and random effects for chain) performs significantly better than a model with only length as a fixed predictor (Table 2, Similarity with length control). Hence, length alone cannot account for the increase in learnability of the sequences.

3.3 Dispersion

If all the sequences in a set look alike, it can become easier for a subject to reproduce them accurately. We hypothesise that sequence *sets* of later generations are more learnable because they have less internal variation. We test whether indeed within-set dispersion decreases over time, correlating with the increasing accuracy shown above. Sets of sequences do become less disperse, but the decrease is robust only when going from the initial random state to the first generation of participants; later generations keep a steady dispersion measure of $\sim$0.5 (Figure 4a).

	Model	Estimate	Std. Error	χ^2	Pr ($> \chi^2$)
1	Similarity	0.0255	0.00771	5.28	0.0215
2	Similarity with length control	0.023	0.00708	5.17	0.023
3	Length	-0.249	0.0952	3.99	0.0459
4	Set dispersion	-0.00333	0.00155	4.36	0.0367
5	Sequence dispersion	-0.0153	0.00188	65.3	6.56e-16
6	Compression	-0.00263	0.00092	7.45	0.00633
7	Identifiability	0.000545	0.00019	8.16	0.00428
8	Boundary identifiability	0.0209	0.00204	101	9.71e-24

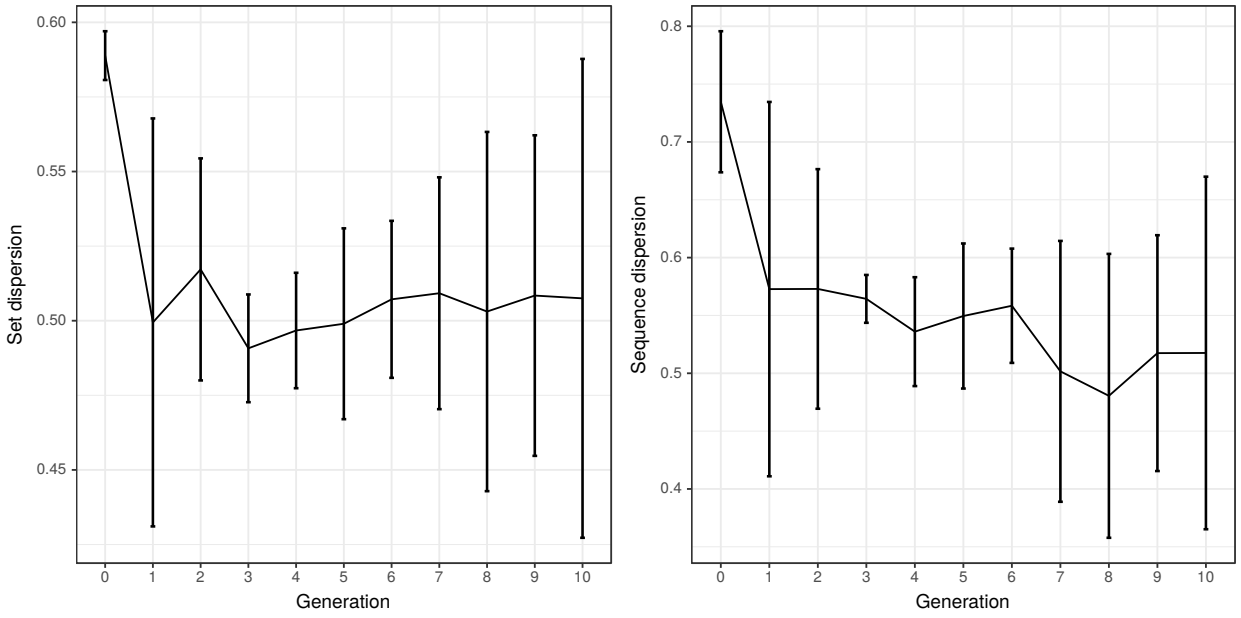
Table 2: Results of the full mixed models compared to the correspondent null model where the predictor of interest (generation) has been removed.

Adding generation as a predictor of set dispersion significantly improves the explanatory power of the null mixed effects model (Table 2, Set dispersion). Nevertheless, the effect disappears if the initial random state (generation 0) is removed, indicating that the decrease takes place as soon as a human subject intervenes, but is not amplified as a function of iterated learning (a variant of the models in Table 2 excluding the initial generation is reproduced in Table 5 of the Supplementary Information).

The decrease of sequence-internal dispersion, however, shows a more robust cumulative effect over generations, as shown in Figure 4b. The first and second halves of sequences resemble each other more in later generations, indicating that each subject increases the amount of repetition of sequence-internal patterns. In Table 5 of the Supplementary Information we confirm that generation remains a significant predictor of the decrease in sequence-internal dispersion even when the computer-produced generation is excluded from the analysis. This means that within-sequence dispersion does not decrease categorically, but shows a cumulative effect.

3.4 Compression

The evolution of compressibility resembles that of within-set dispersion: human-produced sets of sequences are more compressible than the randomly generated ones. However, once the compression ratio drops with the first participant, it does not further decrease in a robust way across chains (Figure 5a).



(a) Within-set dispersion.

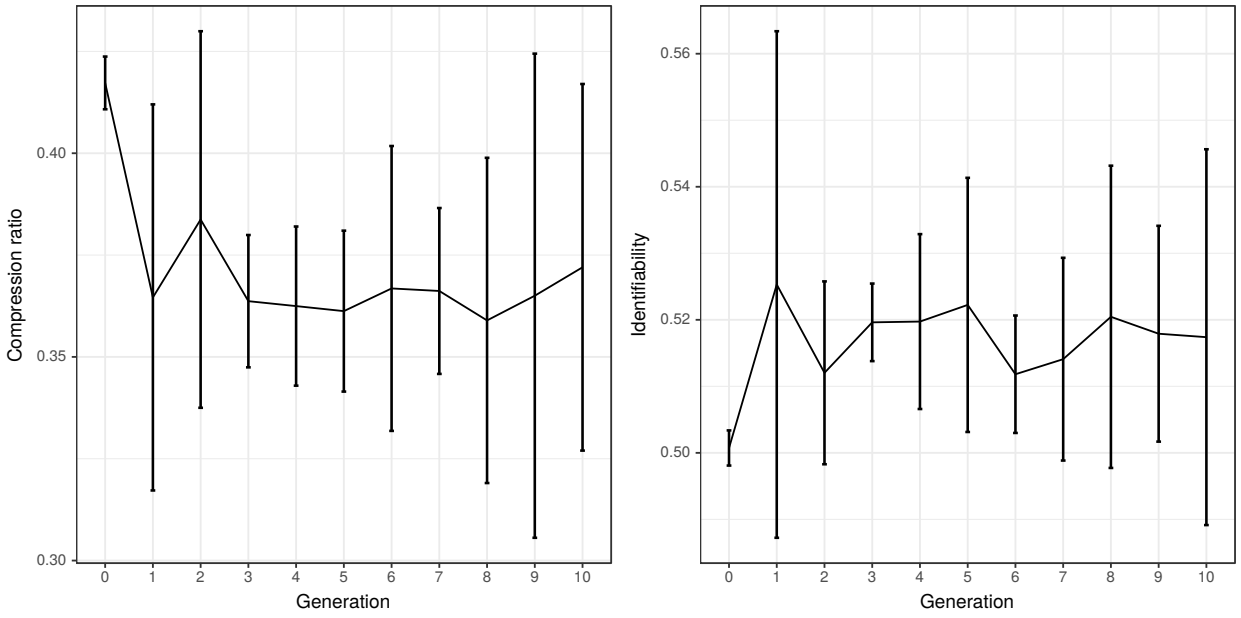
(b) Within-sequence dispersion.

