Feature economy vs. logical complexity in phonological pattern learning[1]

Klaas T. Seinhorst[2]
Amsterdam Center for Language and Communication,
University of Amsterdam

Complexity has been linked to ease of learning. This article explores the roles of two measures of complexity – feature economy and logical complexity – in the acquisition of sets of signs, taken from a small sign language that serves as an analogue of plosive inventories in spoken language. In a learning experiment, participants acquired data sets that varied in feature economy and logical complexity. The results from this study suggest that ease of learning is best predicted by logical complexity, and that a considerable number of learners unintentionally reduce the complexity of their input.

1 Introduction

The contributions in this volume present various perspectives on the notion of complexity, illustrating the wide array of applications this term has in linguistics. Perhaps the most widely known example from phonology is found in the description of syllable structure, where onsets and codas are called complex if they contain more than one segment. However, many other interpretations are possible: Maddieson (2009), for instance, argues that a phonological alternation is more complex when it is less predictable. The present article focuses on two specific quantifications of complexity, namely feature economy and logical complexity (or incompressibility), and compares them as predictors for ease of learning in a phonological acquisition task.

The structure of this article is as follows: section 2 discusses the possible role of complexity in phonological acquisition, and introduces the measures of feature economy and logical complexity; section 3 describes the experimental stimuli and procedure; section 4 presents the results; the conclusion and discussion form section 5.

[1] I'm greatly indebted to Paul Boersma and Silke Hamann for invaluable advice and fruitful discussion, to Dirk Jan Vet for technical assistance, to an anonymous reviewer for their excellent suggestions, and to all participants for their time and effort.
[2] Correspondence: Spuistraat 134, 1012 VB Amsterdam; seinhorst@uva.nl

2 Complexity in the acquisition of feature combinations

Phonological segments are often regarded as bundles of features. For instance, the combination of [–continuant], [+bilabial] and [–voiced] describes the segment /p/, and only /p/. Such features are not merely useful descriptive tools, but they have psychological reality in the speaker-listener (a.o. Chládková 2013). Features are commonly used to analyse the internal structure of phoneme inventories, both in spoken language and sign language. Pressures of articulatory/gestural ease and perceptual distinctiveness play a major role in the typology of such inventories (for spoken language: Passy 1890; Martinet 1955, 1968; Boersma 1998; Boersma and Hamann 2008; for sign language: Crasborn 2001; Mathur and Rathmann 2001; Ann 2008; Ormel, Crasborn and Van der Kooij 2013), but cognitive constraints operate on the typology of phoneme inventories as well. For instance, it has often been noted that sound systems disprefer gaps; Martinet (1968) ascribes the sparsity of such systems to cognitive factors. In terms of feature economy – a principle stating that languages tend to maximally combine their phonological features (De Groot 1931; Martinet 1955; Clements 2003, 2005) – Martinet would predict that more economical inventories are easier to learn; and what is easier to learn, is more likely to be cross-linguistically frequent (Kirby and Hurford 2002; Christiansen and Chater 2008; Chater and Christiansen 2010).

2.1 Learning of category structures: non-linguistic stimuli

In experimental psychology, the learning of classes of feature combinations has been investigated since at least the early 1960s (a.o. Shepard, Hovland and Jenkins 1961; Nosofsky et al. 1994; Feldman 2000). These experiments use a set of 8 stimuli whose properties are described in terms of three binary features. This data set can be divided into two mutually exclusive classes of 4 stimuli in different ways, all of which may be – through rotation and/or mirroring – reduced to one of six possible so-called category structures.[3] These are Types I–VI in Figure 1. The stimuli that are included in a class are drawn as black circles, those that are not included as white circles.

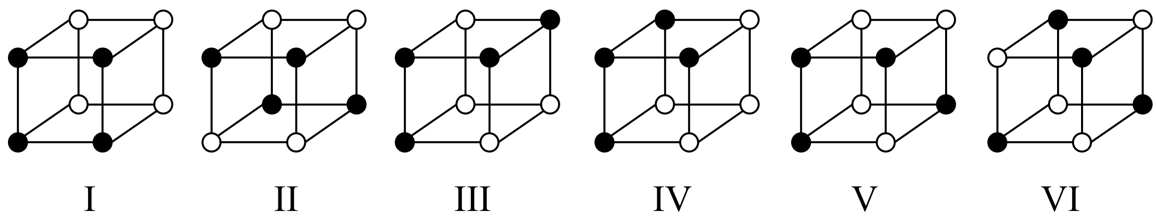


Figure 1. *The six category structures from Shepard et al. (1961).*

[3] I will speak of 'classes' rather than 'categories', because the latter term will later be used to refer to phonological categories.

The three features are represented in the three dimensions: they are binary because they can only take on two values (i.e. in the figure: front vs. back, left vs. right, top vs. bottom). Suppose that the features are shape (square vs. triangle), size (small vs. large) and colour (black vs. white). If the stimuli are divided according to Type I, this division could look like Figure 2a.; divisions of Types II and VI could look like Figures 2b. and c. respectively.

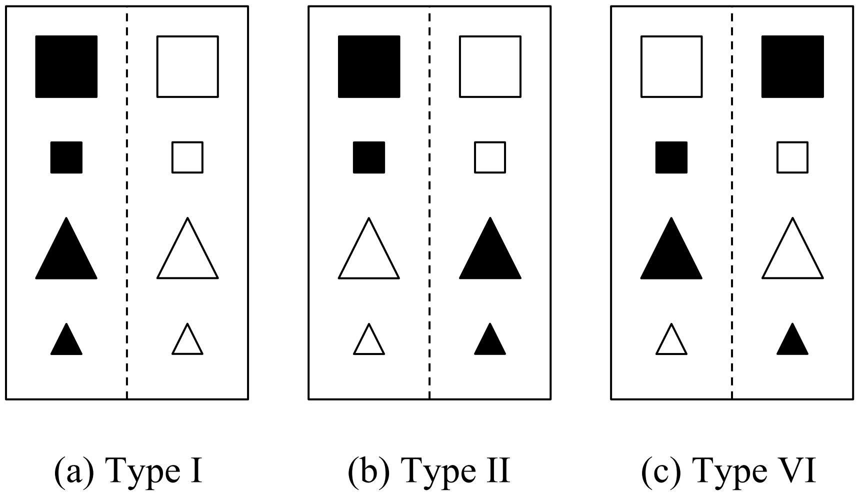


Figure 2. *Examples of stimulus divisions from different Shepard types.*

Shepard et al. (1961: 3) presume that higher Type numbers are more difficult to learn and remember. In order to classify stimuli from a Type I division as belonging to either the left class or the right one, two features can be disregarded: in the example shown in Fig. 2(a), shape and size are irrelevant. For the Type II division from Fig. 2(b), only size is irrelevant, and for Type VI divisions, none of the features can be ignored.

