# REGULARITY IN MAPPINGS BETWEEN SIGNALS AND MEANINGS

MÓNICA TAMARIZ & ANDREW D. M. SMITH
*Language Evolution and Computation Research Unit, Linguistics and English Language,*
*University of Edinburgh, 40 George Square, Edinburgh, EH8 9LL*
*monica@ling.ed.ac.uk / andrew@ling.ed.ac.uk*

We combine information theory and cross-situational learning to develop a novel metric for quantifying the degree of regularity in the mappings between signals and meanings that can be inferred from exposure to language in context. We illustrate this metric using the results of two artificial language learning experiments, which show that learners are sensitive, with a high level of individual variation, to systematic regularities in the input. Analysing language using this measure of regularity allows us to explore in detail how language learning and language use can both generate linguistic variation, leading to language change, and potentially complexify language structure, leading to qualitative language evolution.

## 1. Introduction

Croft (2000)'s evolutionary model of language change proposes that language is made up of multiple linguistic patterns, which can be differentially replicated across communities and over time, and thereby adapt to their environment. We investigate one potential functional source of such adaptation, namely the ease with which patterns of mapping between signals and meanings can be learnt.

Recent research focuses on the inherent stochasticity of language learning (Bod, Hay, & Jannedy, 2003); children make use of statistical regularities in their linguistic input to learn phonemic contrasts (Maye, Werker, & Gerken, 2002), word boundaries (Jusczyk, Goodman, & Baumann, 1999; Saffran, Newport, & Aslin, 1996) and basic syntactic dependencies (Gómez, 2002). Regularity helps us to learn the specific mappings between meanings and signals: sharing a linguistic label increases the degree to which meanings are perceived to be more similar (Sloutsky, Lo, & Fisher, 2001), and reliable co-occurrence with labels enhances the perceptual salience of features of referent meanings (Schyns & Murphy, 1994). Patterns of frequency of use also play a crucial role in the entrenchment of linguistic constructions and in the persistence of linguistic irregularity (Bybee & Hopper, 2001).

Few efforts, however, have been made to *quantify* the systematicity or regularity in linguistic knowledge. Our main aim in this paper is to propose such a measure, which can be used to examine how this regularity impacts on the learnability of languages and on their use. In Section 2, we present a novel measure of quantifying linguistic regularity, based on the confidence in the signal-meaning mappings that learners can derive from their experience over multiple episodes of language use. In Section 3, we use the measure in two artificial language learning experiments, and examine how learning is affected by regularities in the input. Finally, we briefly discuss the ramifications for language change and evolution.

## 2. Quantifying Linguistic Regularity

Researchers in evolutionary linguistics often make a distinction between compositional and holistic languages (Kirby, 2002; Brighton, 2002). In a compositional language, the meaning of a signal is a function of the meanings of elements of the signal and of the way those elements are arranged together. Symmetrically, the signal encoding a meaning is a function of the signals that encode elements of the meaning. In a holistic language, by contrast, there is no such relationship: the whole signal stands for the whole meaning. Human languages, however, are neither wholly compositional nor wholly holistic, but contain constructions of both types, and many with intermediate behaviour. Recent formulations of grammar (Langacker, 1987; Croft, 2001), indeed, use this insight to represent all linguistic knowledge in a large lexicon of constructions, or form-meaning pairings of varying levels of generality, ranging from very general compositional rules to idiosyncratic holistic idioms.

From an evolutionary point of view, it would be beneficial to compare languages in terms of their level of compositionality, to explore the conditions under which they become more systematic and can sustain complexity. Despite this, useful measures of systematicity are not available; among the very few attempts to measure language compositionality was K. Smith (2003), who used the correlation of similarity between signals with similarity between meanings, but only by considering signals and meanings holistically, and thus failing to isolate the effects of meaningful elements of signals and irreducible aspects of meanings. We aim here to fill this gap, by describing a gradient measure to quantify the regularity of mapping ($RegMap$) between signals and meanings. This measure is based on the cross-situational co-occurrence (Siskind, 1996; Smith, Smith, Blythe, & Vogt, 2006) of signal and meaning components in the language; it is bidirectional, and can thus be used to quantify both the regularity of the mapping from signals to meanings and vice versa; it can also be applied at many different levels of linguistic analysis, from measuring the regularity with which a particular morpheme encodes a component of meaning, to the overall regularity of the entire system.