Figure 4: Evolution of dispersion within sequence sets and within individual sequences, over generations. Error bars indicate the 95% confidence interval.

Adding generation as a predictor of set compressibility significantly improves the explanatory power of the null mixed effects model when including the computer-generation (Table 2, Compression), but not when excluding it (Table 5, Supplementary Information). If we compare the compression ratio of the random sets, to those produced by the participants, we obtain a mean difference of 0.051 ($t = 12.676, p = 5.027e - 13$).

3.5 Identifiability

Overall, sequences from later generations are more identifiable as belonging to their chain (Figure 5b). This suggests that at least some of the strategies by which chains develop structure are chain-specific. The average identifiability index for the initial random generation approximates 0.5; i.e. within-group similarity is as high as across-group similarity. Some of the increase in identifiability can be attributed to the effect of generation (Table 2, Identifiability).



(a) Compression.

(b) Identifiability.

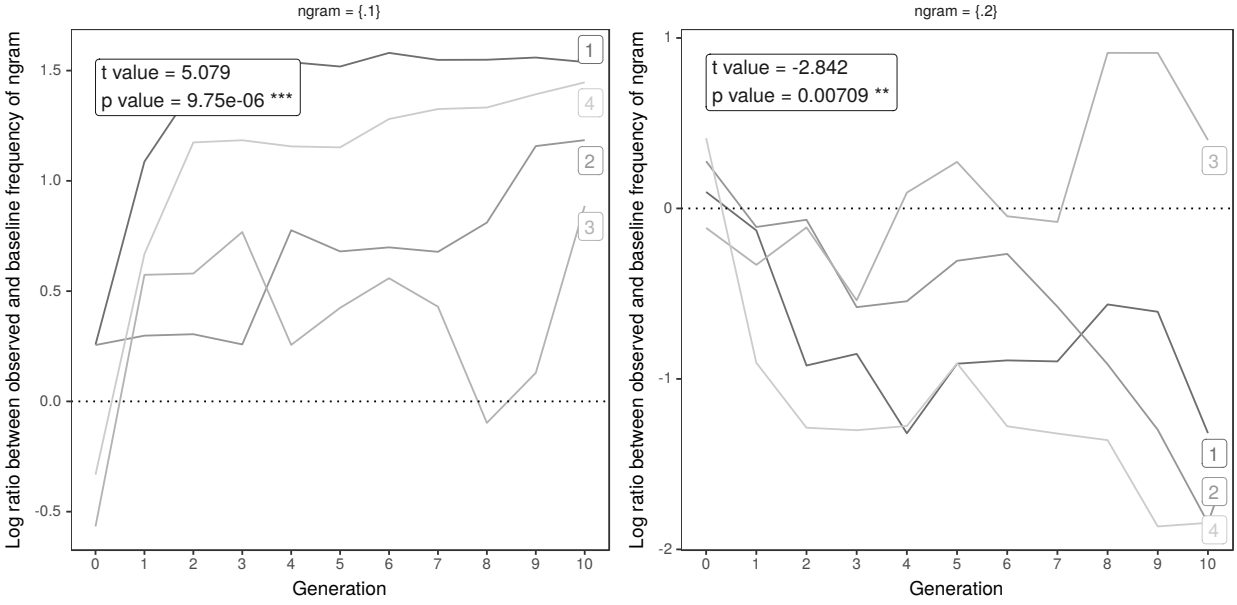
Figure 5: Evolution of sequence compression and identifiability over generations. Error bars indicate the 95% confidence interval.

3.6 Interesting patterns

For convenience during the analyses, syllables are encoded with the integers {1, 2, 3, 4} corresponding to the syllables {*ban, bi, ta, tin*}. Besides, the start and end of sequences are encoded by a special boundary symbol {.}, so a bigram like {.1} represents the single syllable *ban* opening a sequence. When referring to sequences of syllabic features (i.e. sequences of onsets, nuclei or codas), we use the corresponding orthographic symbols, as shown in Table 1; hence, the bigram {bb} represents a succession of two syllables sharing the same *b* onset (*ban* or *bi*).

We have investigated distinctive patterns at four different levels: at the level of the syllable, and at three sub-syllabic dimensions: onset, nucleus and coda (see Section 2.2 for details on the syllabic structure). Table 3 displays the ngrams of size 2, 3 and 4 which are increasingly over-represented over generations.[6] Note that all of these contain a boundary symbol, meaning that they belong to the sequence-initial or sequence-final contexts. Figure 6a illustrates the increase in popularity for the opening pattern {.1} (*ban*) over generations.

[6]The Bonferroni-corrected significance thresholds differ for syllabic patterns ($\alpha = 9.92e-05$) and sub-syllabic feature patterns ($\alpha = 8.93e-04$); see Section 2.4.4.



(a) Global increase of {.1} (*ban*). **(b)** Idiosyncratic increase of {.2} (*bi*).

Figure 6: Frequency of the sequence-opening patterns {.1} (*ban*) and {.2} (*bi*) relative to the baseline. A log ratio above 0 means that the pattern is over-represented in that generation. Each line in the plot represents how the relative frequency of the pattern develops within each of the four transmission chains. The summary statistics indicate the global trend for that pattern across all four chains.

The emergent pervasiveness of this opening can be observed by comparing the first and last generations of chain 1 (Figures 11 and 12 of the Supplementary Information). More generally, by visually inspecting the sets of sequences plotted as phase-space diagrams (Figure 13 of the Supplementary Information), we can notice how the initial generations produce all syllable combinations with a similar frequency, while later participants persist on a few ngrams, reflected in the emerging geometric patterns (Ravignani 2017).

In some instances, an individual chain can develop a preference for an ngram, while the general trend of the other three chains is to gradually disprefer the pattern. These chain-specific patterns are listed in Table 4, and Figure 6b shows the evolution of a sample ngram.