Shepard et al. carried out two experiments. In the first experiment, learners were shown the individual stimuli and replied to them with one of two response categories, after which they received feedback on their response. The experiment was completed when participants had given 32 consecutive correct responses. In the second experiment, subjects were asked to formulate the rules they thought underlay the division, and two weeks later were instructed to recreate the division from memory. The results of both experiments reflect the increasing difficulty of the six types: learners perform best on Type I category structures, worse on divisions of Type II, even more poorly on Types III–V, and worst on Type VI. For instance, many participants indicated that they had learned Type VI divisions by rote. In an experiment using the same six types, Griffiths, Christian & Kalish (2008) presented learners with three out of four stimuli from a class, then asked them to complete the set. In an iterated learning paradigm, they found that Type I became increasingly frequent over generations.

There are (at least) two ways of quantifying complexity in the Shepard types: we can compute their feature economy indices, and their logical complexities. Table 1 lists all feature economy indices E, using a computation similar to Hall (2007: 176)'s "Exploitation" measure: feature economy is computed by dividing the number of categories in a class (always 4) by the product of the number of shape, size and colour distinctions within a class. This product yields the total number of possible categories given the distinctions made; E thus expresses to what extent a type makes use of its full "potential". The columns in Table 1 show the number of categories in each class, the number of shape distinctions, the number of size distinctions, the number of colour distinctions, and the feature economy index. Note that the measure of economy does not necessarily correlate with the number of features that are relevant to learners in the tasks.

Table 1. *Feature economy indices* E *of Shepard, Hovland and Jenkins' six types.*

Type	*categories*	*shapes*	*sizes*	*colours*	*E*
I	4	2	2	1	1.0
II	4	2	2	2	0.5
III	4	2	2	2	0.5
IV	4	2	2	2	0.5
V	4	2	2	2	0.5
VI	4	2	2	2	0.5

If feature economy is related to ease of learning, higher values of E should correspond to better performance in a learning task. On the basis of this measure, however, we would not expect Shepard, Hovland and Jenkins' results: it erroneously predicts similar scores for Types II–VI.

Feldman (2000) calls upon a different measure of complexity: he suggests that participants' scores can be predicted by logical (Boolean) complexity or incompressibility, and that logically simple data sets are easier to learn. Table 1 from Feldman (2000: 631) is presented here as Table 2. For each type, this table lists the disjunctive normal form (a summation of the members of the class), the minimal formula (the shortest possible description of the set) and the logical complexity lc of a class. This measure of complexity has been quantified as the number of literals in the minimal formula. More compressible sets can be represented with a shorter minimal formula, and are hence less complex. a, b and c are the three binary dimensions or Boolean variables that can either have value 0 (e.g. $\neg a$, written here as a') or 1 (e.g. a). ab means $a \wedge b$.

Table 2. *Logical complexities* lc *of Shepard, Hovland and Jenkins' six category structures. The letters a, b and c indicate the relevant three features.*

Type	disjunctive normal form	minimal formula	*lc*
I	$a'b'c'+a'b'c+a'bc'+a'bc$	a'	1
II	$a'b'c'+a'b'c+abc'+abc$	$ab+a'b'$	4
III	$a'b'c'+a'b'c+a'bc'+ab'c$	$a'(bc)'+ab'c$	6
IV	$a'b'c'+a'b'c+a'bc'+ab'c'$	$a'(bc)'+ab'c'$	6
V	$a'b'c'+a'b'c+a'bc'+abc$	$a'(bc)'+abc$	6
VI	$a'b'c'+a'bc+ab'c+abc'$	$a(b'c+bc')+a'(b'c'+bc)$	10

Because all divisions are symmetric, both classes in a type (i.e. both halves in Figs. 2(a)-(c)) have the same complexity index.

Table 3 allows for easy comparison between the two complexity indices for each type:

Table 3. *All complexity indices of Shepard, Hovland and Jenkins' six category structures.*

Type	*E*	*lc*
I	1.0	1
II	0.5	4
III	0.5	6
IV	0.5	6
V	0.5	6
VI	0.5	10

Note that logical complexity correctly predicts the hierarchy found by Shepard, Hovland and Jenkins (1961); it also predicts the hierarchy found in Feldman's (2000) experiments with smaller subsets from Shepard types.

2.2 Learning of category structures: linguistic stimuli

The learning experiments discussed above suggest that ease of learning in a non-linguistic acquisition task is better predicted by complexity than by feature economy, and Kirby et al. (2015) postulate that pressures of increasing compressibility have played a crucial role in the evolution of language as well. So far, fairly little experimental data on the role of complexity in phonological acquisition is available. Saffran and Thiessen (2003) suggest that infants have less difficulty acquiring sound patterns over which a phonological generalisation can be drawn than they have with non-generalisable stimuli; Pater and Staubs (2013) provide a computational model of the learning of plosive inventories across generations with a grammar based on feature economy, and show that inventories with a high feature economy emerge as a result of the iterated learning process; Moreton, Pater and Pertsova (2015) use the Shepard types to investigate the learning of phonological alternations, but do not exactly replicate the order found in non-linguistic experiments.

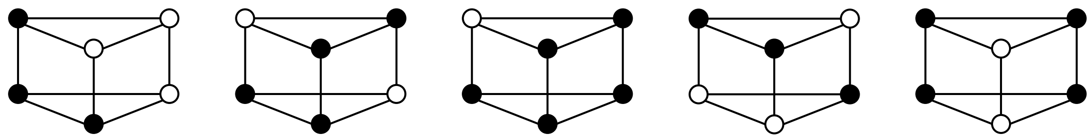
Any experiment that investigates language acquisition must differ fundamentally from experiments as described in §2.1, because of the nature of the language learning process: language acquisition does not involve a dichotomy between categories that are present in the language vs. those that are absent. The language learner is only confronted with positive evidence about the sounds in his language, and does not have any knowledge of feature combinations that the language lacks. That means that he does not hear non-native phonemes (which would be similar to Shepard et al.'s design), nor is he exposed to only part of a sound system (which would be similar to Griffiths et al.'s design). Remember, however, that the complexity indices given in Tables 1-3 are identical for both classes in a Shepard division: if a class of stimuli is absent, that does not change the complexity of the remaining class.

