A Formal Model of Lexical Negation in Discrete Communication

Mikołaj Piotr Golecki
Université Paris Cité, CNRS, LLF
mikolaj.golecki@etu.u-paris.fr

Timothée Bernard
Université Paris Cité, CNRS, LLF
timothee-bernard@u-paris.fr

Abstract

Natural languages distinguish between objects satisfying a predicate and those satisfying its complement, often using a simple lexical negation. In emergent communication, however, a system may separate positive and negative meanings without developing a single negation marker: polarity may be tied to the thing being negated, distributed across multiple symbols, or reflected only in accidental correlations. We propose an information-theoretic account of negation applicable to discrete communication systems. We first study these metrics on toy languages, showing how they can be used to detect various patterns and how these are indeed related to negation. We then apply them to languages emerging in a signalling game with set-complement relations, under pressures known to favour compositionality. The results suggest that these pressures can produce high-scoring polarity-sensitive features, but not necessarily a compositional encoding of negation. More generally, we highlight both the usefulness and the limits of targeted semantic diagnostics for analysing structure in emergent languages.

1 Introduction

In natural language, the fact that the meanings of many complex expressions seem to be computable using only the meanings of their immediate subparts and the way they are combined is known as *compositionality* (Frege, 1892; Montague et al., 1970; Partee, 1995; Fodor and Lepore, 2002). Recent work shows that compositionality is not uniformly realised across language use, and that complex expressions in natural language retain a high degree of arbitrariness (Baroni, 2019; Dankers et al., 2022; Sathe et al., 2024). Limited forms of composition are also attested in some animal communication systems (Crockford and Boesch, 2005; Girard-Buttoz et al., 2022; Arnon et al., 2025).

What is distinctive about human language is therefore not simply the presence of regular combination, but its scale (Bortolato et al., 2023), systematicity (Pinker, 2003; Hurford, 2003), and cultural transmission (Kirby et al., 2008; Arnon and Kirby, 2024). Compositionality may be crucial in this respect, since finite, memorable components and reusable rules make it possible to express an infinite range of meanings.

In recent years, substantial work has investigated the emergence of language-like communication systems, including the conditions under which structured and compositional languages arise both in human participants (Kirby et al., 2008, 2014) and computational settings (a.o. Lazaridou et al., 2018; Dessì et al., 2019; Chaabouni et al., 2020; Baroni, 2019; Tucker et al., 2022; Bernard and Mickus, 2023; Akkerman et al., 2024). So far, this work has mostly focused on the emergence of *content words*: labels for objects, properties, spatial relations, commands for movement (e.g. *big blue bug*, *big red cube (to the) top left*, *go right*). Far fewer studies have investigated the emergence of functional or logical elements (Steinert-Threlkeld, 2020; Todd et al., 2020; Korbak et al., 2020), such as quantifiers, determiners, or negation. This is an important gap because content words are commonly interpreted as sets, and their composition via set intersection (e.g. $[\![\text{blue bug}]\!] = [\![\text{blue}]\!] \cap [\![\text{bug}]\!]$) or, equivalently, logical conjunction (e.g. $[\![\text{blue bug}]\!](x) = [\![\text{blue}]\!](x) \wedge [\![\text{bug}]\!](x)$) – the resulting expressions are said to be *trivially compositional* (Schlenker et al., 2016; Zuberbühler, 2018; Steinert-Threlkeld, 2020). By contrast, operators such as negation and quantification introduce non-intersective composition: the meaning of “not blue”, for instance, cannot be computed by intersecting the meaning of “blue” with any set. In this article, we focus on negation as a minimal instance of non-intersective operator.

Note however that a language that can express some predicates (or sets) and their negation (or complement) might not encode negation composi-

Third Workshop on the Bridges and Gaps between Formal and Computational Linguistics, pages 135–147
July 11, 2026 ©2026 Association for Computational Linguistics

tionally, or not with a single dedicated lexical item or syntactic construction. For instance, negation may be conflated with some predicates (cf. *present* and its negation *absent*) or distributed across multiple lexical items (French *ne ... pas* negation). Distinguishing between these possibilities is necessary for understanding emergent languages, comparing how artificial and natural languages encode negation, characterising typological variation in negation, and analysing how models represent semantic structure.

In this paper, we first propose information-theoretic metrics – *negation strength*, *homogeneity*, *value entanglement*, and *value conservation* – for detecting some forms of negation encoded in discrete communication systems, where signals are finite sequences of symbols. These metrics distinguish constructions that encode polarity compositionally from those that conflate polarity with predicate content, arise from distributed items, or reflect incidental correlations with polarity. We then apply these metrics on controlled toy languages to illustrate which kinds of negative constructions they can be used to detect. Finally, we present a signalling game that crucially revolves around set complementation and apply our metrics on languages obtained in an exploratory setting to study the emergence of negation. The associated experiment serves to illustrate the metrics for emergent language and provides preliminary validation; however, the results remain inconclusive regarding the types of pressure that reliably favour the development of a compositional encoding of negation. Code is available at https://github.com/TimotheeMickus/LEMURIA/.

2 A Formal Model of Negation in Discrete Communication

To express the negation of a predicate, languages sometimes use a different, arbitrary name (ex: *leave* for the negation of *stay*), a prefix (ex: *im* in *impossible*), a preposition (ex: *without*), a syntactic rule involving a dedicated lexical item (ex: *not*), and we can think of many other attested and non-attested solutions (such a simply changing the position of the predicate in the utterance, repeating the name of the predicate, or spelling it backward). In this paper, we focus on forms of negation expressed via one or more elements, be them words, parts of words, or any other sequences of symbols. For simplicity, we call these forms of negation *lexical negations* (in slight contrast with the standard terminology that takes the use of *not* in English to be a form of syntactic negation) to distinguish them from, among other constructions, the use of complex syntactic rules (e.g. ones involving movement or repetition).