Chain 3, for instance, develops a preference for sequences starting with the syllable *bi* or two light syllables (i.e. syllables without a final [n]), while the general tendency is to decrease these openings (Table 4) in favour of a pattern with a heavy syllable like *ban* (Table 3). On the sequence-final context, there is a global tendency to end with the nuclei pattern {ia.}, but chain 4 idiosyncratically favours the ending {iaa.}.

	Feature	Context	Pattern	t statistic	Pr ($> t$)
1	syllable	initial	.1	5.08	9.75e-06
2	onset	initial	.b	6.82	3.82e-08
3	onset	initial	.bb	4.36	9.13e-05
4	onset	final	t.	3.76	0.000561
5	nucleus	initial	.aii	4.45	7.05e-05
6	coda	initial	.n-n	3.63	0.000817
7	coda	final	-.	4.16	0.000169

Table 3: Patterns which increase robustly across generations. The numeric codes follow the syllables in alphabetical order: 1 = *ban*, 2 = *bi*, 3 = *ta*, 4 = *tin*. Sequences of onsets, nuclei or codas are indicated with their corresponding letters: {*b, t*}, {*a, i*}, and {*n,* -}, where the hyphen represents a coda-less syllable.

	Context	Chain	Pattern	Chain's t	Global t	Pr ($> t$)
1	internal	4	134	2.48	-2.42	0.0203
2	initial	3	.2	3.12	-2.84	0.00709
3	initial	2	.4	2.33	-2.29	0.0273
4	initial	3	.--	2.27	-2.1	0.0423
5	initial	2	.41	5.68	-2.19	0.0349
6	final	4	iaa.	2.44	-2.76	0.00884

Table 4: Ngrams with a significant increase in preference in one chain, coupled with a global trend to disprefer the pattern. Chain-specific t values are the result of linear regressions on a single chain; global t values are computed with a mixed model including all chains.

As the previous results reveal, distinctive patterns tend to emerge at the boundaries of sequences, but the over-represented opening syllables seem to differ from the closing ones. By running the identifiability analysis (Section 2.4.2) on the opening and closing unigrams, we can test whether sequence-initial syllables progressively become more similar to each other, and more unlike the sequence-final syllables. Figure 7 indicates that, indeed, opening and closing syllables become increasingly polarised as a function of generation. Opening syllables start off being indistinguishable from closing syllables (mean identifiability at generation 0 = 0.49), and exhibit a steady divergence over generations which proves robust across all four chains (Figure 7 and Table 2, Boundary identifiability).

4 Discussion

The starting point of the experiment are sequences which randomly alternate the syllables *ban, bi, ta, tin*. After a process of iterated learning involving four chains of transmission and ten generations of participants, the original sequences become (1) easier to recall, (2) shorter, (3) more structured. As suggested by the analysis of significant patterns and sequence identifiability, some of the emerging regularities are common to all four chains, while others are chain-specific.

Tendencies which emerge across the board are most likely attributed to biases shared by all the participants. These biases can be related to (1) basic aspects of human cognition involved in sequence perception and recall, (2) phonological properties of the Dutch language, which all participants speak natively.

Regarding basic cognitive biases, we can highlight that all chains become consistently more compressible and less disperse than the initial random sets. This is the result of a number of syllabic patterns gaining popularity at the expense of others. Hence, we can infer that sequences are not processed as unitary entities; instead, sub-patterns within sequences are recognised and reproduced, leading to an increasingly uneven distribution of ngrams.

The data suggest that participants are engaging in a chunking strategy to deal with the task at hand, i.e. instead of processing syllables individually, they group them into chunks, just the way people memorise e.g. telephone numbers (for an overview of chunking as a cognitive process, see Gobet et al. 2001). The participants, essentially, are asked to remember a sequence too long to fit in working memory, and then reproduce it. Working memory can hold around four items (Mathy & Feldman 2012, Chen & Cowan 2005), yet, crucially, items need not be unitary but can contain further items within themselves. This effectively can expand our working memory capacity to a span of tens of items (Ericsson, Chase & Faloon 1980). We apply chunking strategies unconsciously, and even 14-month-old infants combine chunks into super-chunks under experimental conditions in order to expand the limits of working memory (Rosenberg & Feigenson 2013).

Given this readiness to divide temporal sequences, it is unsurprising that human music and poetry rely heavily on segmenting and repeating motifs (cf. Tierney, Russo & Patel 2011, Rubin

1995). Moreover, this aspect of cognition is not restricted to humans; a number of bird species (e.g. bullfinches, nightingales) learn and reproduce sound sequences, and are shown to engage in chunking too (Nicolai et al. 2014).

Besides chunking being pervasive within human cognition, it also manifests some particularities within the domain of verse templates. Everyday speech is also segmented into e.g. intonational phrases, phonological words and syllables; in contrast with everyday speech, however, verse sections and subsections typically display an added level of regularity or numeric control. When speaking, the number of syllables or accents in a sentence is not fixed, but it is when creating e.g. a sonnet. Regarding bigger chunks, many verse traditions produce songs or poems by generating pairs of lines, linked together by a common beginning (Kara 1970 for Mongolian), a common rhyme (Hanson 2006 for English) or a semantic parallel (Fox 2014 for Rotinese). By introducing this kind of regularities, verse constituents prove easier to recall compared to similar segments of non-verse speech (Rubin 1995 for an extensive overview of the effects of rhyme). After ten generations, the sequences in the current experiment show an increased regularisation of the boundary syllables (Figure 10), which alters the initial sequences to make them more similar to lines of verse, than to speech utterances. More precisely, the lack of semantic content in our sequences of syllables makes this material more comparable to the patterns of mnemonic syllables illustrated under Example 2.

Regarding the general chunking bias, unlike in the previous experiment (Cornish, Smith & Kirby 2013), we do not observe a cumulative effect on within-set dispersion and compressibility: the effect appears in the very first participant of a chain, and then remains stable along the following generations. Given that both experiments only run over ten generations, we do not know whether the dispersion, for instance, would continue to decrease or remain stable in further generations. Alternatively, a strong-enough bias can be observed within a single generation, as shown in the case of writing systems (Morin 2018).