This article reports on an exploratory study that compares the measures of feature economy and logical complexity in a phonological acquisition task. Like the experiments with Shepard types, it makes use of category structures, but in this experiment they mirror the structure of cross-linguistically frequent plosive inventories. Plosive inventories were chosen because all spoken languages seem to make use of this type of speech sound (Maddieson 1984; Mielke 2008). Many languages have at least a three-way place of articulation contrast, and most implement an additional voicing contrast. The combinations of the cross-linguistically preferred features are listed in Table 4:

Table 4. *Cross-linguistically frequent feature combinations in plosive inventories.*

	[bilabial]	[alveolar]	[velar]
[–voiced]	/p/	/t/	/k/
[+voiced]	/b/	/d/	/g/

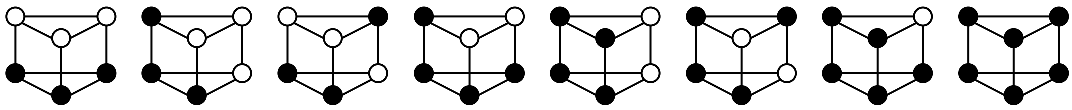
The translation of these features into a Shepard type-like representation yields an equilateral triangular prism: in one feature plane (voicing), there are two values, in the other (place) there are three, equidistant from each other. Any inventory can be represented within this prism, cf. Figure 3. Suppose that the bottom plane contains all voiceless segments, the top plane all voiced ones; of the three vertical edges, the left one connects the labials, the central one connects the alveolars, and the right one connects the velars. The top left vertex thus indicates [b], the bottom right vertex corresponds to [k]. Figures 3(a)-(e) show how a number of randomly chosen systems can be represented within the prism: black circles indicate categories that the system does contain, white circles correspond to categories that the system does not contain.



(a) “b p t” (b) “p t d g” (c) “p t d k g” (d) “b d k” (e) “b p g k”

Figure 3. *Feature representations of phoneme inventories*.

Languages generally employ between three and six of these segments, which means that a total of $\binom{6}{3}+\binom{6}{4}+\binom{6}{5}+\binom{6}{6} = 20 + 15 + 6 + 1 = 42$ different plosive inventories can be drawn from this set. Through rotation and mirroring, each and any of these 42 inventories can be reduced to one of the eight category structures depicted in Figure 4.



I II III IV V VI VII VIII

Figure 4. *The eight possible types of plosive category structures*.

Any inventory with three categories must belong to one of the Types I–III; any inventory with four categories is of one of the Types IV–VI; Type VII describes all five-category systems, and Type VIII comprises the full set. For instance, the inventory in Fig. 3(a) is of Type II; that in 3(b) of Type VI; that in 3(c) of Type VII; that in 3(d) of Type III; and that in 3(e) of Type V.

2.3 Feature economy, logical complexity, and gaps

For each of the six types in Fig. 4, Table 5 lists the number of categories *C*, the number of voicing feature values *V*, the number of place feature values *P*, and an index of feature economy E ($E = \frac{C}{VP}$, as above).

Table 5. *Feature economy indices* E *for all eight category structures (*C *= number of categories,* V *= number of voicing feature values,* P *= number of place feature values)*.

Type	*C*	*V*	*P*	*E*
I	3	1	3	1.0
II	3	2	2	0.75
III	3	2	3	0.5
IV	4	2	3	0.67
V	4	2	2	1.0
VI	4	2	3	0.67
VII	5	2	3	0.83
VIII	6	2	3	1.0

Table 6 presents the disjunctive normal forms, minimal formulae and logical complexities of the eight types. The voicing feature is represented as *a*, and can have one of two values *a* and *a´*; the place feature is represented as *b*, and can have one of three values *b*, *b´* and *b″*.

Table 6. *Logical complexities* lc *for all eight category structures*. *The letters* a *and* b *indicate the relevant two features.*

Type	disjunctive normal form	minimal formula	*lc*
I	*ab+ab´+ab″*	*a*	1
II	*ab+ab´+a´b*	*ab´+b*	3
III	*ab+ab´+a´b″*	*a(b+b´)+a´b″*	5
IV	*ab+ab´+ab″+a´b´*	*a+a´b´*	3
V	*ab+ab´+a´b+a´b´*	*b+b´*	2
VI	*ab+ab´+a´b+a´b″*	*ab´+a´b″+b*	5
VII	*ab+ab´+ab″+a´b+a´b´*	*a+a´(b+b´)*	4
VIII	*ab+ab´+ab″+a´b+a´b´+a´b″*	*A [all]*	1

Table 7 lists the values of both complexity measures for all eight types.

Table 7. *Feature economy indices* E *and logical complexities* lc *for all eight category structures.*

Type	*E*	*lc*
I	1.0	1
II	0.75	3
III	0.5	5
IV	0.67	3
V	1.0	2
VI	0.67	5
VII	0.83	4
VIII	1.0	1

In general, feature economy seems to increase as logical complexity decreases, but the correlation between them is not perfect ($\rho = -.866$). Compare, for instance, Types IV and VI, which have the same *E* but different *lc* values, or Types III and VI, which have the same *lc* but different *E* values. If learnability is positively correlated with feature economy, hierarchy (i) is to be expected, in ascending order of difficulty ('$x < y$' means '*x* is easier to learn than *y*'; if logical complexity predicts learnability, we expect hierarchy (ii).