In this section we first characterise lexical negation in a discrete communication channel through multiple information-theoretic measures. We then use these measures to define metrics aimed at quantifying negation; optimising said metrics over a set of constructions is a manner of detecting negation. Finally, we discuss examples and cases where the proposed metrics fail to capture intuitively valid encodings of negation, for instance when negation is expressed through multiple lexical items or when correlations unrelated to negation lead to false positives.

2.1 Desiderata for Detecting Negation

Consider a set U and a language $\mathcal{L}$ such that the elements of $\mathcal{L}$, called *signals*, are interpreted as subsets of U. For example, we could have a set of objects U, a signal $\mathrm{r} \in \mathcal{L}$, and r interpreted as the set $[\![\mathrm{r}]\!] \subseteq U$ of red objects. In such a setting, set complementation (in U) can be interpreted as negation: $U \setminus [\![\mathrm{r}]\!]$ is the set of non-red objects; in what follows, we note this set $\neg[\![\mathrm{r}]\!]$. If $\mathcal{L}$ contains many pairs of signals denoting a set and its complement, then these signals might use a lexical item that one would intuitively describe as a negation. How to detect this lexical item automatically?

Here we assume a set of *values* $\mathcal{V}$ (e.g. colours, such as blue and red) in bijection with the objects in U; in fact, we assume that each value is a singleton of U and vice-versa (e.g. each object is of a single colour, and no two distinct objects are of the same colour). We then define the set of *messages* to be $M = \mathcal{V} \cup \{\neg v \mid v \in \mathcal{V}\}$. For instance, if $\mathcal{V} = \{\{o_r\}, \{o_g\}, \{o_b\}\}$, corresponding intuitively to three colours (red, green, and blue), then $M = \{\{o_r\}, \{o_g\}, \{o_b\}, \{o_r, o_g\}, \{o_r, o_b\}, \{o_g, o_b\}\}$.

We assume that the signals are finite sequences of symbols; there is a finite *alphabet* Σ such that $\mathcal{L} \subset \Sigma^*$. We also assume that each signal expresses a message: $\forall s \in \mathcal{L},\ [\![\mathrm{s}]\!] \in M$, so $[\![\mathrm{s}]\!] \subseteq U$.

We call boolean random variables on signals *features* (e.g. the presence of a specific symbol), and our goal is to automatically detect features that are lexical negations as described above. To formalise this notion, we define V and N, the two random variables on messages such that $\forall v \in \mathcal{V}$,

- $V(v) = V(\neg v) = v$,
- $N(v) = +$ and $N(\neg v) = -$.

V is the *value variable* and N the *polarity variable*.

If we consider a set of colours and English restricted to expressions such as *red*, *not red*, *blue*, *not blue*, we observe that the presence of *not* entirely determines N and gives no indication on V. We note this feature $X_A \equiv \texttt{not}$; $X_A = 1$ if *not* appears in the signal, 0 otherwise. In information-theoretic terms (Shannon (1948); Weaver (1953); see also Cover and Thomas (2006) for an introduction) and noting $I(X;Y)$ for the *mutual information* between random variables X and Y, $I(X_A;N)$ is maximal and $I(X_A;V)$ is null.

We argue, however, that a feature X can qualify as a negation even if $I(X;N)$ is not maximal. That $I(X;N)$ be maximal corresponds to a form of *homogeneity*. Imagine a variant of English containing *not*, *red*, and *not-blue* in its vocabulary, such that *red* and *not not-blue* denote red and blue objects respectively, and that *not red* and *not-blue* denote not red and not blue objects respectively. Here, the presence of *not* gives, by itself, no indication on N, but just like in the first language, X_A fully determines N *given* V. In this language, X_A is a non-homogeneous negation. These considerations lead to the following definitions:

Negation Strength

$$n = \frac{I(X;N \mid V)}{H(N \mid V)} \tag{1}$$

Normalising on (conditional) entropy $H(N \mid V)$ coerces the value into a score in $[0,1]$. Candidate features for negation can be selected by optimising this metric.

While we do not take homogeneity to be a necessary property of negation, it is still an interesting one. We quantify it as follows:

Negation Homogeneity

$$h = 1 - \frac{I(X;V \mid N)}{H(V \mid N)} \tag{2}$$

If for some values, the feature X is used to indicate that $N{=}+$, and is used for other values to indicate that $N{=}-$, then given the value of N, knowing whether $X = 1$ helps determining the value of V; in other words, X carries information about V given N. The quantity above is 1 when X is used homogenously and decreases for features that are non-homogeneous.

Negation strength alone can be used to detect negation in simple cases such as the language:

$$\begin{aligned} [\![\text{red}]\!] &= red, \\ [\![\text{blue}]\!] &= blue, \\ [\![\text{not red}]\!] &= \neg red, \\ [\![\text{not blue}]\!] &= \neg blue \end{aligned} \tag{A}$$

Still considering the presence of *not*, X_A perfectly encodes negation and has a negation strength $n = 1$; similarly with $[\![\text{not not-blue}]\!] = blue$ instead of $[\![\text{blue}]\!]$ and $[\![\text{not-blue}]\!] = \neg blue$ instead of $[\![\text{not blue}]\!]$ in the language. However, we are also interested in *composite* features, of the form $X \equiv X_1 \vee X_2 \vee \cdots X_n$ where each X_i is the presence of a lexical item. Consider the following language:

$$\begin{aligned} [\![\text{red}]\!] &= red, \\ [\![\text{blue}]\!] &= blue, \\ [\![\text{not red}]\!] &= \neg red, \\ [\![\text{no blue}]\!] &= \neg blue \end{aligned} \tag{B}$$

In this language, $X_B \equiv (\texttt{not} \vee \texttt{no})$ has maximal negation strength, in contrast with both `not` alone and `no` alone. Considering composite features is useful when negation is lexicalised not by one but by multiple lexical items. But as explained below, when considering composite features, maximal negation strength does not guarantee that the feature under discussion indicates the presence of a negation in the intuitive sense.