This earlier decrease and stabilisation compared to the colour experiment may stem from a greater difficulty in the task. This can force the participants to focus on less detail, effectively boosting the chunking effect. Crucially, the experiments differ in the modality used for stimuli

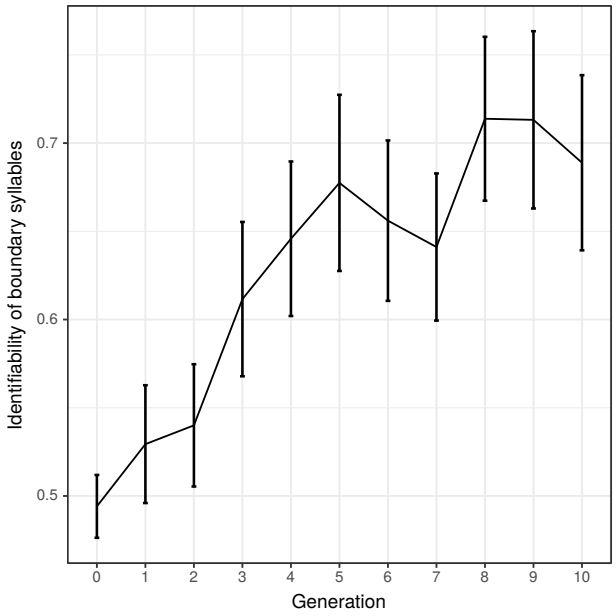


Figure 7: Identifiability of syllables as sequence openers or closers. All four chains averaged. Error bars indicate 95% confidence interval.

presentation (visual vs auditory), and the input method used by the subjects (visual cues vs no cues), which can make the task more challenging. Nonetheless, the increase in reproduction accuracy (i.e. learnability, Figure 3a) is gradual, indicating individual participants do introduce some facilitation effect, which then accumulates across generations. Even if a sudden increase in structure occurs specifically within the first generation, the accuracy improvement shown by subsequent generations demonstrates that at least some cumulative effect typical of cultural evolution plays a role in modifying the syllable sequences.

A structural measure where a cumulative effect does take place is the gradual divergence of opening and closing syllables (Figure 7); syllables get specialised in all chains by being increasingly employed either as sequence openers or as sequence closers. On the one hand, the fact that a specialisation takes place can still be driven by some aspect of general cognition; on the other hand, the specific macro-phonotactics of which syllables are preferred on the left or right boundary are arguably language-specific.

In this experiment, we hypothesise that a Dutch bias drives the emerging preference for starting sequences with the syllable *ban* (more generally, a heavy syllable), and ending sequences with the syllable *ta* (more generally, a light syllable). This can be formulated in terms of an at-

traction for e.g. starting with a heavy syllable, or inversely in terms of sequence-opening light syllables being systematically mistaken for heavy ones. One kind of support for this bias comes from the properties of the Dutch lexicon: heavy syllables attract stress, and most content words have initial stress (van Heuven & Hagman 1988). Another kind of evidence is provided by acquisition data: children learning Dutch produce mostly disyllabic words starting with a stressed syllable (Fikkert 1994). This trochaic bias has been described for other Germanic languages (Pater 1997), but some non-Germanic languages like Hebrew or Portuguese show, respectively, either no preference between iambic or trochaic disyllables (Santos 2003), or a preference in the opposite direction (Segal & Kishon-Rabin 2012). Follow-up experiments can exploit these differences to set apart general biases from those related to particular phonological systems.

In this respect, it is crucial to re-run the current experimental design with speakers of other languages in order to confirm plausible language-specific effects such as the preferential opening of sequences with a heavy syllable. Further, such connection between the phonological properties of a language and the emerging pseudo-verse templates would be in agreement with language-based theories of poetic metre (e.g. Hanson & Kiparsky 1996, Golston & Riad 2000), according to which "the possible versification systems for a language" can be derived "from its phonology" (Hanson & Kiparsky 1996:288).

So far, we have only discussed tendencies which consistently appear in all four chains. Certain patterns, nevertheless, gain preference in a single chain, while the other three follow the opposite direction (Table 4). We can refer to these as *arbitrary preferences*, since, if they were determined by general cognitive or linguistic biases, all four chains should have developed them. Instead, we can think of these biases as pressures which shape the pool of possible patterns in a particular direction. The pressures get amplified in the process of cultural evolution, but can not explain in a deterministic way the exact patterns which will prevail. Even in a non-creative task as the one we present here, individual subjects move the syllable sequences in idiosyncratic directions, some of which are picked up and amplified by further generations.

The emergence and evolution of verse patterns in the world's languages can be conceptualised in this way. Among the virtually infinite combinations of e.g. syllables, phonological features,

or drum patterns, our shared cognitive and linguistic background creates a biased baseline. Only a subset of these combinations is used as the basis to create songs and poems, and this subset is continually innovated and filtered in the process of cultural evolution.

5 Conclusion

In the present study we show that random sets of syllables develop an increasingly systematic structure through iterated learning, where individuals try to reproduce the stimuli produced by the predecessor. Cognitive, linguistic and other subject-specific biases shape the sequences in particular ways, while the learn-and-reproduce procedure allows small biases to have a cumulative effect over generations. Both the emergent features and the iterated learning mechanism resemble aspects of versification pervasive in human societies, making this paradigm a suitable one to model the emergence of verse templates.

Acknowledgements

We thank Ruby Sleeman for patiently recording the stimuli, and the participants for their time. The article has benefited from insightful comments by Marc van Oostendorp, three anonymous reviewers, and valuable editorial work by Seán Roberts. The usual disclaimers apply.

Funding

This research was made possible thanks to the project *Knowledge and culture* (Horizon grant 317-70-010) funded by NWO (Dutch Organisation for Scientific Research).

References

Bates, D. et al. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48.

Boyd, R. & P. J. Richerson. 1985. *Culture and the evolutionary process*. Chicago: University of Chicago Press.

Brown, S. & J. Jordania. 2011. Universals in the world's musics. *Psychology of Music* 41(2). 229–248.

Chen, S. F. & J. Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on association for computational linguistics* (ACL '96), 310–318. Stroudsburg, PA, USA: Association for Computational Linguistics.

Chen, Z. & N. Cowan. 2005. Chunk limits and length limits in immediate recall: a reconciliation. *Journal of experimental psychology, learning, memory, and cognition* 31(6). 1235–1249.

Clayton, M. 2000. *Time in Indian music: rhythm, metre, and form in North Indian rāg performance*. Oxford University Press.

Conklin, D. 2010. Discovery of distinctive patterns in music. *Intelligent Data Analysis* 14(5). 547–554.

Cornish, H., K. Smith & S. Kirby. 2013. Systems from sequences: an iterated learning account of the emergence of systematic structure in a non-linguistic task. In M. Knauff et al. (eds.), *Proceedings of the 35th annual conference of the cognitive science society*, 340–345. Austin (TX): Cognitive Science Society.

de Groot, A. W. 1936. De structuur van het vers. *De Nieuwe Taalgids* 30. 197–212.

deCastro-Arrazola, V., P. van Kranenburg & B. Janssen. 2015. Computational textsetting analysis of Dutch folk songs. In O. Adam & D. Cazau (eds.), *Proceedings of the 5th international workshop on folk music analysis*, 51–55. Paris (France): Association Dirac.