(i) I = V = VIII < VII < II < IV = VI < III

(ii) I = VIII < V < II = IV < VII < III = VI

Fig. 4 clearly shows that most types, i.e. all except for I, V and VIII, have a number of gaps in the system. These gaps correspond to categories that could have existed given the relevant features of the inventory, but that are in fact absent. Both experimental data and observations about natural language acquisition suggest that learners' errors tend to favour regular systems, i.e. systems without gaps (a.o. Singleton and Newport 2004; Hudson Kam and Newport 2005; Reali and Griffiths 2009; Ferdinand et al. 2013). This tendency may be attributed to inductive biases: in the learning process, hypotheses favouring regular systems may have higher a priori probabilities and are thus more likely to be selected in acquisition. Such errors are readily explained in terms of complexity: filling a gap always increases the feature economy of a system, and in most cases increases its compressibility (the only exceptions would be a Type III system changing into a Type VI system, keeping its logical complexity at 5, and a Type IV system changing into a Type VII system, raising its logical complexity from 3 to 4).

3 Stimuli and method

This section describes an experimental paradigm that aims to assess the learnability of the types from Fig. 4. This paradigm should shed light on two questions: firstly, whether learnability differs between types, and secondly, whether learnability is better predicted by feature economy or by logical complexity.

3.1 Stimuli

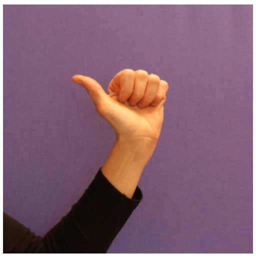
In experimental investigation of phonological acquisition, one runs the risk of interference from participants' language background with the learning task. For instance, if one of the categories to be learnt is absent from the participant's language, he is likely to map tokens from that (foreign) category to a different, native category (a.o. Lisker 2001 on the perception of Polish sibilants by English listeners; for Dutch listeners' perception of [g], see Schuttenhelm 2013). Also, if the segments in the stimulus set are a subset of the participant's segment inventory, he could simply draw upon part of his knowledge: any gaps in the subset do not correspond to gaps in the learner's inventory, which makes it difficult to probe learnability issues in the emergence of a new feature system. One solution to this problem is the use of linguistic stimuli in a different modality, i.e. signs.[4] This strategy has been employed more often, exactly in order to avoid influences from the extant language system: Smith, Abramova and Kirby (2012), for instance, use sign language to investigate how the encoding of semantic features emerges in a preset meaning space (viz. manner and path in descriptions of movement). In this experiment, however, an influence of the participant's language background still seems likely, as a result of the semantic component of the task; for the investigation of phoneme inventories no transfer is expected, because knowledge of features in sign language is independent from knowledge of features in spoken language.

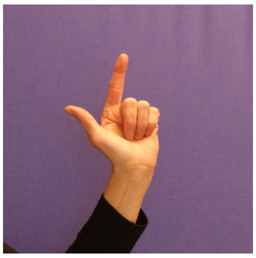
For the present learning experiment, an artificial sign language was composed that consists of six signs. In analogy with the 'basic' plosive set from Table 3, these signs can be described as a combination of two features: a binary thumb opposition feature, and a ternary handshape feature. The thumb could be either opposed or unopposed; the three handshapes are (1) zero fingers pointing up (a clenched fist); (2) the index finger pointing up; (3) four fingers, i.e. all except the thumb, pointing up. These features and their values were chosen because they seem perceptually sufficiently distinct. The handshape feature is very commonly used in sign languages, with many more feature values than the three presented here. Thumb opposition does not seem to be distinctive in sign languages (they rather use e.g. movement), but in terms of implementing and running the experiment, a static feature would be easier to use than a dynamic one.

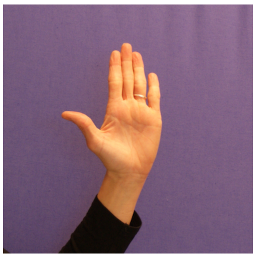
---

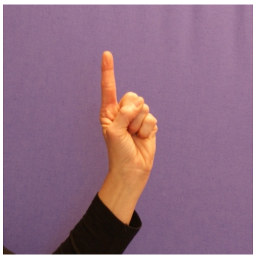
[4] Thanks to Anne Baker and Roland Pfau for this suggestion.

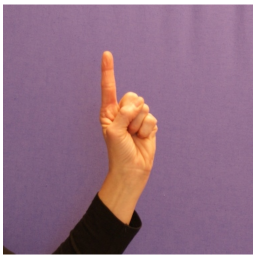
Examples of the six signs are shown in Figure 5. Hereafter I will notate them as a feature combination composed of two numbers: firstly the handshape feature, given as the number of fingers pointing up (0, 1 or 4), and secondly the thumb opposition feature, expressed as a Boolean variable (0 = thumb unopposed, 1 = thumb opposed).


/0 0/


/1 0/


/4 0/


/0 1/


/1 1/

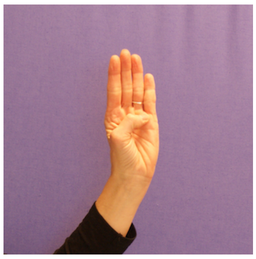

/4 1/

Figure 5. *The six phonemes of the artificial sign language.*

A female signer was photographed producing each sign ten times, so that the data set contained variability. This was done to replicate the lack of invariance with which every learner of a new sound system is faced, and from which he has to induce the intended discrete categories that are relevant in his language.

3.2 Method

48 adults participated in the experiment (38 female, 10 male), none of whom had any prior knowledge about any sign language. Each participant was trained on one of the category structures from Fig. 4, so there were 6 learners per type. The six signs from Fig. 5 were distributed among the eight types as shown in Table 8 ('+' means "present in the input", '–' means "absent from the input"). In this distribution, the categories were scattered across the types fairly evenly, to reduce any influence of salience differences between the signs.

Table 8. *The distribution of signs across types, as used in the experiment.*

	categories					
Type	/0 0/	/0 1/	/1 0/	/1 1/	/4 0/	/4 1/
I	–	+	–	+	–	+
II	–	–	+	+	+	–
III	–	+	–	+	+	–
IV	+	+	+	–	+	–
V	–	–	+	+	+	+
VI	+	+	+	–	–	+
VII	+	+	+	+	–	+
VIII	+	+	+	+	+	+

The experiment was run in ED, a freeware application similar to E-Prime (Vet 2013). Each experiment started with a training phase, using pictures of geometric shapes as stimuli, in which the participant was familiarised with the task. After this phase, the acquisition of the sign language began. The participant was shown a photo of a token for 2000 ms; after exposure, a 'Next' button appeared under the photo, which the participant had to click to proceed to the next stimulus. The stimuli were presented in random order. The categories that constitute a type appeared in the input 30 times, so every photo was shown three times. A Type V learner thus saw a total of 120 photos. The number of photos shown per category was kept constant between types, rather than the total number of photos. Although this does not necessarily reflect natural language acquisition, it avoided the possibility that participants with larger types would be able to count the number of photos, as well as the risk that experiments with smaller types would be monotonous while learners of larger types would receive too little input to perform the task accurately. An average experiment, including the training phase and debriefing, lasted between 15 and 30 minutes (depending on type size).