Consider for instance the following language:

$$\begin{aligned} [\![\text{red}]\!] &= red, \\ [\![\text{blue}]\!] &= blue, \\ [\![\text{der}]\!] &= \neg red, \\ [\![\text{eulb}]\!] &= \neg blue. \end{aligned} \tag{C}$$

The feature $X_C \equiv \texttt{der} \vee \texttt{eulb}$ has maximal negation strength, however it does not indicate the existence of what we would call a lexical negation. More generally, if there is a bijection between messages and signals, the feature that is the disjunction of all of the signals used for negative messages will always have maximal negation strength (and some other features do too), even if the language is entirely arbitrary and no lexical negation exists.

To help us detect lexical negations, given some composite feature $X \equiv X_1 \vee X_2 \vee \cdots \vee X_n$, we note T the random variable $(X_1, X_2, \cdots, X_n)$ (whose values are in $\{0,1\}^n$). We now define:

Value Entanglement

$$t = \frac{I(T; V \mid X{=}1)}{H(V \mid X{=}1)} \tag{3}$$

$I(T; V|X{=}1)$ quantifies how much knowing which of the X_is are true tells us about V when $X{=}1$. For X_C in language C, t is maximal, revealing that the disjunct features of X_C together fully predict V. The metrics n and t provide solid ground for detecting negation; when negation strength is high and value entanglement is low, we argue that X is probably a lexical negation marker.

However, zero or low value-entanglement may be too strict a criterion. Indeed, value entanglement can be non-zero for cases of class-specific or *typed* negation[1] (it is the case for X_B in B). For instance, a simplified version of German might use the negation `kein` for nouns and `nicht` for adjectives. Consider the language including

$$\begin{aligned} [\![\text{kein Würfel}]\!] &= \neg cube, \\ [\![\text{kein Kugel}]\!] &= \neg sphere, \\ [\![\text{nicht rot}]\!] &= \neg red, \\ [\![\text{nicht blau}]\!] &= \neg blue, \end{aligned} \tag{D}$$

and their non-negated counterparts. For $X_D \equiv (\texttt{kein} \vee \texttt{nicht})$, $n = 1$ and $t = 0.5 > 0$, while X_D is arguably a high-quality negation marker. We believe that it is a negation marker because while T does carry information about V, this information is entirely redundant with what is found elsewhere in the signals (in the nouns and adjectives).

Let us formalise this idea. Given that we consider features of the form $X \equiv X_1 \vee X_2 \vee \cdots X_n$ where X_i is the random variable indicating the presence of some symbol $x_i \in \Sigma$, we can define R the random variable whose value for a signal s is the sequence of symbols obtained from s by removing all occurrences of any of $x_1, x_2, \cdots, x_n$. For instance, if $X \equiv \texttt{not}$, for the signal *not a white crow*, $R =$ *a white crow*. We can now define:

Value Conservation

$$c = \frac{I(R; V \mid X{=}1)}{H(V \mid X{=}1)} \tag{4}$$

Value conservation quantifies how much the signals, "beyond X" (so to speak), encode V. High n but low c indicate that the feature X under discussion separates the meaning space according to N while the remainder carries little value information. In unrestricted feature spaces, search may exploit such distributions to construct non-minimal or weakly interpretable candidates. We dub this risk *feature fabrication*.

Finally, we distinguish two notions of negation for a candidate feature X, based on the previously defined information-theoretic metrics. First, X is a *strong negation* when $n \approx 1$ and $t \approx 0$, i.e. when it reliably encodes polarity without relying on value-specific lexical items. Second, X is a *weak negation* when $n \approx 1$ and $c \approx 1$, i.e. when it reliably encodes polarity possibly relying on value-specific items but then that encode value-information that is redundant with that carried by the rest of the signals. These two notions are captured by the following definitions:

Strong Negativity

$$F_s = 2\frac{n(1-t)}{n + (1-t)} \tag{5}$$

Weak Negativity

$$F_w = 2\frac{nc}{n + c}. \tag{6}$$

2.2 Further Analysis

We now examine how the proposed metrics behave on other controlled languages.

Negation is typologically diverse in natural language (Dryer, 2013) and its realisation may intuitively reflect the strong/weak distinction proposed above. A relatively accessible example of this is French sentential negation, which in its literary form is often expressed by a preverbal particle *ne* together with a postverbal negative element such as *pas* (generic), *jamais* (*never*), *aucun* (*none*), etc. In simplified French with sentences such as [[x ne est pas un chat]] and [[x ne est aucun chat]] (*x is no/not a cat*), $X_{ne} = \texttt{ne}$ is a perfect, homogenous negation feature while $X_{pas \vee aucun} \equiv (\texttt{pas} \vee \texttt{aucun})$ is a weak, homogenous negation feature, although both appear necessary for expressing negation; the difference is that $X_{pas \vee aucun}$ carries additional information beyond polarity information (see scores in Appendix).

In contrast, by alleviating the constraint on marking nouns or adjectives in D, we obtain the language D' $= \{\text{nicht}, \text{kein}, \emptyset\} \times \{\text{Würfel}, \text{Kugel}, \text{rot}, \text{blau}\}$. The feature $X \equiv (\texttt{nicht} \vee \texttt{kein})$ in this language is a perfect negation marker, since *nicht* and *kein* are perfectly synonymous.

[1]This kind of negation is attested for natural language, see for instance Japanese (Nyberg, 2012, 6.3.1).

Note that in some extreme cases of *incomplete languages*, that is languages that do not contain a signal for all messages, value entanglement and value conservation may be undefined for some features that one might want to call a negation. Considering for instance

$$\begin{aligned} [\![\text{cube}]\!] &= cube, \\ [\![\text{not cube}]\!] &= \neg cube, \\ [\![\text{sphere}]\!] &= sphere, \\ [\![\text{pyramid}]\!] &= pyramid, \end{aligned} \tag{E}$$

feature $X \equiv \texttt{not}$ does look like a negation feature, but t and c are here undefined because $H(V \mid X = 1) = 0$.

Finally, the occurrence of negation might be correlated to other changes. Consider for instance the case of simple past sentences in English, in which the use of negation *not* triggers the apparition of auxiliary *did* and a change in verb form (e.g. from *I went* to *I did not go*). The presence of the past auxiliary or of some verb forms is thus correlated with that of the negative item, which has an impact on the scores of the corresponding features, even though only *not* would intuitively be called a negation. Still, this is not a serious issue, because the distributions of these features correlated with negation are significantly different from that of the actual negation *not*.