Dell, F. & M. Elmedlaoui. 2008. *Poetic meter and musical form in Tashlhiyt Berber songs*. Köln: Rüdiger Köppe.

Ericsson, K., W. G. Chase & S. Faloon. 1980. Acquisition of a memory skill. *Science* 208(4448). 1181–1182.

Euba, A. 1990. *Yorùbá drumming: the dùndún tradition*. Vol. 21. E. Breitinger, Bayreuth University.

Fabb, N. 1997. *Linguistics and literature*. Blackwell Publishers.

Fabb, N. & M. Halle. 2008. *Meter in poetry: a new theory. With a chapter on Southern Romance meters by Carlos Piera*. Cambridge University Press.

Fikkert, P. 1994. *On the acquisition of prosodic structure*. Holland Institute of Generative Linguistics.

Fox, J. J. 2014. *Explorations in semantic parallelism*. ANU Press.

Gailly, J.-l. & M. Adler. 2016. Zlib compression library.

Gobet, F. et al. 2001. Chunking mechanisms in human learning. *Trends in cognitive sciences* 5(6). 236–243.

Golston, C. & T. Riad. 2000. The phonology of Classical Greek meter. *Linguistics* 38(1). 99–167.

Gordon, M. K. 2006. *Syllable weight: phonetics, phonology, typology*. New York: Routledge.

Hanson, K. 2006. Rhyme. In K. Brown (ed.), *Encyclopedia of language and linguistics, 2nd ed.* 605–616. Elsevier.

Hanson, K. & P. Kiparsky. 1996. A parametric theory of poetic meter. *Language* 72(2). 287–335.

Heuven, V. J. van & M. de Jonge. 2011. Spectral and temporal reduction as stress cues in Dutch. *Phonetica* 68(3). 120–132.

Jouad, H. & B. Lortat-Jacob. 1982. Les modèles métriques dans la poésie de tradition orale et leur traitement musical. *Revue de musicologie* 67(1-2). 174–197.

Kara, G. 1970. *Chants d'un barde mongol*. Budapest: Akadémiai Kiadó.

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.

Knight, R. 1984. Music in Africa: the Manding contexts. *Performance practice: Ethnomusicological perspectives* (12). 53.

Kolmogorov, A. N. 1963. On tables of random numbers. *Sankhyā: The Indian Journal of Statistics, Series A*. 369–376.

Lerdahl, F. & R. Jackendoff. 1983. *A generative theory of tonal music*. Massachusetts Institute of Technology.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics–Doklady* 10(8). 707–710.

Lomax, A. 1976. *Cantometrics: an approach to the anthropology of music: audiocassettes and a handbook*. Berkeley: University of California Extension Media Center.

Mathy, F. & J. Feldman. 2012. What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition* 122(3). 346–362.

Matthews, C., G. Roberts & C. A. Caldwell. 2012. Opportunity to assimilate and pressure to discriminate can generate cultural divergence in the laboratory. *Evolution and Human Behavior* 33(6). 759–770.

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.

Morin, O. 2016. *How traditions live and die*. New York: Oxford University Press.

Morin, O. 2018. Spontaneous emergence of legibility in writing systems: the case of orientation anisotropy. *Cognitive science* 42(2). 664–677.

Moyle, R. M. 1979. *Songs of the Pintupi: musical life in a central Australian society*. Australian institute of aboriginal studies.

Nicolai, J. et al. 2014. Human melody singing by bullfinches (*Pyrrhula pyrrula*) gives hints about a cognitive note sequence processing. *Animal cognition* 17(1). 143–155.

Opie, I. & P. Opie. 1959. *The lore and language of schoolchildren*. Oxford: Oxford University Press.

Pater, J. 1997. Minimal violation and phonological development. *Language Acquisition* 6(3). 201–253.

Ravignani, A. 2017. Visualizing and interpreting rhythmic patterns using phase space plots. *Music Perception: An Interdisciplinary Journal* 34(5). 557–568.

Ravignani, A., T. Delgado & S. Kirby. 2016. Musical evolution in the lab exhibits rhythmic universals. *Nature Human Behaviour* 1. 0007.

Reina, R. 2013. *Karnatic rhythmical structures as a source for new thinking in Western music*. Amsterdamse Hogeschool voor de Kunsten, Conservatorium van Amsterdam Doctoral dissertation.

Rosenberg, R. D. & L. Feigenson. 2013. Infants hierarchically organize memory representations. *Developmental science* 16(4). 610–621.

Rubin, D.-C. 1995. *Memory in oral traditions: the cognitive psychology of epic, ballads, and counting-out rhymes*. Oxford University Press.

Santos, R. S. 2003. Bootstrapping in the acquisition of word stress in Brazilian Portuguese. *Journal of Portuguese Linguistics* 2(1). 93–114.

Savage, P. E. et al. 2015. Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences* 112(29). 8987–8992.

Segal, O. & L. Kishon-Rabin. 2012. Evidence for language-specific influence on the preference of stress patterns in infants learning an iambic language (Hebrew). *Journal of Speech, Language, and Hearing Research* 55(5). 1329–1341.

Stone, R.-M. 1985. In search of time in African music. *Music Theory Scpectrum* 7. 139–148.

Tamariz, M. 2017. Experimental studies on the cultural evolution of language. *Annual Review of Linguistics* 3. 389–407.

Tierney, A. T., F. A. Russo & A. D. Patel. 2011. The motor origins of human and avian song structure. *Proceedings of the National Academy of Sciences* 108(37). 15510–15515.