We illustrate the method by exploring the regularities in the miniature artificial language shown in Table 1. In this language, meanings are represented in a three-dimensional meaning space $\{$COLOUR, SHAPE, INSET$\}$, with three different values on each dimension, giving the language 27 possible meanings in total. Each meaning is paired with a signal (shown in the cells of the table), which is also made up of three dimensions, or syllables $\{\sigma_1, \sigma_2, \sigma_3\}$. We can see that the signal

Table 1. A language with near-perfect compositionality. Values in syllables 1, 2 and 3 encode values on the meaning dimensions colour, shape and inset respectively, with the exception of the highlighted elements.

| | | | COLOUR | | | |
|---|---|---|---|---|---|---|
| | | *blue* | *red* | *yellow* | | |
| SHAPE | *square* | tulo**ga** | kilodi | pelodi | *cross* | INSET |
| | | tuloga | kiloga | peloga | *dot* | |
| | | tulobe | kilobe | pelobe | *star* | |
| | *hexagon* | tumudi | kimudi | **tu**mudi | *cross* | |
| | | tumuga | kimuga | pemuga | *dot* | |
| | | tumube | kimube | pemube | *star* | |
| | *oval* | tunadi | kinadi | penadi | *cross* | |
| | | **ki**naga | **pe**naga | **tu**naga | *dot* | |
| | | tunabe | kinabe | penabe | *star* | |

and meaning dimensions map onto each other almost perfectly; the first syllable encodes colour, the second, shape and the third, inset. Only a few elements (highlighted in the table), do not conform to this encoding, so these break the perfect compositionality of the language. On the other hand, the language is clearly far from holistic, as there remains a large degree of regularity in the signal-meaning mappings. How can we quantify this systematicity? We start by calculating how regularly a single signal dimension encodes a given meaning dimension, and then scale this up to measure $RegMap$ for the entire language.

### 2.1. *RegMap from a signal dimension to a meaning dimension*

In developing $RegMap$, we make use of humans' "cognitive preference for certainty and for robust, redundant descriptions" (Pierrehumbert, 2006, p.81), basing our metric on *redundancy*, namely the degree of predictability, order or certainty in a system. Redundancy is defined mathematically as the converse of entropy, as measured over a finite set of mutually independent variants (Shannon, 1948).

Consider, then, the different variants in the first syllable in the language shown in Table 1, namely {`ki, pe, tu`}, and how they co-occur with the variants {*red, blue, yellow*} of the meaning dimension COLOUR depicted in the columns of the table. For each signal variant $s$ in dimension $\sigma_1$, we can calculate the relative entropy and thence its redundancy $R_s$ across meaning variants:

$$R_s = 1 - \frac{-\sum p_{s,m} \times log(p_{s,m})}{log(N_m)}, \qquad (1)$$

where $N_m$ is the number of different values on the meaning dimension (here COLOUR), and $p_{s,m}$ is the probability that signal variant $s$ and meaning value $m$ co-occur. $R_s$ effectively reflects how certain we are that a signal variant in $\sigma_1$ unambiguously encodes one COLOUR variant (Table 2).

In calculating the regularity of mapping for the whole signal dimension $\sigma_1$, we

Table 2. Co-occurrences of the signal variants of $\sigma_1$ and the meaning values of COLOUR in the language shown in Table 1.