3 Application to Emergent Languages

In order to study the emergence of negation, we designed the signalling game (Lewis, 1969) described below. Many turns of the game are played in order to train two artificial agents – the *asker* and the *retriever* – who develop a shared language over the course of their training.

3.1 Game Definition

As above, we start by considering a set of arbitrary values $\mathcal{V}$ (e.g. $\{blue, red, \dots\}$, written $\{v_1, v_2, \cdots, v_{|\mathcal{V}|}\}$ below) in bijection with a set of objects; we see each value as a predicate on objects satisfied by exactly one of them. We then define the set of messages as the set of predicates $\mathcal{V} \cup \{\neg v \mid v \in \mathcal{V}\}$ – where $\neg v$ stands for $\lambda x.\ \neg v(x)$ –, containing each value and its logical negation; these messages/predicates are each assigned a unique index. We call predicates in $\mathcal{V}$ *positive* and predicates in $\{\neg v \mid v \in \mathcal{V}\}$ *negative*. Positive predicates are thus each satisfied by a single object while negative predicates are each satisfied by all objects save one.

At each turn of the game, the index i of a predicate is drawn uniformly at random. The asker observes i and produces a signal, which is a finite sequence of symbols. The retriever receives this signal along with a set of objects and must determine which of them satisfy the predicate corresponding to index i. Note that the objects are entirely symbolic; in our setting, there is one object per value in $\mathcal{V}$ and observing an object for the retriever means observing the associated value. The asker is shown only a categorical index because any representation reflecting the syntactic or semantic structure of the predicates would bias the asker toward a structured language (Słowik et al., 2021; Lazaridou et al., 2018) while our goal is to see whether set interactions as experienced by the retriever can induce compositional behaviour.

Both agents are implemented as neural networks, each built primarily around an LSTM (used as a decoder for the asker and as an encoder for the retriever). The asker includes a *predicate embedding matrix* and when it observes an index i, it accesses its i^{th} vector. Similarly, the retriever includes a *value embedding matrix* and when it observe an object, it accesses the embedding of the value associated with this object. The asker is trained using REINFORCE (Williams, 1992): for each object correctly recognised by the retriever, the asker's parameters are updated to stabilise the association between the observed index and the signal produced (and conversely, for each incorrectly recognised object). The retriever is trained using standard supervised learning to classify each input object as positive or negative given the signal it receives. Through repeated training, the asker can associate a meaning (a predicate/set of objects) with each index, a meaning that it learns to describe with signals that the retriever learns to interpret. In this situation, it is entirely possible that the agents develop a non-compositional language where each index is associated with an arbitrary, non-decomposable signal. Nevertheless, under certain constraints – limited memory, restricted symbol inventory – it might become advantageous for them to develop a compositional language.

The presence of semantically encoded negation in the game creates a communicative pressure (or advantage) for the development of a language that will encode negation systematically; a holistic language is by definition composed of one lexical item per message that the two agents have to agree on, here twice the number of values, while the development on a lexical negation can reduce the set of lexical item to only one per value plus one for negation.

3.2 Experiments

We conduct two experiments involving pressures known to incentivise compositionality in emergent language due to learning advantages and memory limitations (Spike, 2017, §2.5).

First, we introduce a *retriever reset* period n: every n training epochs the retriever is randomly reinitialised (while the asker is untouched). Kirby et al. (2008, 2014) show that compositionality presents a learning advantage when language is transmitted across generations, more so than within a stable population. In our case, we predict that the retriever regularly "forgetting" the language incentivises the agents to develop a compositional encoding of negation. We call an *epoch* a sequence of 250 *batches*, each batch corresponding to 2048 rounds of the game played in parallel, with two objects per predicate, exactly one satisfying it. We train the agents for 50 epochs across spaces of 2×4, 2×16, and 2×64 predicates (below, we often count in terms of predicate pairs, i.e. in terms of $|V|$) and retriever reset period of 1, 4, 10 and 50 epochs (the latter value is equivalent in practice to no retriever reset). Three randomly initialised runs are done per setting. Signal length is capped at 10 and alphabet size at the number of predicates +1 (for an end-of-sequence symbol 0), to allow arbitrary codes.

Then, we repeat the experiment with 2×64 predicates and add an *alphabet penalty* to the asker's training loss, i.e. a term that penalises alphabet usage.[2] The higher its value, the stronger is the pressure for the asker to use as few unique symbols as possible, possibly trading off game performance. A strong enough alphabet penalty would, among other things, encourage the asker to be more consistent, avoiding to use multiple symbols interchangeably (as if they were synonyms). In particular, encoding negation with a shared marker is then more economical than using multiple alternatives.

[2]Within a batch of rounds, for each round, the asker receives a penalty for each symbol occurrence in the signal it produces in this round. All symbols, globally in the batch, lead to the same penalty (set as a hyperparameter λ_{alph}) and this penalty is distributed equally over all occurrences of the same symbol.

We report the value of the metrics defined in Section 2 for the features that maximise them when the agents game performance is the highest (one evaluation is performed after each training epoch), defined as the accuracy of the retriever. We focus on this epoch because the metrics are only meaningful for a language that solves the communication task, and because, this study being exploratory, we do not track the full learning trajectory but aim to demonstrate the application of our metrics on successful protocols. Importantly, we do not examine all possible features (which would be intractable) but implement a bounded expansion search: after all non-composite features are built, growth is iterative, only the top-k (here: 72) features (ranked by negation strength) are kept as parents of the next expansion round, and disjunctions are only built up to 8 disjuncts. These should be interpreted as evidence only *within* this tractable subspace rather than exhaustive optima over all possible features. Appendix A indicates which epochs were analysed for each setup. For reference, Figure 3 there compares the scores discussed here with those at the final epoch: for the 64-pair setup, from which our main observations are drawn, the two are close (F_s differences below 0.06, with no systematic direction), indicating that the selected epoch is not an unrepresentative point in training.

