École Normale Supérieure de Lyon

Bachelor of Mathematics

Final internship report

Optimal lexicalization

Claire RONG

under the supervision of

Benjamin SPECTOR (CNRS, ENS)
and **Emmanuel CHEMLA** (CNRS, ENS)

Institut Jean Nicod
Department of Cognitive Studies
École Normale Supérieure de Paris
May - July 2023

Contents

Acknowledgements

I am very grateful to my co-supervisor during the internship, Emmanuel Chemla, for suggesting many of the topics addressed here and for dedicating much of his time to provide guidance and documentation for this report. My deepest thanks go to my supervisor, Benjamin Spector, for accepting me as an intern and for his constant support throughout the internship. He also gave me the opportunity to attend seminars at Institut Jean Nicod, generously dedicated his time to satisfy my curiosity about various questions in linguistics, and was instrumental in my decision to pursue a master's degree in linguistics. Lastly, I would like to thank my friends Marguerite, Pauline, and Nicolas for reviewing and providing feedback on this work.

1 Introduction

Understanding how meanings are conveyed through messages is a fundamental aspect of communication, and we wish to find the most efficient ways to achieve this. Specifically, this paper aims to determine the optimal way to assign meanings to a set of messages in a highly idealized setting.

Central to our discussion is the concept of "expected utility," a key notion in decision theory and economics. Expected utility allows us to quantify the desirability or value associated with uncertain events or outcomes – the outcome here being the messages we choose to say. In this context, we examine how expected utility measures the efficiency of a language and can guide the allocation of meanings to messages in an optimal manner[1]. To introduce the problems we are dealing with, let us start with a simple scenario:

[1]This way of measuring the efficiency of a language has roots in information theory and has been recently developed for linguistic pragmatics, giving rise to a formal framework known as the Rational Speech Act model (cf Bergen, Levy, and N. Goodman 2016 and N. D. Goodman and Stuhlmüller 2013). This model defines a notion of the utility of a message in a given situation for a speaker. With extra assumptions, one can then think about the expected utility of a language as a whole and try to explain certain universal tendencies as arising from maximization of this expected utility. This perspective is for instance adopted in Enguehard and Spector 2021, who offer an account for why a quantifier such as 'not all' is never lexicalized across languages.

Two messages[2] m_1 and m_2 are used to communicate events e_1 and e_2 which have prior probabilities of p_1 and p_2 with $p_1 > p_2$, and let us call p_{12} the probability of their conjunction – that is, e_1 and e_2 being both true. In information theory, the "information content" of an event with probability p is given by $-\log(p)$. (We will write log for logarithm to the base e.) The function $p \mapsto -\log(p)$ is decreasing over $[0, 1]$. Hence m_2 which has a lower probability conveys more information than m_1 because it is *more precise* or *less common* than m_1. To understand this intuitively, let us take two examples from everyday language. First imagine that m_1 = 'mammal' and m_2 = 'rabbit'. The messages are not disjoint here. Rabbits are part of the mammals, therefore m_2 is *more precise*. We can also imagine two messages that are disjoint, for instance m_1 ='right-handed' and m_2 = 'left-handed'. There are much more right-handed people than left-handed people. Most of the time, being right-handed is not "worth mentioning". Lefties are *less common*, so saying that someone is left-handed is more informative.

Consider a rational speaker who wants to maximize information transmission and can only use (at most) one message. The term 'rational' will be mathematically defined further on, for this example it just means that the speaker has a goal in mind (to maximize efficiency) when they communicate and doesn't choose their messages at random. They will use m_2 when m_2 is true, irrespective of the truth-value of m_1, because m_2 is more informative than m_1. They will use m_1 whenever m_1 is true and m_2 is false. When both messages are false they will say nothing. Therefore, the expected value (in a probabilistic sense) of the information the speaker will convey i.e. the expected utility is

$$-p_2 \log(p_2) - (p_1 - p_{12}) \log(p_1)$$

How should we choose p_1, p_2 and p_{12} so as to maximize this quantity ?

With no assumptions about independence, the answer is $p_1 = p_2 = \frac{1}{e}$ and $p_{12} = 0$. It contradicts the assumption that $p_1 > p_2$ but it shows that the optimal lexicalization for two messages and one 'spot' to say a message is to have two incompatible messages having the same probability $\frac{1}{e}$. These values correspond to a utility of $\frac{2}{e}$.

If we constraint the messages to be independent, then we will try to maximize

$$-p_2 \log(p_2) - (1 - p_2)p_1 \log(p_1)$$

This is achieved by taking $p_1 = \frac{1}{e}$ and $p_2 = \frac{1}{e^{1+\frac{1}{e}}}$, corresponding to a utility of $\frac{1}{e} + \frac{1}{e^{1+\frac{1}{e}}}$ which is less than in the non-independent case.