Because of the linguistic nature of the experiment (cf. §2.2), the methodology had to differ fundamentally from similar experiments with non-linguistic stimuli, such as Shepard et al. (1961), with two mutually exclusive classes, or Griffiths et al. (2008), in which participants were only exposed to part of the division. Instead, participants were asked to carry out a frequency estimation task. After the learning phase, a screen appeared showing pictures of eight signs: the six signs from the language plus two control signs. The control signs were not possible signs of the sign language: in one, the thumb and index fingers created a circular shape while the remaining fingers pointed up (used to convey the meaning "fine"), in the other the little finger and thumb pointed up, the remaining fingers down. Juxtaposed to each of the signs were sliders, whose leftmost position was labelled "not at all", and whose rightmost position was labelled "very often". There were no ticks along the slider, in order to avoid a preference for the values associated with these ticks. The initial position of the slider was random and has not been recorded. Participants adjusted the positions of the eight sliders to indicate the relative frequency with which each of the signs had appeared in the input. The test subjects did not have to produce the signs themselves, because the experiment intends to investigate the learning of systems with different internal structures, without focusing on the roles of perceptual distinctivity and gestural ease. The chosen task was expected to not only provide insight into the relative learnability of the data set, but also to reveal the nature of any learnability issues, i.e. any systematicity with which participants might disregard and/or add categories.

4 Results

The learnability of a type is considered to be indicated by its average error score: the more errors participants make on a type, the lower its learnability. The error score has been operationalised as follows. The scale on which participants indicated estimated frequency was discretised in 100 steps of equal size. The left-hand end of the slider was assigned the value 0, the right-hand end was assigned the value 100. The discretisation in 100 steps was done in order to make the slider move smoothly on the participant's computer screen, but since the sliders did not appear on the screen big enough for them to actually produce the smallest possible difference (i.e. 1 on the scale), their indicated values were rounded off to the nearest multiple of 5.

Some participants did not use the full range of the scale: in these cases, the highest indicated frequency of any category was scaled up to 100, and the other frequencies were adjusted accordingly. The lowest indicated frequency, if not zero, was not set to be zero, as this value has an absolute interpretation: as opposed to any non-zero, zero means that the participant has not seen a sign at all. For all six signs, the difference between rounded estimated frequency and input frequency was computed (the latter being either 0 or 100, for categories that are absent or present in the input, respectively). The average of these six differences is the error score.

As an example of frequency scaling and the computation of the error scores, consider Table 9, containing possible responses of a fictional Type VII learner. The highest indicated frequency is 94, which is rounded off to 95. This latter value is scaled up to 100, and the other indicated frequencies are multiplied by $\frac{100}{95}$ as well.

Table 9. *An example of frequency scaling and error score computation*.

sign	indicated frequency (raw)	indicated frequency (rounded)	indicated frequency (scaled)	input frequency	misestimation
/0 0/	94	95	100	100	0
/0 1/	80	80	84	100	16
/1 0/	83	85	89	100	11
/1 1/	91	90	95	100	5
/4 0/	8	10	11	0	11
/4 1/	90	90	95	100	5
	average misestimation = error score:				8

None of the participants indicated having seen the control signs. The average error scores per category are reported in Table 10 and presented visually in Figure 6.

Table 10. *Error scores per type: averages and standard deviations*.

Type	feature economy	logical complexity	average error score	st. dev. of error score
I	1.0	1	4	3
II	0.75	3	2	4
III	0.5	5	12	12
IV	0.67	3	6	5
V	1.0	2	9	8
VI	0.67	5	40	12
VII	0.83	4	15	11
VIII	1.0	1	8	6

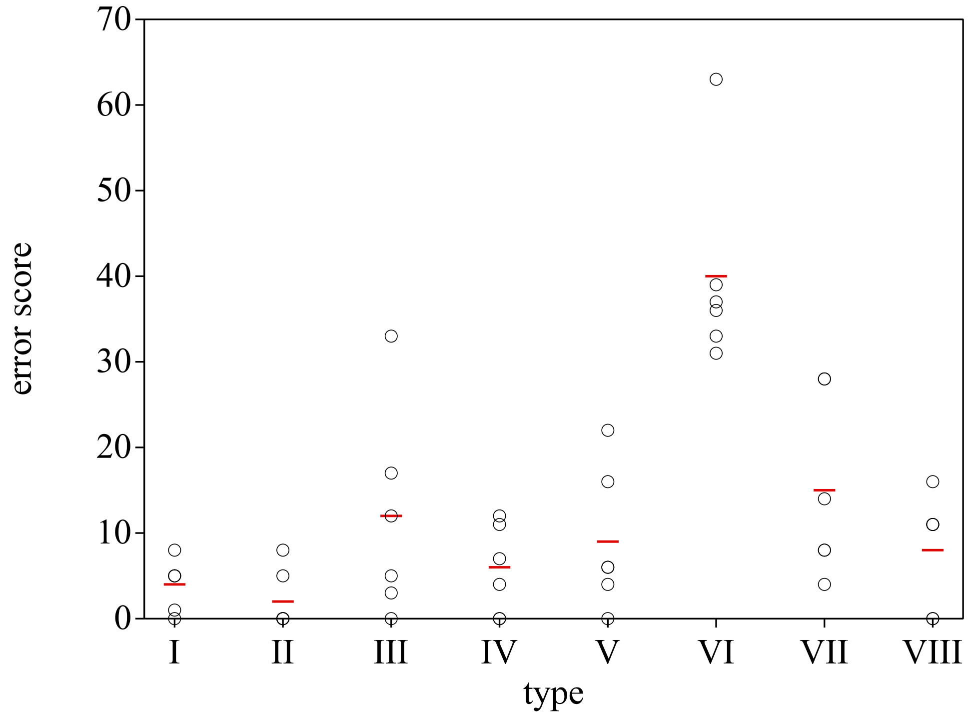


Figure 6. *Error scores per type. Circles indicate individual scores: red bars indicate average error scores.*