| | | COLOUR | | | | | |
|---|---|---|---|---|---|---|---|
| | | *blue* | *red* | *yellow* | $R_s$ | $F_s$ | $RF_s$ |
| | `ki` | 1 | 8 | 0 | 0.682 | 9 | 6.142 |
| $\sigma_1$ | `pe` | 0 | 1 | 7 | 0.657 | 8 | 5.256 |
| | `tu` | 8 | 0 | 2 | 0.545 | 10 | 5.445 |

need to consider the $R$ values for every variant. Following usage-based models (Barlow & Kemmer, 2000), we also assume that frequency plays a crucial role in linguistic entrenchment, and hence the level of regularity which can be attributed to a construction. We therefore multiply the redundancy of each signal variant by its frequency in the language ($F$), obtaining a weighted redundancy value ($RF$). We now define $RegMap$ for a signal dimension $S$ with respect to a meaning dimension $M$ as the sum of $RF$ for each signal variant $s$, divided by the sum of frequencies for each variant[a]. This is further adjusted to take account of any discrepancy $d$ between the number of variants in $S$ and the number of variants in $M$, where $d$ is the greater of these divided by the lesser:

$$RegMap(S \rightarrow M) = \frac{\sum(RF_s)}{\sum(F_s)} \times \frac{1}{d} \tag{2}$$

Substituting the data from Table 2 into Eq. 2, therefore, yields a value for $RegMap(\sigma_1 \rightarrow \text{COLOUR})$ of $16.843/27 \times 1/(3/3) = 0.623$.

## 2.2. *RegMap for the entire language*

Table 3. $RegMap(S \rightarrow M)$ for all dimension pairs in the language.

| | | M | | | | | |
|---|---|---|---|---|---|---|---|
| | | COLOUR | SHAPE | INSET | $R_S$ | $F_S$ | $RF_S$ |
| | $\sigma_1$ | 0.623 | 0.008 | 0.008 | 0.881 | 0.639 | 0.563 |
| S | $\sigma_2$ | 0.000 | 1.000 | 0.000 | 1.000 | 1.000 | 1.000 |
| | $\sigma_3$ | 0.008 | 0.007 | 0.890 | 0.910 | 0.905 | 0.825 |

Table 3 shows $RegMap$ values for all combinations of signal and meaning dimensions, calculated using Eq. 2. Note that when $RegMap = 1$, there is an unambiguous representation of the meaning dimension by the signal dimension (e.g. $RegMap(\sigma_2 \rightarrow \text{SHAPE})$); when $RegMap = 0$, there is no information at all about the meaning dimension in the signal dimension (e.g. $RegMap(\sigma_2 \rightarrow \text{COLOUR})$). The values in Table 3 can be used to estimate the regularity of the whole language. First, we use Eq. 1 again, substituting signal and meaning dimensions for

---

[a] Each word occurs once here, so the sum of frequencies is the number of words in the language.

signal and meaning variants, to calculate the redundancy for a signal dimension $R_S$ across all meaning dimensions. This value is again weighted by the sum of all the $RegMap$ values for the signal dimension, yielding a modified redundancy value $RF_S$; this is averaged across all signal dimensions and again adjusted for any discrepancy $D$ between the number of signal dimensions $N_S$ and the number of meaning dimensions $N_M$ to produce a $RegMap$ value for the whole language:

$$RegMap(L_{S \to M}) = \frac{\sum(RF_S)}{N_S} \times \frac{1}{D} \tag{3}$$

It is important to re-emphasise that directionality in the mappings between signals and meanings is assumed in these calculations, and therefore that $RegMap(L_{S \to M})$, as illustrated in the exposition above, will not necessarily yield the same value as $RegMap(L_{M \to S})$ for the same language $L$. The latter measure can be calculated exactly as described above, with the co-ocurrence matrices in Tables 2 and 3 transposed before application of the equations.

## 3. Miniature artificial language learning experiments

We hypothesise that signal and meaning components which map each other systematically are more likely to be learnt and replicated than those with higher levels of ambiguity or uncertainty. To investigate this, we conducted two experiments using an artificial language learning task (Gómez & Gerken, 2000) with artificial languages structured like the one in Table 1, but with different $RegMap$ levels, as detailed in Table 4. 40 participants (14 males, 26 females; all students in their 20s) were randomly assigned to the four conditions; they were recruited through the Edinburgh University Careers website, and each paid £5 for participation.