One may also wonder what the most beneficial addition (in terms of maximizing expected utility) to our set of two messages would be: a negation or a third message? At first glance, negation would seem more beneficial because we would then be able to communicate 4 messages: $m_1, m_2, \neg m_1$ and $\neg m_2$. However, a simple reasoning proves that negation is in fact never more useful than an additional message (but this is only valid in the case where only one message can be used). Among the 3 most informative messages of the 4, one is the negation of another, and so on every occasion at least one of the 3 is true. Hence the least informative message of the 4 cannot be used at all because in any situation at least one of the other three messages is true and will be chosen instead. This means that we are going to use only 3 messages, with the constraint that one is the negation of another. The best lexicalization that can be achieved under this constraint can also be achieved by a system with 3 messages under no specific constraint. So the best lexicalization with [2 messages + negation] is not better than the best lexicalization with 3 messages.

[2] A 'message' can simply be understood as a symbol, a signal. For instance, in animal communication, that could be a gesture or a vocal signal denoting a predator.

As a side note, if the listener is "smart", they will understand that m_2 is false when the speaker uses m_1, so in a sense the message is now interpreted as meaning "m_1 and not m_2". Likewise, the listener will understand from "silence" that both are false. Then the expected information conveyed by the language would be

$$-p_2 \log(p_2) - (p_1 - p_{12}) \log(p_1 - p_{12}) - (1 - (p_1 + p_2 - p_{12})) \log(1 - (p_1 + p_2 - p_{12}))$$

We can define three new probabilities $p'_2 = p_2, p'_1 = p_1 - p_{12}$ and $p'_3 = 1 - (p_1 + p_2 - p_{12})$ that correspond respectively to three disjoint events: e_2, $e_1 \wedge \neg e_2$, $\neg e_1 \wedge \neg e_2$. It will be shown further on that the best lexicalization assigns a probability of $\frac{1}{3}$ to each of the disjoint events. Solving for the original p_i's, we get $p_1 = \frac{2}{3}$, $p_2 = \frac{1}{3}$ and $p_{12} = \frac{1}{3}$.

We will tackle these kinds of questions for various setups, taking into account the following parameters:

1. N, the number of (atomic) messages the language contains
2. S, the number of messages that can be used in one utterance
3. whether the messages are probabilistically independent
4. whether the language includes a negation

2 Theoretical framework

- We fix once and for all a certain probability space (Ω, F, P), with the condition that for any real number x in $[0, 1]$, there is an event E (i.e. an element of F) such that $\mathsf{P}(E) = x$. We call **worlds** the elements of Ω.
- We consider a language with a vocabulary consisting of a set of N propositional atoms $a_1, \ldots, a_N$ and possibly a negation notated $\neg$ (we will consider both languages with and without negation.)
- A **literal** is either an atom a or the concatenation of the negation symbol with an atom ($\neg a$). If the language does not have negation, a literal is simply an atom.
- S (for 'spots') denotes the number of literals from $\mathcal{L}$ that can be uttered on a particular occasion. The set of messages $M(\mathcal{L})$ of this language is the set of all subsets of literals that contain at most S literals.
- A **semantics** σ (or **lexicalization**) for $\mathcal{L}$ is a function from $M(\mathcal{L})$ to events with positive probability [3] such that (if the message is a singleton, e.g. $\{l\}$, then instead of writing $\sigma(\{l\})$, we simply write $\sigma(l)$): if the language includes negation, then for every atom a, $\sigma(\neg a) = \Omega \backslash \sigma(a)$; the event corresponding to a message is the intersection of all events corresponding to each literal

[3] The restriction to events with positive probability is not necessary but makes a few things easier, and can be made without loss of generality. This is because we are interested in the optimal semantics, and a semantics in which a message m denotes an event with probability 0 is such that the contribution of m and $\neg m$ to the expected utility of a language will necessarily be 0 ($-1 \times \log(1) = 0$ and under the convention $-0 \times \log(0) = 0$), so that it can never be worse to instead map m to some event with positive probability.

in the message i.e. if m is a message $\{l_1, \ldots, l_k\}$, then $\sigma(m) = \bigcap_{1 \leq i \leq k} \sigma(l_i)$. As a reminder, those events are –in this setting– measurable subsets of $[0, 1]$. Saying that a message is true at a given world (i.e. a real number in $[0, 1]$) is simply saying that this world belongs to the event corresponding to the message.