3.3 Results

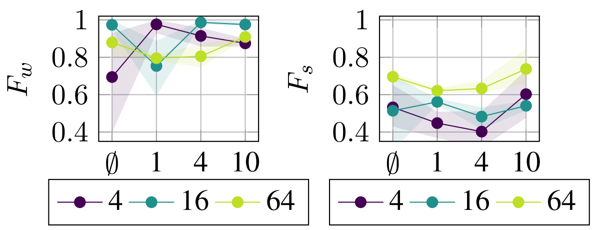

Figure 1: Mean top X F-scores with min–max range over 3 runs, by retriever reset period, with one curve per number of predicate pairs.

Retriever reset. Figure 1 show mean negation F-scores (F_s and F_w) with min-max values over the 3 seeds, depending on the number of predicate pairs and retriever reset period. Recall that a period of 50 effectively amounts to no retriever reset being performed. According to our definition of weak and strong negation, the weak kind dominates overall, though F_s is highest in the 64-pair setup. Within this setup, the 10-epoch reset period yields the highest mean F_s among the four tested intervals, based on a limited number of seeds. In these runs, the absence of retriever reset (period 50) does not appear to systematically lower the negation score of the top items.

Table 1 provides an example of the realisation of a relatively strong negation feature X ($F_s \approx 0.85$) from this 64-predicate pair setup. This language effectively uses $|\Sigma| = 89$ symbols, and the negation feature is a disjunction over 8 items.

v	**signal for** v	**signal for** $\neg v$
v_1	`114 0`	**`4`** `71 4 15 46 0`
v_2	`114` **`4`** `114 0`	`102 28 96 2 26 0`
v_3	`114 114 14 65 68 43 65 0`	**`126`** `2 88 20 88 0`

Table 1: Selected signal pairs where $X \equiv (1 \vee 126 \vee 3 \vee 36 \vee 4 \vee 41 \vee 83 \vee 85)$ occurs (on the negated, non-negated, and negated side, respectively – see bold symbols). Each line shows the signal produced for each of two predicates; a positive predicate on the left and the corresponding negative predicate on the right.

The size of the disjunction makes the resulting feature difficult to interpret as a compact lexical marker. At the same time, its scores are consistent with a relatively strong negation; n and c are both high (around 0.9), and value entanglement is low (around 0.18) – the symbols on which X relies do express some information beyond negation, but not much. It seems to us that the use of many different symbols here is due to a lack of incentive for the asker to reduce its use of the alphabet at its disposal.

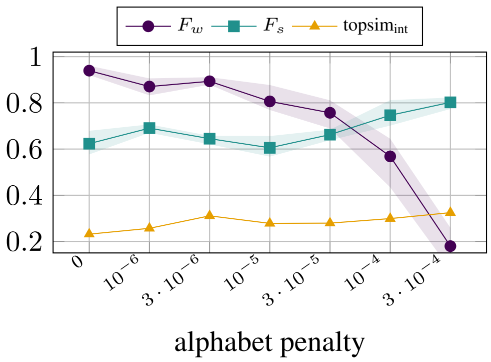

Figure 2: Mean top-1 negation scores with min–max range over 3 runs and mean message-length-normalised Levenshtein topographic similarity, by alphabet penalty, for a 64-predicate-pair setup with retriever reset period=10 over 50 epochs.

Alphabet penalty. We apply an alphabet penalty on the setup with the highest-ranking negation in terms of F_s (64 predicate pairs, retriever reset period = 10) and observe how this affects the languages that then emerge. We report the results for seven non-zero values for the hyperparameter λ_{alph}, up to $3 \cdot 10^{-4}$. Relative to the baseline ($\lambda_{\text{alph}} = 0$, see Table 2), game accuracy remains near-perfect up to $\lambda_{alph} = 10^{-4}$, but drops substantially at $3 \cdot 10^{-4}$; for even larger tested values, the penalty becomes so strong as to lead the asker to produce empty signals (starting and ending with the end-of-sequence symbol 0).[3] This indicates that a moderate pressure toward low alphabet use does not prevent successful communication, while a stronger pressure leads to degenerate strategies.

Across our selected range of λ_{alph} values, F_w decreases overall as the penalty increases (save a non-monotone dip at the lowest non-zero value), with a sharp drop at higher values. By contrast, F_s shows a non-monotone pattern across the range: the mean over 3 runs decreases initially with alphabet penalty, and then starts rising, dominating F_w for the two highest penalty values. This is consistent with high alphabet penalties reducing value entanglement, yielding a stronger negation. Taken together, these trends point to a trade-off. Alphabet penalty constrains vocabulary use, which correlates with a decrease in value entanglement (t, associated with higher F_s) and a decrease in value conservation (c, associated with lower F_w): symbol reuse then appears to promote a more general polarity marking at the cost of compositional separability of predicate-value information. At the highest tested penalty the emergent language approaches degeneracy and the F_s gain there may be a consequence of vocabulary compression and not actual emergence.

Figure 2 also shows topographic similarity (Brighton and Kirby, 2006), the correlation between distances between messages on the one hand and distances between the signals produced to convey the messages on the other hand. Topographic similarity (*topsim*) is computed as Spearman's ρ between pairwise message distances – defined here as the cosine distance between the asker's predicate embeddings – and pairwise signal distances – defined here as the normalised Levenshtein distance on signals – for the language obtained at the end of a training epoch.[4] We include it as a reference

[3] See the previous footnote concerning λ_{alph}. For comparison, the base reward is $\pm\frac{1}{2}$ for each object (in)correctly classified by the retriever.