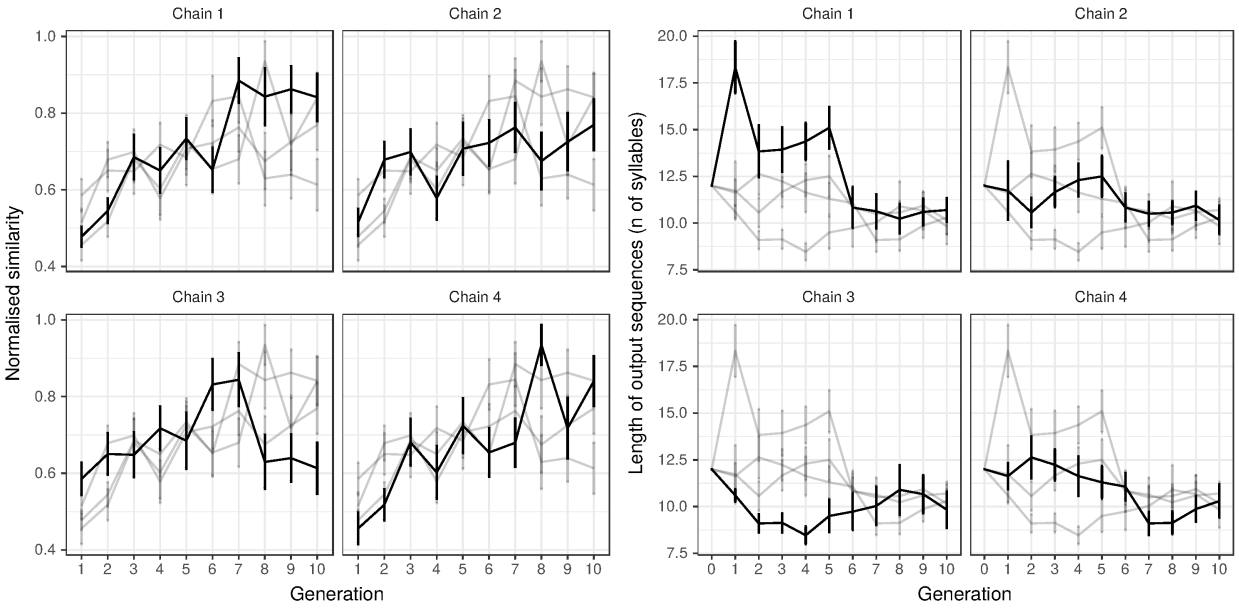
Van der Hulst, H. 1984. *Syllable structure and stress in Dutch*. Foris.

van Heuven, V., P. J. Hagman, et al. 1988. Lexical statistics and spoken word recognition in Dutch. *Linguistics in the Netherlands*. 59–68.

Winter, B. & M. Wieling. 2016. How to analyze linguistic change using mixed models, growth curve analysis and generalized additive modeling. *Journal of Language Evolution* 1(1). 7–18.

Supplementary Information

In the following figures, we plot the evolution of the different metrics for each chain separately. This enables the tracking of global and local trends with more detail. Note that the error bars of the main-text plots indicated the 95% confidence interval based on the average of all 4 chains, while the following intervals are based on 30-sequence sets of single chains. Each of the plots under Figures 8, 9 and 10 contains four facets, each corresponding to one chain of transmission. For each of these facets, the chain of interest is highlighted, while the lines for the other three chains are greyed out and shown as reference.

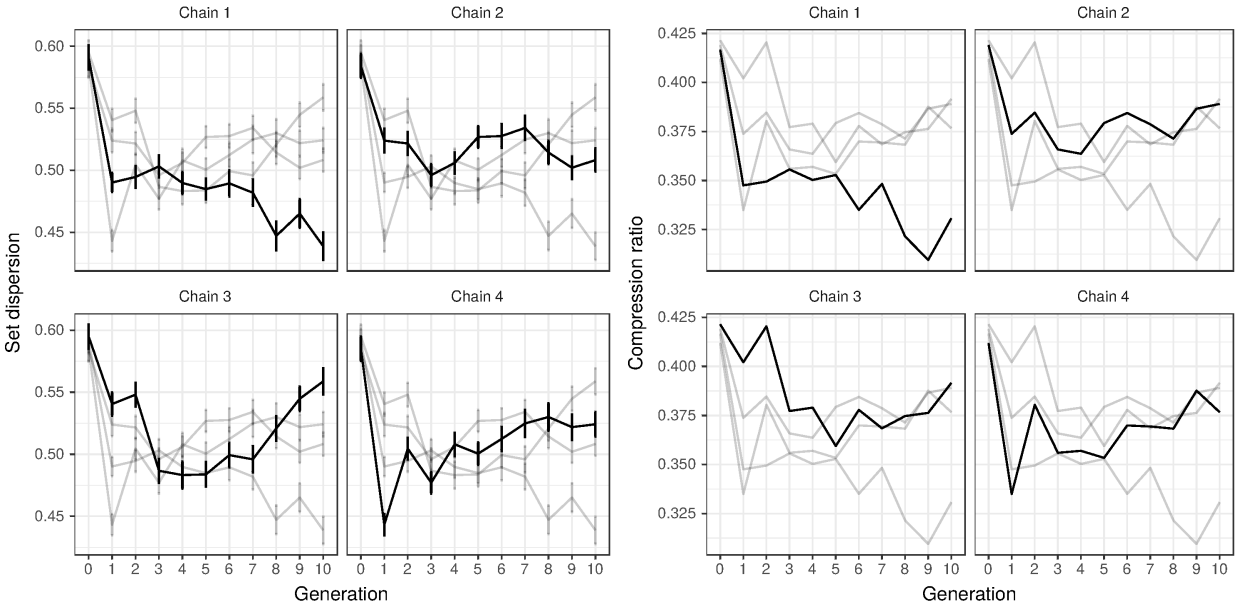


(a) Learnability. **(b)** Length.

Figure 8: Evolution of sequence learnability and length over generations. Error bars indicate the 95% confidence interval.

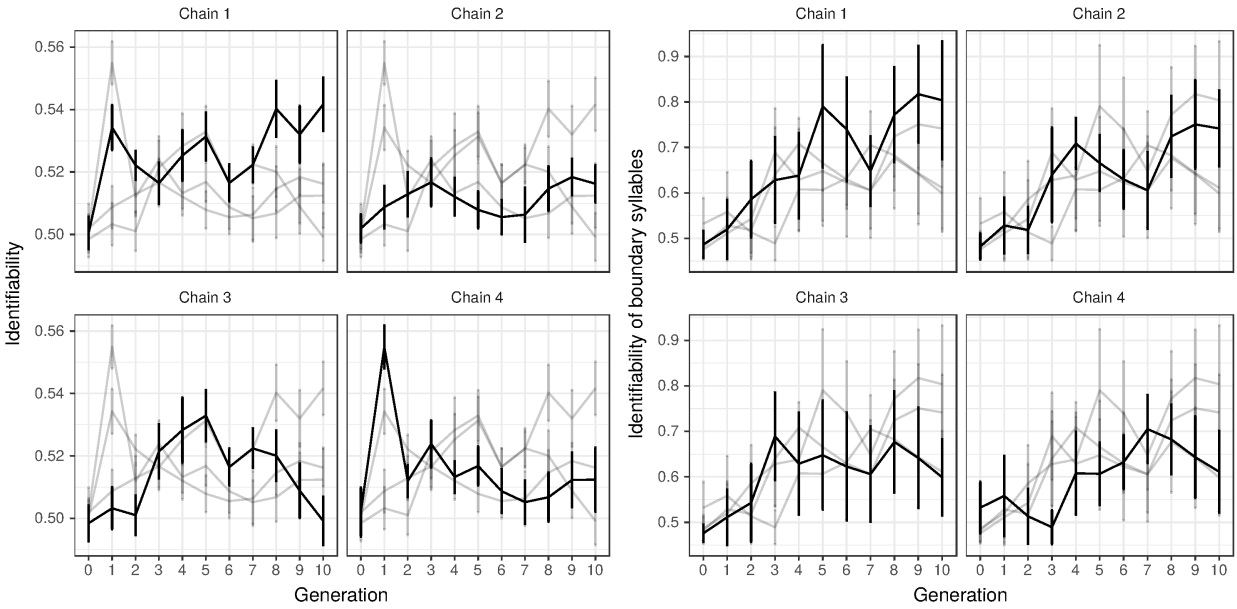
	Model	Estimate	Std. Error	χ^2	Pr ($> \chi^2$)
1	Length	-0.282	0.158	2.34	0.126
2	Set dispersion	0.000718	0.00129	0.311	0.577
3	Sequence dispersion	-0.00848	0.00217	15.1	0.000102
4	Compression	-0.000432	0.000803	0.288	0.591
5	Identifiability	-0.000323	0.000219	2.18	0.14
6	Boundary identifiability	0.0189	0.00245	58.4	2.15e-14