Table 4. $RegMap$ values for the four conditions in Experiments 1 and 2.

| | Language 1 | Language 2 | Language 3 | Language 4 |
|---|---|---|---|---|
| $RegMap(L_{S \to M})$ | 0.143 | 0.455 | 0.754 | 1.00 |
| $RegMap(L_{M \to S})$ | 0.154 | 0.468 | 0.754 | 1.00 |

**Experiment 1. $RegMap$ from Signals to Meanings** Participants were asked to learn the meanings of words in an artificial language as best they could. During training, object-label pairs were presented on a computer monitor one at a time, and participants proceeded to the next pair by clicking the mouse in their own time (training duration: mean 10.2 mins, range 6.8-14.5). The whole language was shown three times, with breaks between each. Participants were then tested on the same objects they had seen in the training phase, and asked to type in the corresponding words for each object in the language they had learnt. We measured how well the structure of the signals produced by the participants mapped to the structure of the meanings provided (i.e. $RegMap(S \to M)$).

**Experiment 2. *RegMap* from Meanings to Signals** The experimental setup was identical to Experiment 1, except that in the testing phase participants saw screens showing one of the labels and all the objects; they were asked to click on the object that they thought corresponded to the label. In this experiment, we measured $RegMap(M \rightarrow S)$, or how well the meanings participants chose reflected the structure of the signals provided. Since the results of both experiments are comparable, they are presented and discussed together in the following sections.

**Results** We examine $RegMap$ for individual signal dimensions (syllables) with respect to the different meaning dimensions. For each signal and meaning dimension, Figure 1 shows the change in $RegMap$ between the input and output languages, for both signal and meaning dimensions. Signal and meaning dimensions show similar, but not identical, distributions. The three signal distributions are significantly different (one-factor ANOVA: $F = 19.554, d.f. = 2; p < 0.001$), as are the three meaning distributions (one-factor ANOVA: $F = 21.742, d.f. = 2; p < 0.001$).

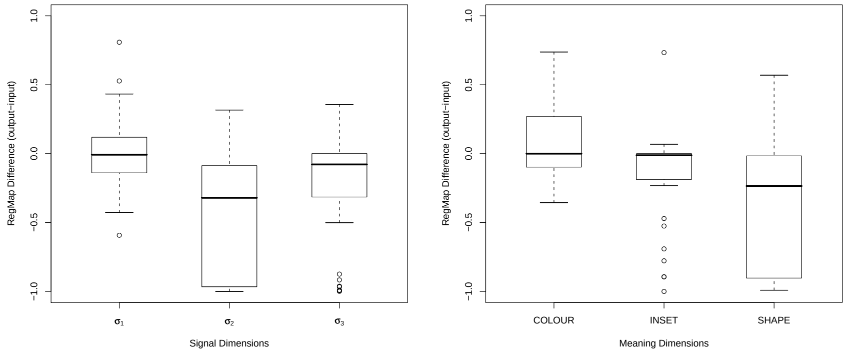

Figure 1. Change in $RegMap$ between input and output languages, by signal dimension (left) and meaning dimension (right). Plot shows inter-quartile range and median change.

Figure 2 shows $RegMap$ for the output languages plotted against $RegMap$ for the input languages provide to participants. Visual inspection of the plots in Figure 2 reveals a very high degree of individual variation, as all participants in each vertical row of data were exposed to exactly the same input language. Nevertheless, there is a significant effect of $RegMap$ for the input language on the resultant $RegMap$ in the output language, both for signals to meanings (single factor ANOVA: $F = 21.581; d.f. = 3; p < 0.001$) and for meanings to signals (single factor ANOVA: $F = 36.848; d.f. = 3; p < 0.001$).

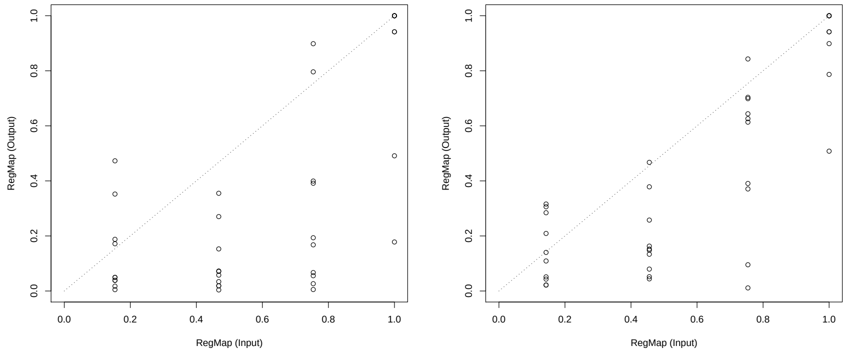

Figure 2. $RegMap(M \rightarrow S)$ (left) and $RegMap(S \rightarrow M)$ (right) showing the languages produced by participants as a function of the $RegMap$ of their input language. Vertically arranged datapoints (left to right) are from participants trained on languages 1-4; each point corresponds to one individual.