There is an effect of type on error score [$F(7, 40) = 12.412$, $p < .001$], so there are statistically significant differences in learnability between types. Because of the unequal variances and fairly small sample sizes for each type, Dunnett's T3 test was chosen for post-hoc tests, which reveal that the main effect is due to Type VI: participants perform significantly poorer on this type than on Types I–V and VIII. The differences between all other types are not statistically significant.[5]

4.1 Feature economy versus logical complexity

Figures 7 and 8 plot the error scores as a function of feature economy and logical complexity, respectively. An inverse relation is expected between feature economy and error score, whereas the error score is presumed to increase with logical complexity.

[5] The p values from the pairwise comparisons with Type VI: $p_{\text{I-VI}} = .007$; $p_{\text{II-VI}} = .005$; $p_{\text{III-VI}} = .043$; $p_{\text{IV-VI}} = .006$; $p_{\text{V-VI}} = .011$; $p_{\text{VI-VII}} = .062$; $p_{\text{VI-VIII}} = .008$.

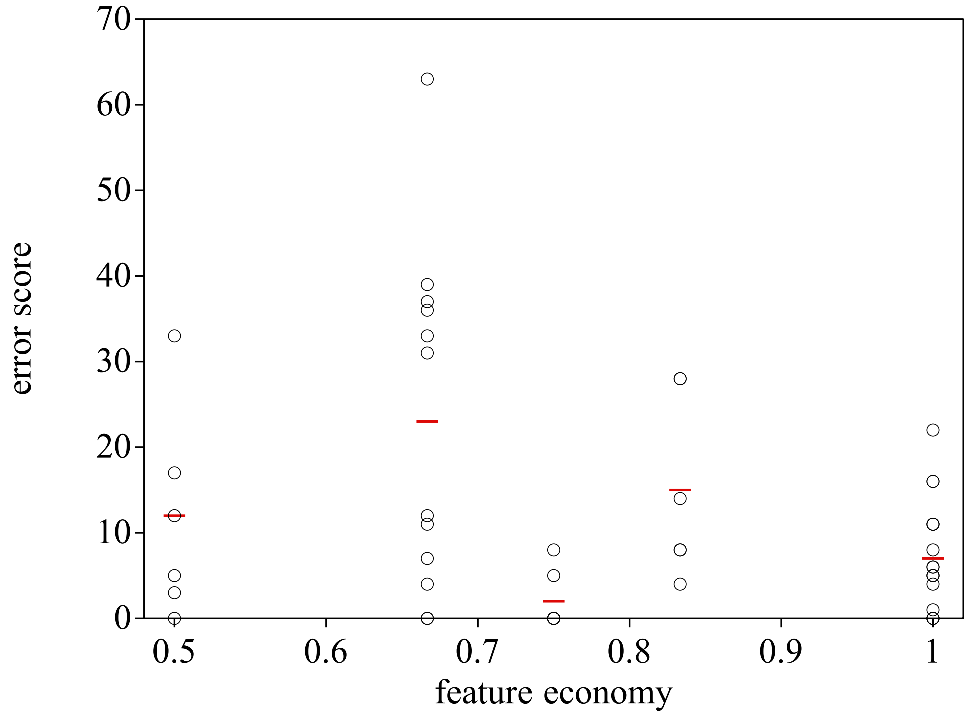


Figure 7. *Error scores as a function of feature economy. Circles indicate individual scores: red bars indicate average error scores.*

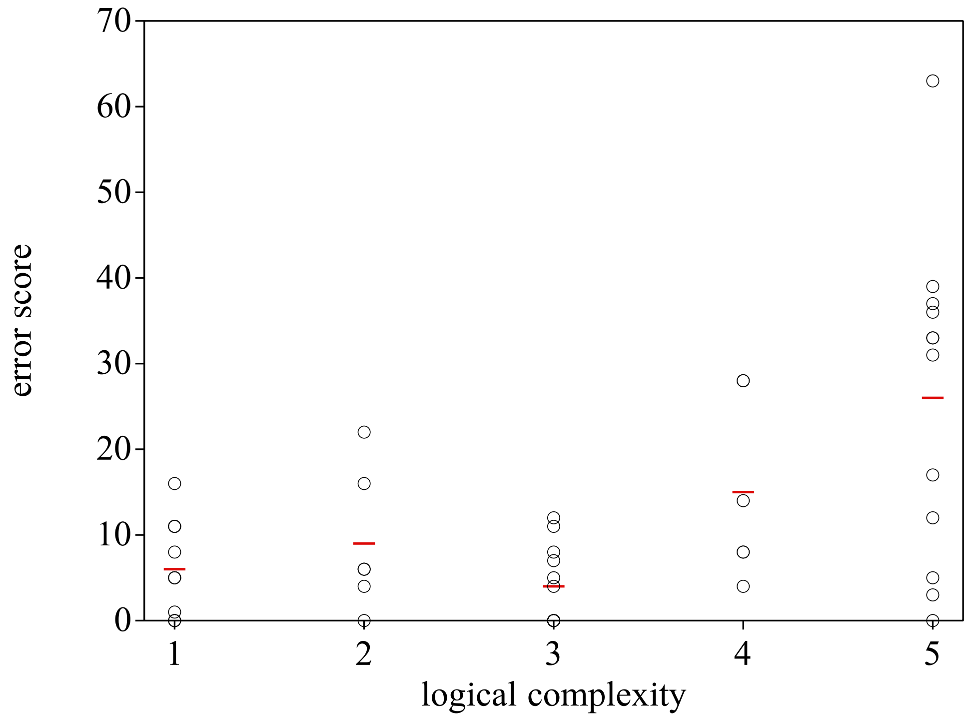


Figure 8. *Error scores as a function of logical complexity. Circles indicate individual scores: red bars indicate average error scores.*

A linear regression with feature economy as the sole predictor of error score does not yield statistically significant results ($\beta_E = -0.268$, $t = -1.888$, $p = .065$); a linear regression with logical complexity as the sole predictor of error score does ($\beta_{lc} = 0.520$, $t = 4.130$, $p < .001$). Steiger's Z (Steiger 1980) reveals that the difference between both coefficients is statistically significant ($\rho_{E,lc} = .866$, $n = 48$, $Z = 3.7$, $p < .001$), meaning that logical complexity predicts error scores significantly better than feature economy.