[4] Because we define message distances as distances between the model's learned predicate embeddings, we call the resulting topographic similarity *intensional*. Instead of these learned embeddings, an *extensional* alternative would consist in using the boolean vectors indicating, for a given list of objects, which ones satisfy the predicates. We do not find this

alpha. penalty	F_w	ΔF_w	F_s	ΔF_s	acc.	Δacc.
baseline	0.909	+0.000	0.741	+0.000	1.000	+0.000
0	0.939	+0.030	0.623	-0.118	0.999	-0.000
10^{-6}	0.871	-0.038	0.691	-0.050	0.999	0.000
$3 \cdot 10^{-6}$	0.893	-0.016	0.645	-0.096	0.999	-0.001
10^{-5}	0.806	-0.103	0.605	-0.136	1.000	0.000
$3 \cdot 10^{-5}$	0.757	-0.152	0.662	-0.079	0.994	-0.006
10^{-4}	0.568	-0.341	0.746	+0.005	0.990	-0.010
$3 \cdot 10^{-4}$	0.179	-0.730	0.802	+0.061	0.842	-0.158

Table 2: Game accuracies and top X mean F-scores for vocabulary penalty values relative to the baseline from the first experiment at properties $= 64$ and retriever reset interval $= 10$. Deltas are alpha-pen. minus baseline.

point because it is common in emergent language work. In the present experiments, topographic similarity shows only limited variation across alphabet-penalty conditions. This is consistent with previous work reporting that topographic similarity fails to detect non-trivial forms of compositionality such as the one involved with negation (Korbak et al., 2020).

v	**signal for** v	**signal for** $\neg v$
v_1	`28 0`	**`99 99`** `0`
v_2	**`99 99`** `0`	`93 0`
v_3	**`78`** `0`	**`99 99`** `0`

Table 3: Selected signal pairs where $X \equiv (78 \vee 99)$ occurs, in a game with $\lambda_{alph} = 3 \cdot 10^{-4}$.

The following examples suggest that in some settings the current feature-selection procedure can recover high-scoring, composite features, whose interpretation as compositional negation markers remains unclear.

The case at $\lambda_{alph} = 3 \cdot 10^{-4}$ illustrates the vocabulary compression effect. The emergent language is not entirely degenerate, but still highly constrained: symbol `99`, involved in the identified feature $X \equiv (78 \vee 99)$, is also the most frequent symbol, accounting for about 45% of all tokens. Table 3 shows that X often separates negative from positive values, but non-homogeneously and in fact non-systematically. In the first row, X is active only on the negative side through `99`; in the second, only on the positive side through `99`; and in the third, both on the positive side through `78` and the negative side through `99`. Thus, X tracks polarity to a certain extent (n is high) without corresponding to a clean lexical negation item.

The language obtained at $\lambda_{alph} = 10^{-4}$ (Table 4) yields much better game accuracy, but also heavily makes use of a few frequent symbols in many signals, including the symbol `117`, involved in the detected composite feature. We note however that the second pair in Table 4 contains the feature on both sides; the other pairs contain it only on one side. The feature's high F_s suggests a strong negation according to our definitions, but its realisation is not uniform across pairs, which makes it difficult to interpret as a lexical marker.

v	**signal for** v	**signal for** $\neg v$
v_1	**`78`** `74 0`	`15 15 0`
v_2	**`117 117`** `4 0`	**`117`** `3 0`
v_3	`29` **`34`** `29 0`	`29 29 0`

Table 4: Selected signal pairs where $X \equiv (117 \vee 34 \vee 36 \vee 78)$ occurs, in a game with $\lambda_{alph} = 10^{-4}$.

4 Discussion

The toy-language analysis supports the intended interpretation of the metrics and shows that they distinguish different negation encodings and degenerate features. In the emergent languages, the main difficulty is to determine whether the highest-scoring candidates correspond to interpretable lexical items. In some successful settings, strong negation is realised by large disjunctions, and qualitative analysis shows that several disjunctions differing by only one or two disjuncts score almost identically. This pattern suggests that our search procedure may recover distributed or non-minimal realizations of negation, rather than the compact markers one might hope to observe. The present exploratory study cannot determine whether this reflects actual distribution in the emergent language or over- or under-expansion in the search procedure.

This suggests that feature selection is also an important issue, beyond metric design. Bounding disjunction size makes the search tractable and may reduce the risk of over-expansion, if it is indeed relevant. Conversely, if a large disjunction is ob-

alternative notion of distance satisfying; for instance, for two distinct values v_1 and v_2, it takes v_1 and v_2 to be very close, and v_1 and $\neg v_1$ to be maximally distant.

tained by gradual expansion of an initially good candidate, and each added item only marginally improves negation strength, the score may be only partially fabricated rather than entirely spurious. Investigating this trade-off more directly appears to be an important direction for future work.

The fact that topographic similarity does not target the encoding of any specific semantic operation (in particular negation) justifies the use of targeted metrics such as ours. Meaning-form correlation may miss local structure tied to a specific semantic operation. At the same time, high negation metric scores do not imply that the language as a whole is compositionally organised. Global and targeted measures may capture different aspects of structure and should be interpreted jointly.

5 Limitations

This study has several limitations. This framework is designed for discrete communication channels with symbol-based features, and does not directly extend to settings where polarity is encoded in continuous representations. The formal setup is limited to atomic unary predicates and their complement/negation, and therefore does not cover richer semantic phenomena such as scope interactions. The empirical study is exploratory and computationally bounded, with few runs and a restricted top-k pruning strategy, rather than exhaustive feature discovery. The current search space is restricted to single symbol-presence features and their disjunctions, and does not consider conjunctions of features, repetitions, and more complex combinations. High-scoring candidates in emergent languages remain difficult to interpret, since disjunctions may reflect genuine distributed encoding, feature fabrication, or both. We do not yet establish principled thresholds for how high t or c must be, nor how strongly they ought to be weighted relative to n, in deciding whether a feature can count as a "good" negation marker. Moreover, candidate selection for feature expansion depends on negation strength, but other metrics, namely t, could play an important role in selection. Many computational pressures associated with the emergence of compositionality have not been considered; in particular, an additional penalty on signal length could complement the alphabet penalty.

References

Daniel Akkerman, Phong Le, and Raquel G. Alhama. 2024. The Emergence of Compositional Languages in Multi-entity Referential Games: from Image to Graph Representations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18713–18723, Miami, Florida, USA. Association for Computational Linguistics.