Table 5: Results of the full mixed models compared to the correspondent null model where the predictor of interest (generation) has been removed. These models only include human-generated data, i.e. the initial generation has been excluded.



(a) Dispersion. **(b)** Compression.

Figure 9: Evolution of sequence dispersion and compression over generations. Error bars indicate the 95% confidence interval.



(a) Sequence identifiability.

(b) Boundary identifiability.

Figure 10: Evolution of sequence identifiability, and identifiability of syllables as sequence openers or closers, over generations. Error bars indicate the 95% confidence interval.

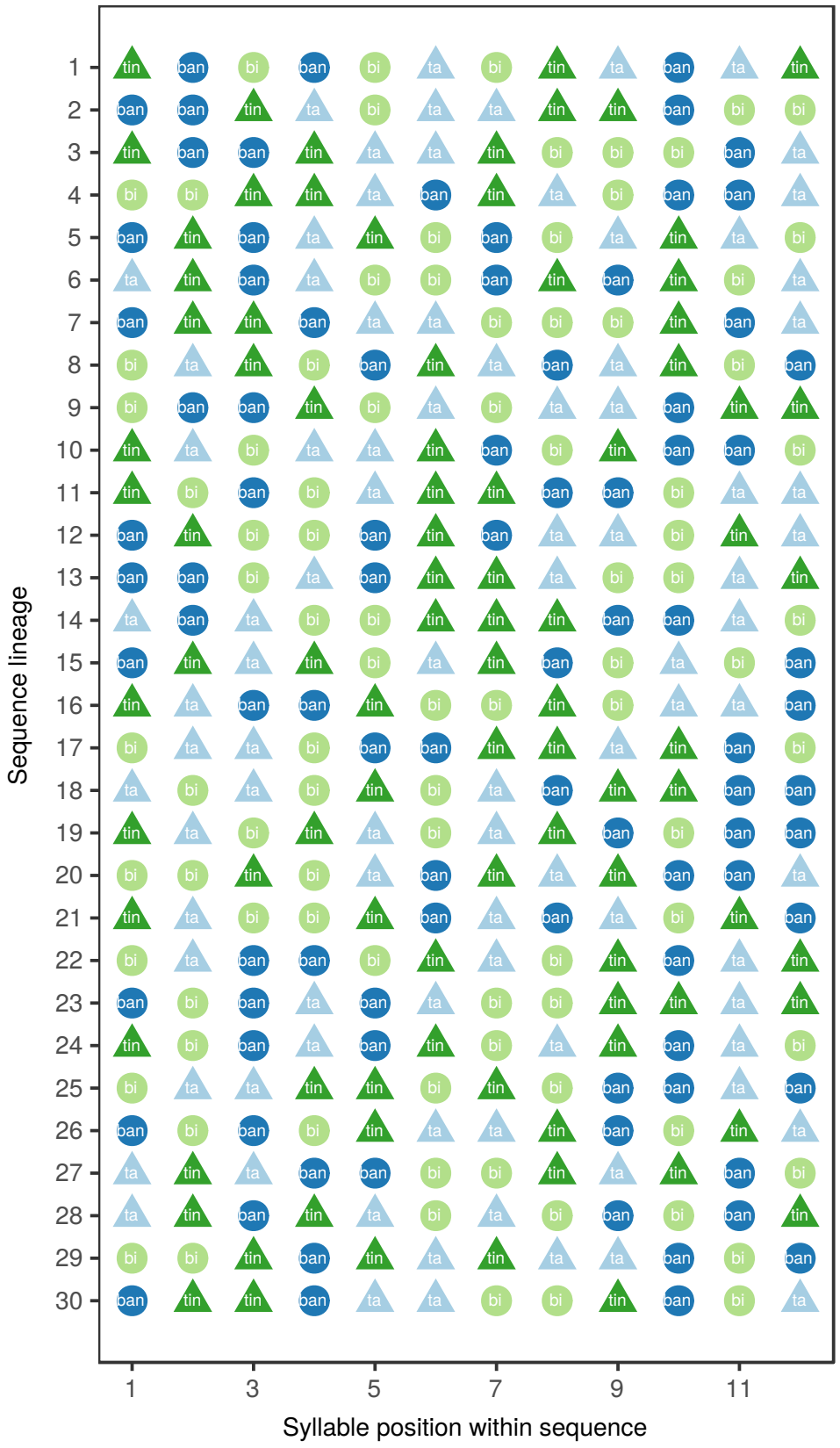


Figure 11: All 30 computer-generated sequences (i.e. initial state) of chain 1.