The fit of the linear models may be improved by including the size of the learnt type as a factor. An influence of number of categories on error score is feasible, as the experiments took longer for larger types, and thus posed larger demands on participants' memory and attention spans. Indeed, a linear regression in which both feature economy and number of categories predict error score yields statistically significant results [$F(2, 45) = 3.626$, $p = .035$], but again, logical complexity is the statistically significantly better predictor of the two ($\rho_{E,nc} = .372$, $\rho_{lc,nc} = .583$, $\rho_{E,lc} = .868$, $n = 48$, $Z = 3.17$, $p = .001$).

Although these results cannot be compared directly to findings from experiments with Shepard types, they do indicate a crucial role of logical complexity in phonological learning. Contrary to Martinet's (1968) and Pater and Staubs' (2013) assumption, and in line with Feldman (2000) and Griffiths et al. (2008), logical complexity seems to be a better predictor of ease of learning than feature economy; apparently logical complexity impedes acquisition success.

4.2 Error patterns: reduction of complexity

This subsection zooms in on participants' response patterns. Table 11 reveals how many subjects indicated having seen any of the eight types, and the proportion of correct classifications. Note that this summary of subjects' responses abstracts away from the estimated proportions. By means of illustration, consider the fictional Type VII learner from Table 9: although they only indicated a very low frequency (raw score 8) for the sign that was absent from their input, the fact that they indicated having seen it at all would assign their response to Type VIII. There is still a large difference between this system and the input to Type VIII learners, in which all six signs occur equally often.

Table 11. *Subjects' responses per type. Type changes that added one or more categories are given in bold, type changes that removed one or more categories are given in italic.*

		response									
		I	II	III	IV	V	VI	VII	VIII	oth.	p_{corr}
input	I	6									1
	II		6								1
	III	*1*		4						*1*	.67
	IV				6						1
	V					5			**1**		.83
	VI						0		**6**		0
	VII							5	**1**		.83
	VIII								6		1

The diagonal that runs from the top left to bottom right contains all correct classifications. Any number outside this diagonal is an error. In total, the learners have made ten errors: one Type III learner reported having seen only two signs, i.e. this participant's response does not correspond to one of the eight types and has been put in the column 'other'; another Type III learner responded with Type I; all six Type VI learners responded with Type VIII; one Type VII learner also indicated having seen a Type VIII system. In the second Type III case, this means that the learner exchanged one category they had seen for another they had not seen; for Types VI and VII, it means that learners introduced only unseen categories in their responses.

We see that the errors, except for one, favour only Types I and VIII, i.e. the types without gaps: the participants show strongly regularising behaviour. This also entails a stark decrease of the cumulative logical complexity in the entire data set. The five columns of Table 12 list the complexity indices of the eight types; the number of participants who learnt them (n_{before}; the participant from the 'other' column in Table 9 was not taken into consideration, so Type III has only five learners); the number of participants who selected them after learning (n_{after}); and the contribution per type to the cumulative complexity before and after learning (lc_{before} and lc_{after}, respectively). For each type, $lc_{before} = lc \cdot n_{before}$ and $lc_{after} = lc \cdot n_{after}$. These measures reveal how much every type adds to the cumulative complexity. We see that the demise of Type VI reduces the complexity by 30, whereas the growth of Type VIII adds 8.

Table 12. *Cumulative logical complexity in the in- and output* (lc = *logical complexity,* n_{before} = *number of participants who received this type as input,* n_{after} = *number of participants who chose this type as output,* $lc_{before} = lc \cdot n_{before}$, $lc_{after} = lc \cdot n_{after}$).

Type	*lc*	n_{before}	n_{after}	lc_{before}	lc_{after}
I	1	6	7	6	7
II	3	6	6	18	18
III	5	5	4	25	20
IV	3	6	6	18	18
V	2	6	5	12	10
VI	5	6	0	30	0
VII	4	6	5	24	20
VIII	1	6	14	6	14
total				139	107

The cumulative complexity decreases from 139 to 107, i.e. by 23.0%. This decrease is most likely unintentional, as there was no reason for participants to actively implement it; also, some participants remarked that the task was easy, while they had in fact selected a different type than their input.

5 Conclusion and discussion

In a learning experiment, the roles of two measures of complexity in language acquisition were probed: feature economy and logical complexity. The stimuli were combinations of phonological features, like phonemes in spoken language; however, because we were interested in the acquisition of a new feature system, we used a small language of six signs as the stimulus set. These signs could be described in terms of a binary thumb opposition feature and a ternary handshape feature, as analogues of the voicing feature and place of articulation feature in spoken language, respectively. Participants learnt one of eight data sets (called "types"), consisting of three, four, five or six signs, and subsequently carried out a proportion estimation task. The types differed in their feature economies and logical complexities.

The error scores on this task are better predicted by the logical complexity of the set than by its feature economy. The majority of errors favoured regular systems, i.e. systems without gaps; because the complexity of such systems is low, the overall complexity in the data set as a whole was considerably reduced.

In future research, a larger-scale version of this study will be done with a larger participant sample. Additionally, in the future experiment the features will have been distributed among learners perfectly evenly (the distribution in the current study left some room for improvement; cf. Table 8), to minimise the influence of differences in perceptual salience between signs. These experiments will investigate the learning of both signed and spoken stimuli, using the same eight types from the current study. This way we can establish the robustness of the learnability effect attested here, and assess the role of the added task of feature induction in the sign language acquisition task. In addition, a comparison with typological data will be drawn, to see whether the experimental results comply with spoken-language data; if so, this would provide empirical evidence for the hypothesis that the typology of sound systems is constrained by considerations of complexity. In a similar vein, it would be interesting to also compare the results with data from other types of inventories that can be described in terms of features, such as pronominal systems (with properties like number, person, gender, in-/exclusive, proximity etc.).