Inbal Arnon and Simon Kirby. 2024. Cultural evolution creates the statistical structure of language. *Scientific Reports*, 14(1):5255.

Inbal Arnon, Simon Kirby, Jenny A. Allen, Claire Garrigue, Emma L. Carroll, and Ellen C. Garland. 2025. Whale song shows language-like statistical structure. *Science*.

Marco Baroni. 2019. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1791):20190307.

Timothée Bernard and Timothee Mickus. 2023. So many design choices: Improving and interpreting neural agent communication in signaling games. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8399–8413, Toronto, Canada. Association for Computational Linguistics.

Tatiana Bortolato, Roger Mundry, Roman M. Wittig, Cédric Girard-Buttoz, and Catherine Crockford. 2023. Slow development of vocal sequences through ontogeny in wild chimpanzees (pan troglodytes verus). *Developmental Science*, 26(4):e13350.

Henry Brighton and Simon Kirby. 2006. Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial Life*, 12(2):229–242.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. 2020. Compositionality and Generalization In Emergent Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4427–4442, Online. Association for Computational Linguistics.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience.

Catherine Crockford and Christophe Boesch. 2005. Call combinations in wild chimpanzees. *Behaviour*, 142(4):397–421.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. 2022. The Paradox of the Compositionality of Natural Language: A Neural Machine Translation Case Study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175, Dublin, Ireland. Association for Computational Linguistics.

Roberto Dessì, Diane Bouchacourt, Davide Crepaldi, and Marco Baroni. 2019. Focus on What's Informative and Ignore What's not: Communication Strategies in a Referential Game. *arXiv preprint*. ArXiv:1911.01892 [cs].

Matthew S. Dryer. 2013. Negative morphemes (v2020.4). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Jerry A Fodor and Ernest Lepore. 2002. *The compositionality papers*. Oxford University Press.

Gottlob Frege. 1892. On sense and reference.

Céline Girard-Buttoz, Emiliano Zaccarella, Tania Bortolato, and 1 others. 2022. Chimpanzees produce diverse vocal sequences with ordered and recombinatorial properties. *Communications Biology*, 5:410.

James R. Hurford. 2003. The language mosaic and its evolution. In Morten H. Christiansen and Simon Kirby, editors, *Language Evolution*, pages 38–57. Oxford University Press.

Simon Kirby, Hannah Cornish, and Kenny 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.

Simon Kirby, Tom Griffiths, and Kenny Smith. 2014. Iterated learning and the emergence of language. *Current Opinion in Neurobiology*, 28:108–114.

Tomasz Korbak, Julian Zubek, and Joanna Rączaszek-Leonardi. 2020. Measuring non-trivial compositionality in emergent communication. *arXiv preprint*. ArXiv:2010.15058 [cs].

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. 2018. Emergence of linguistic communication from referential games with symbolic and pixel input. *Preprint*, arXiv:1804.03984.

David Lewis. 1969. *Convention: A Philosophical Study*. Harvard University Press, Cambridge, MA.

Richard Montague and 1 others. 1970. Universal grammar. *1974*, pages 222–46.

Joacim Nyberg. 2012. Negation in japanese. Bachelor's thesis, Lund University, Lund, Sweden.

Barbara H. Partee. 1995. Lexical semantics and compositionality. In Lila R. Gleitman and Mark Liberman, editors, *An Invitation to Cognitive Science: Language*, pages 311–360. MIT Press, Cambridge, MA.

Steven Pinker. 2003. Comprehension and production in language evolution. In Morten H. Christiansen and Simon Kirby, editors, *Language Evolution*, pages 16–37. Oxford University Press.

Abhijit Sathe, Evelina Fedorenko, and Noga Zaslavsky. 2024. Language use is only sparsely compositional: The case of english adjective-noun phrases in humans and large language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Philippe Schlenker, Emmanuel Chemla, Anne M. Schel, James Fuller, Jean-Pierre Gautier, Jeremy Kuhn, Dunja Veselinović, Kate Arnold, Cristiane Cäsar, Sumir Keenan, Alban Lemasson, Karim Ouattara, Robin Ryder, and Klaus Zuberbühler. 2016. Formal monkey linguistics. *Theoretical Linguistics*, 42(1-2):1–90.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Agnieszka Słowik, Abhinav Gupta, William L. Hamilton, Mateja Jamnik, Sean B. Holden, and Christopher Pal. 2021. Structural inductive biases in emergent communication. *Preprint*, arXiv:2002.01335.

Matthew John Spike. 2017. *Minimal Requirements for the Cultural Evolution of Language*. Ph.D. thesis, The University of Edinburgh.

Shane Steinert-Threlkeld. 2020. Toward the emergence of nontrivial compositionality. *Philosophy of Science*, 87(5):897–909.

Graham Todd, Shane Steinert-Threlkeld, and Christopher Potts. 2020. Learning compositional negation in populations of roth-erev and neural agents. *Preprint*, arXiv:2012.04107.

Mycal Tucker, Roger Levy, Julie A Shah, and Noga Zaslavsky. 2022. Trading off utility, informativeness, and complexity in emergent communication. In *Advances in Neural Information Processing Systems*, volume 35, pages 22214–22228. Curran Associates, Inc.

Warren Weaver. 1953. Recent contributions to the mathematical theory of communication. *ETC: a review of general semantics*, pages 261–281.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256.

Klaus Zuberbühler. 2018. Combinatorial capacities in primates. *Current Opinion in Behavioral Sciences*, 21:161–169.