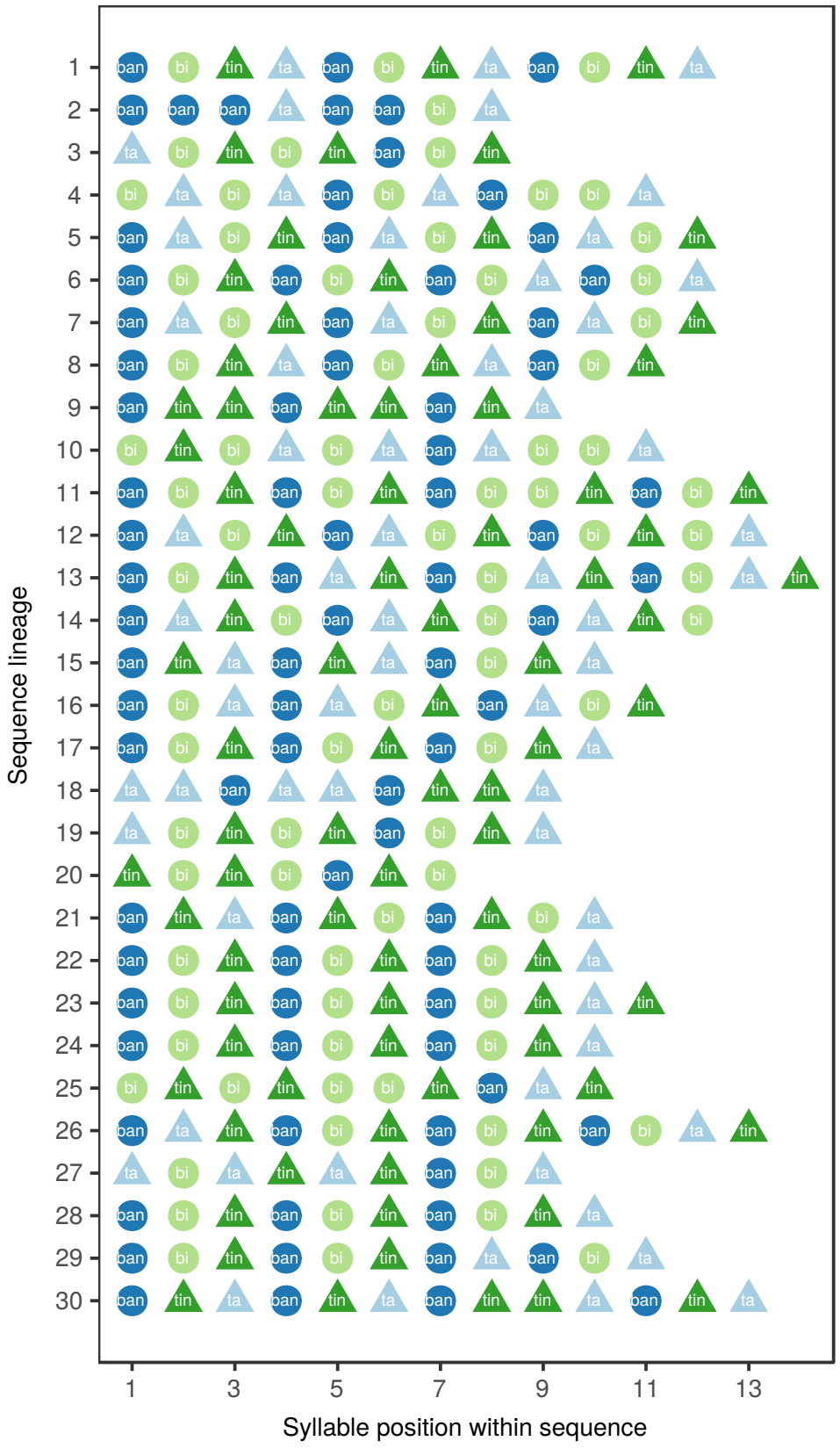


Figure 12: All 30 sequences produced by the last subject in chain 1.

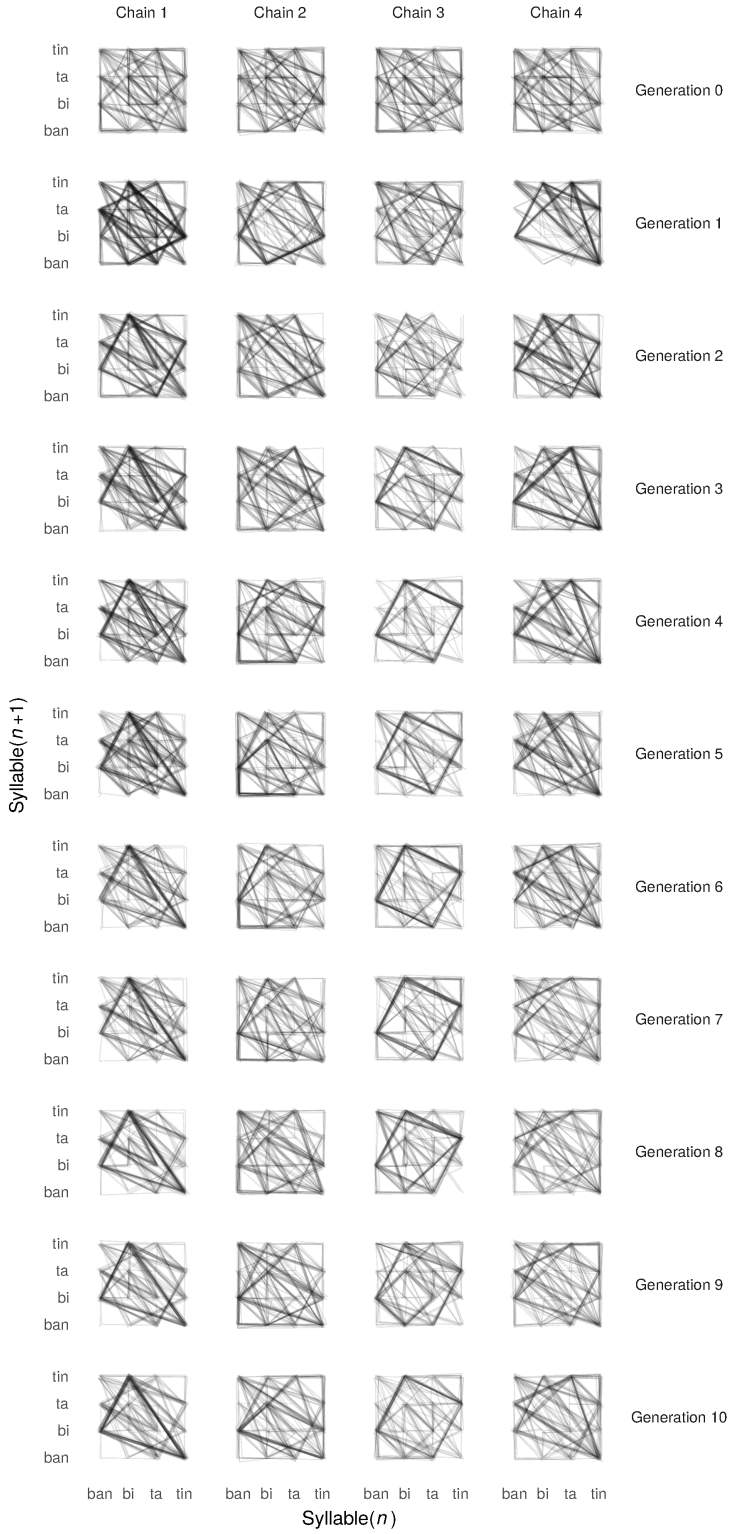


Figure 13: Visualisation of all the sequences as phase-space diagrams (Ravignani 2017). Each syllable receives a coordinate based on its own value (x-axis), and the value of the following syllable (y-axis). Consecutive syllables are connected with a line, and emerging geometric patterns represent often-visited syllabic ngrams.