**Discussion** We note that in all these languages, COLOUR, SHAPE, INSET are mainly encoded in $\sigma_1, \sigma_2, \sigma_3$ respectively, which confounds the cause of the significant differences between signal and meaning dimensions in Figure 1; we plan to adapt the paradigm to explore these effects separately in future studies. Nevertheless, the results provide support to the well-established finding that word beginnings and endings are particularly salient (Jusczyk et al., 1999; Saffran et al., 1996) and that structure in the middle of signals is more susceptible to being lost. Our preliminary results suggest also that participants are sensitive to, and can reproduce, regularities in the mappings between signals and meanings at different levels, without explicit instruction; that there are great individual differences in these abilities and that, in some cases, $RegMap$ is greatly increased.

## 4. Conclusion

We have defined a novel metric to quantify the systematicity of languages, and measured how the metric is affected by individual learning. Learning generates new linguistic variants and thus provides an impetus for language change, yet also, since languages with higher levels of $RegMap$ are learnt with greater fidelity, the kind of learning quantified here offers a potential cultural mechanism for the accumulation of structure in language during cycles of learning from experience and transmission.

## Acknowledgements

Mónica Tamariz holds a Leverhulme Trust Early Career Fellowship; Andrew Smith is supported by Arts and Humanities Research Council Grant AR-112105.

### References

Barlow, M., & Kemmer, S. (2000). *Usage-based models of language.* University of Chicago Press.

Bod, R., Hay, J., & Jannedy, S. (Eds.). (2003). *Probabilistic linguistics.* MIT Press.

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

Bybee, J. L., & Hopper, P. J. (Eds.). (2001). *Frequency and the emergence of linguistic structure.* Amsterdam: John Benjamins.

Croft, W. (2000). *Explaining language change: an evolutionary approach.* Harlow: Pearson.

Croft, W. (2001). *Radical construction grammar: syntactic theory in typological perspective.* Oxford: Oxford University Press.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431-436.

Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, *4*(5), 178-186.

Jusczyk, P. W., Goodman, M. B., & Baumann, A. (1999). Nine-month-olds' attention to sound similarities in syllables. *Journal of Memory and Language*, *40*(1), 62-82.

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In E. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models* (pp. 173–203). Cambridge University Press.

Langacker, R. W. (1987). *Foundations of cognitive grammar: theoretical prerequisites* (Vol. I). Stanford, CA: Stanford University Press.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101-B111.

Pierrehumbert, J. B. (2006). The statistical basis of an unnatural alternation. In L. Goldstein, D. H. Whalen, & C. Best (Eds.), *Laboratory phonology viii* (p. 81-107). Mouton de Gruyter.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: the role of distributional cues. *Journal of Memory and Language*, *35*(4), 606-621.

Schyns, P. G., & Murphy, G. L. (1994). The ontogeny of part representation in object concepts. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 31, pp. 305–349). New York: Academic Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379-423 and 623-656.

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

Sloutsky, V. M., Lo, Y.-F., & Fisher, A. V. (2001). How much does a shared name make things similar? Linguistic labels, similarity, and the development of inductive inference. *Child Development*, *72*(6), 1695-1709.

Smith, K. (2003). Learning biases and language evolution. In *Proceedings of the 15th European Summer School on Logic, Language and Information.*

Smith, K., Smith, A. D. M., Blythe, R. A., & Vogt, P. (2006). Cross-situational learning: a mathematical approach. In P. Vogt, Y. Sugita, E. Tuci, & C. Nehaniv (Eds.), *Symbol grounding and beyond* (p. 31-44). Springer.