A Appendix

msg. pairs	retriever reset interval	n	mean epoch	min	max	unique epochs
4	$\emptyset$	3	0.0	0	0	0
4	1	3	0.7	0	1	0, 1
4	4	3	2.0	0	6	0, 6
4	10	3	0.3	0	1	0, 1
16	$\emptyset$	3	17.7	6	40	6, 7, 40
16	1	3	19.0	2	45	2, 10, 45
16	4	3	23.3	3	49	3, 18, 49
16	10	3	13.0	5	26	5, 8, 26
64	$\emptyset$	3	34.0	27	43	27, 32, 43
64	1	3	46.7	44	49	44, 47, 49
64	4	3	33.0	30	35	30, 34, 35
64	10	3	44.0	35	49	35, 48, 49

Table 5: Epochs analyzed for the finite retriever reset interval-only study F-score plots (best accuracy).

voc. penalty	n	mean epoch	min	max	unique epochs
0	3	32.3	19	49	19, 29, 49
10^{-6}	3	35.0	28	49	28, 49
$3 \cdot 10^{-6}$	3	35.3	29	39	29, 38, 39
10^{-5}	3	30.0	16	49	16, 25, 49
$3 \cdot 10^{-5}$	3	41.3	26	49	26, 49
10^{-4}	3	33.3	16	46	16, 38, 46
$3 \cdot 10^{-4}$	3	48.0	46	49	46, 49

Table 6: Epochs analyzed for the voc-penalty study F-score plots (best accuracy). Message pairs=64, retriever reset interval=10.

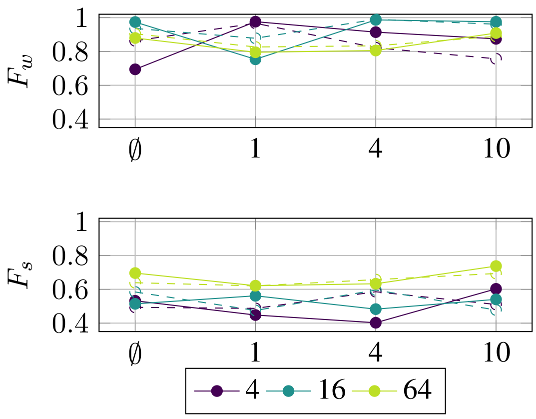

Figure 3: Mean negation F-scores at peak-accuracy epoch (solid, filled) and final epoch (dashed, open), by retriever reset period, with one curve per number of predicate pairs. See Figure 1 for the main result.

Cat sentences				
P0-v0	I am a cat		I am not a cat	
P0-v1	the cat sat on the mat		the cat did not sit on the mat	
P0-v2	the cat is hairy		the cat is not hairy	
P0-v3	curiosity killed the cat		curiosity did not kill the cat	
feat.	n	h	t	c
`not`	1	1	0	1
`did`	0.5	0.75	0	1
`kill`$\vee$`sit`	0.5	0.75	1	1
`cat`	0	1	0	1
French cats				
P0-v0	je suis un chat		je ne suis pas un chat | je ne suis aucun chat	
	I am a cat		*I am not a cat* | *I am no cat*	
P0-v1	le chat a déjà été sur la branche		le chat ne a jamais été sur la branche	
	the cat has already been on the branch		*the cat has never been on the branch*	
P0-v2	le chat a été chez le docteur		le chat ne a pas été chez le docteur	
	the cat went to the doctor		*the cat did not go to the doctor*	
feat.	n	h	t	c
`ne`	1	1	0	1
`pas`	0.41	0.81	0	1
`aucun`$\vee$`jamais`$\vee$`pas`	1	1	0.67	1
`aucun`$\vee$`déjà`$\vee$`pas`	1	0.44	0.67	1
`déjà`	0.3	0.74	$\emptyset$	$\emptyset$
Twisted logic				
P0-v0	I am present | I am not absent		I am not present | I am absent	
P0-v1	I am anxious | I am not calm		I am not anxious | I am calm	
P0-v2	I am here		I am not here	
feat.	n	h	t	c
`absent`$\vee$`calm`$\vee$`not`	0.45	0.94	0.53	0.61
`not`	0.2	0.89	0	1
`absent`$\vee$`calm`	0	0.89	1	0
Untwisted logic				
P0-v0	I am present		I am not present	
P0-v1	I am anxious		I am not anxious	
P0-v2	I am here		I am not here	
P1-v0	I am not absent		I am absent	
P1-v1	I am not calm		I am calm	
feat.	n	h	t	c
`not`	1	0.58	0	1
`I`	0	1	0	1
`anxious`	0	0.69	$\emptyset$	$\emptyset$

Table 7: Natural-language illustration with feature scores.

$\mathcal{L}_1$ **Perfect negation**	+	-
P0-v0	1	4 1
P0-v1	2	4 2
P0-v2	3	4 3

feat.	n	h	t	c
4	1	1	0	1

$\mathcal{L}_2$ **Holistic encoding**	+	-
P0-v0	1	4
P0-v1	2	5
P0-v2	3	6

feat.	n	h	t	c
1∨2∨3	1	1	1	0
4∨5∨6	1	1	1	0
1∨2∨6	1	0.42	1	0
1∨2	0.67	0.71	1	0

$\mathcal{L}'_2$ **Disjoint holistic enc.**	+	-
P0-v0	1	4 7
P0-v1	2	5 8
P0-v2	3	6 9

feat.	n	h	t	c
4∨5∨6	1	1	1	1
4∨5∨9	1	1	1	1
7∨8∨9	1	1	1	1

$\mathcal{L}''_2$ **D. h. enc. with variation**	+	-
P0-v0	1	4 7
P0-v1	2	5 8
P0-v2	3	6 9 | 4 9

feat.	n	h	t	c
4∨5∨6	1	1	0.67	1
4∨6∨8	1	1	0.67	1
2∨4∨6	1	0.44	0.67	1
1∨5∨9	1	0.44	1	1

$\mathcal{L}_3$ **Negation variants**	+	-
P0-v0	1	5 1
P0-v1	2	5 2
P0-v2	3	6 3
P0-v3	4	6 4

feat.	n	h	t	c
5∨6	1	1	0.5	1
5	0.5	0.75	0	1
6	0.5	0.75	0	1

$\mathcal{L}_4$ **Inhomogenous negation**	+	-
P0-v0	1	4 1
P0-v1	4 2	2
P0-v2	3	4 3

feat.	n	h	t	c
4	1	0.42	0	1
1∨2∨3	0	0.42	∅	∅

Table 8: Toy languages with negation-metric scores.