References:

Ann, Jean (2008). *Frequency of occurrence and ease of articulation of sign language handshapes: the Taiwanese example*. Washington, DC: Gallaudet University Press.

Boersma, Paul (1998). *Functional Phonology: formalizing the interaction between perceptual and articulatory drives*. PhD thesis, University of Amsterdam. The Hague: Holland Academic Graphics.

Boersma, Paul and Silke Hamann (2008). The evolution of auditory dispersion in bidirectional constraint grammars. *Phonology* 25: 217–270.

Chater, Nick and Morten Christiansen (2010). Language acquisition meets language evolution. *Cognitive Science* 34: 1131–1157.

Chládková, Kateřina (2013). *Finding phonological features in perception*. PhD thesis, University of Amsterdam.

Christiansen, Morten and Nick Chater (2008). Language as shaped by the brain. *Behavioral and Brain Sciences* 31: 489–558.

Clements, Nick (2003). Feature economy in sound systems. *Phonology* 20: 287–333.

Clements, Nick (2005). The role of features in speech sound inventories. In: Raimy, Eric and Charles Cairns (eds.), *Contemporary views on architecture and representations in phonological theory*, 19–68. Cambridge, MA: MIT Press.

Crasborn, Onno (2001). *Phonetic implementation of phonological categories in Sign Language of the Netherlands*. PhD thesis, Leiden University. Utrecht: LOT.

de Groot, Willem (1931). Phonologie und Phonetik als Funktionswissenschaften. *Travaux du Cercle Linguistique de Prague* 4: 146–147.

Feldman, Jacob (2000). Minimization of Boolean complexity in human concept learning. *Nature* 407: 630–633.

Ferdinand, Vanessa, Bill Thompson, Kenny Smith and Simon Kirby (2013). Regularization behavior in a non-linguistic domain. In: Knauff, Markus, Michael Pauen, Natalie Sebanz and Ipke Wachsmuth (eds.), *Proceedings of the 35th Annual Cognitive Science Society*, 436–441.

Griffiths, Thomas, Brian Christian and Michael Kalish (2008). Using category structures to test iterated learning as a method for revealing inductive biases. *Cognitive Science* 32: 68–107.

Hall, Daniel Currie (2007). The role and representation of contrast in phonological theory. PhD thesis, University of Toronto.

Hudson Kam, Carla and Elissa Newport (2005). Regularizing unpredictable variation: the roles of adult and child learners in language formation and change. *Language Learning and Development* 1 (2): 151–195.

Kirby, Simon and James Hurford (2002). The emergence of linguistic structure: an overview of the iterated learning model. In: Cangelosi, Angelo and Domenico Parisi (eds.), *Simulating the evolution of language*, 121–148. London: Springer Verlag.

Kirby, Simon, Mónica Tamariz, Hannah Cornish & Kenny Smith (forthcoming). Compression and communication drive the evolution of language. *Cognition* 141: 87–102.

Lisker, Leigh (2001). Hearing the Polish sibilants [š ś s]: phonetic and auditory judgments. In: Grønnum, Nina & Jørgen Rischel (eds.), *To honour Eli Fischer-Jørgensen. Travaux du cercle linguistique de Copenhague XXXI*, 226-238. Copenhagen: Reitzel.

Maddieson, Ian (1984). *Patterns of sounds*. Cambridge: Cambridge University Press.

Maddieson, Ian (2009). Calculating phonological complexity. In: Pellegrino, François, Egidio Marsico, Ioana Chitoran and Christophe Coupé (eds.), *Approaches to phonological complexity*, 85–110. Berlin: Mouton de Gruyter.

Martinet, André (1955). *Économie des changements phonétiques: traité de phonologie diachronique*. Berne: Francke.

Martinet, André (1968). *La linguistique synchronique: études et recherches*. Paris: Presses Universitaires de France.

Mathur, Gaurav & Christian Rathmann (2001). Why not GIVE-US: an articulatory constraint in signed languages. In: Dively, Valerie, Melanie Metzger, Sarah Taub & Anne Marie Baer (eds.), *Signed languages. Discoveries from international research*, 1–25. Washington, DC: Gallaudet University Press.

Mielke, Jeff (2008). *The emergence of distinctive features*. Oxford: Oxford University Press.

Moreton, Elliot, Joe Pater and Katya Pertsova (2015). *Phonological concept learning*. Cognitive Science 1–66.

Nosofsky, Robert, Mark Gluck, Thomas Palmeri, Stephen McKinley and Paul Glauthier (1994). Comparing models of rule-based classification learning: a replication and extension of Shepard, Hovland and Jenkins (1961). *Memory & Cognition* 22: 352–369.

Ormel, Ellen, Onno Crasborn & Els van der Kooij (2013). Coarticulation of hand height in Sign Language of the Netherlands is affected by contact type. *Journal of Phonetics* 41: 156–171.

Passy, Paul (1890). *Étude sur les changements phonétiques et leurs caractères généraux*. Paris: Firmin-Didot.

Pater, Joe and Robert Staubs (2013). *Feature economy and iterated grammar learning*. Presentation, 21st Manchester Phonology Meeting, Manchester.

Reali, Florencia and Thomas Griffiths (2009). The evolution of frequency distributions: relating regularization to inductive biases through iterated learning. *Cognition* 111: 317–328.

Saffran, Jenny and Erik Thiessen (2003). Pattern induction by infant language learners. *Developmental Psychology* 39 (3): 484–494.

Schuttenhelm, Lisan (2013). *Perception of the non-native phone* [g] *in Dutch: where lies the VOT boundary?* MA thesis, University of Amsterdam.

Shepard, Roger, Carl Hovland and Herbert Jenkins (1961). Learning and memorization of classifications. *Psychological Monographs* 75 (13).

Singleton, Jenny and Elissa Newport (2004). When learners surpass their model: the acquisition of American Sign Language from impoverished input. *Cognitive Psychology* 49: 370–407.

Smith, Kenny, Katya Abramova and Simon Kirby (2012). *Gesture and the iterated learning paradigm*. Presentation, Amsterdam Gesture Center (Free University), Amsterdam.

Steiger, James (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin* 87: 245–251.

Vet, Dirk Jan (2013). *ED (Experiment Designer)*. www.fon.hum.uva.nl/dirk/ and http://145.18.230.100/dirk/edweb