- Given a world $w \in \Omega$, we say that a message m is **optimal** (for $\{\mathcal{L}, \sigma\}$) in w if $w \in \sigma(m)$ (i.e. m is true in w), and for every message m' which is true in w, $\mathsf{P}(\sigma(m)) \leq \mathsf{P}(\sigma(m'))$.
- A **speaker** s of $\mathcal{L}$ paired with a semantics σ – a speaker of $\{\mathcal{L}, \sigma\}$ for short – is a function from Ω to $M(\mathcal{L})$. We notate m^s_w the image of w under this function i.e. m^s_w is the message that s uses in w.
- A **truthful** speaker of $\{\mathcal{L}, \sigma\}$, is a speaker S of $\{\mathcal{L}, \sigma\}$ such that for every world w, $w \in \sigma(m^s_w)$.
- A **rational** speaker of $\{\mathcal{L}, \sigma\}$ is a speaker S of $\{\mathcal{L}, \sigma\}$ such that for every world w, m^s_w is $\{\mathcal{L}, \sigma\}$-optimal in w.
- Given a pair $\{\mathcal{L}, \sigma\}$, the **utility** (or **informativity**) of a message m in world w, notated $u(m|w)$, is $-\infty$ if m is false in w (i.e. $w \notin \sigma(m)$), and is equal to $-\log(\mathsf{P}(\sigma(m)))$ otherwise.
- For any speaker s of $\{\mathcal{L}, \sigma\}$, the **utility function of s**, notated u^s, is defined by $u^s(w) = u(m^s_w|w)$. If a speaker s is truthful we have, for any w, $u^s(w) = -\log(\mathsf{P}(\sigma(m^s_w)))$.
- The **expected utility** of a speaker s, notated $U(s)$, is defined as the expected value of u^s. For concreteness, we provide an example where Ω is the real interval $[0, 1]$, F is the set of all measurable subsets of $[0, 1]$ and P is the probability distribution associated with the uniform density function f over $[0, 1]$. Then we have $U(s) = \int_0^1 f(w)u^s(w)dw$. If s is truthful, $U(s) = -\int_0^1 f(w)\log(\mathsf{P}(\sigma(m^s_w)))dw$
- The **utility** of a language $\mathcal{L}$ is defined as $\sup_\sigma(U(\mathcal{L}, \sigma))$. It is guaranteed to exist because for every language $\mathcal{L}$ with n messages and every semantics σ for this language, $U(\mathcal{L}, \sigma) \leq \frac{n}{e}$

Proof. A rational speaker of $\{\mathcal{L}, \sigma\}$ is truthful, hence can only use a message m in a world w if m is true in w. When such a speaker uses a message m, the contribution of m to the overall expected utility is at most $-\mathsf{P}(\sigma(m))\log(\mathsf{P}(\sigma(m)))$. Since $x \mapsto -x\log(x)$ reaches its maximum $\frac{1}{e}$ at $x = \frac{1}{e}$, the contribution of a message m to the overall utility is at most $\frac{1}{e}$. If there are exactly n messages in $\mathcal{L}$, $U(\mathcal{L}, \sigma) \leq \frac{n}{e}$ for every σ. □

Consider a rational speaker s who is context sensitive i.e. observes the world then chooses among the true literals the S ones with maximal informativity. Then consider a receiver who starts with the same knowledge (probability distribution) as the speaker. If the goal of language is to exchange information, then an optimal lexicalization should allow a perfectly rational and cooperative speaker s to convey as much true information as possible – that is to say, the expected utility of speaker s should be maximal.

To streamline the notations, we write p_i for $\mathsf{P}(\sigma((a_i)^s_w))$, so there is an implicit dependence of the p_i's (as well as the expected utility) on the lexicalization σ. Likewise, for a message m, we write $\mathsf{P}(m)$ for $\mathsf{P}(\sigma(m^s_w))$.

3 One-spot case ($S = 1$)

We consider a language without negation and will start by examining the optimal lexicalization when no assumption is made about independence.

3.1 Non-independent messages

When there's only one message denoting an event with probability p, the expected utility is $-p\log(p)$ is maximized with $p = \frac{1}{e}$. When there are 2 messages with probabilities p_1 and p_2, the utility $-(p_1\log(p_1) + p_2\log(p_2))$ is maximized with $p_1 = p_2 = \frac{1}{e}$.

We now consider $N > 2$ messages. In the specific case where they are incompatible, it's not possible to assign probability $\frac{1}{e}$ to each message since $3 \times \frac{1}{e} > 1$.

Proposition In the best lexicalization for $N > 2$ mutually incompatible messages, each message has probably $\frac{1}{N}$ (therefore, they exhaust the logical space).

Proof. The function $x \mapsto -x\log(x)$ is increasing from 0 to $\frac{1}{e}$ (and reaches its maximum at $\frac{1}{e}$). If the N messages correspond to events that do not cover the whole logical space i.e. $\sum_{i=1}^{N} p_i < 1$, then at least one of the messages m_i corresponds to an event with probability lower than $\frac{1}{N}$ (because they are disjoint), hence lower than $\frac{1}{e}$ (because $\frac{1}{3} < \frac{1}{e}$). Consider an alternative lexicalization where all messages except m_i denote exactly the same events, but m_i has probability p'_i such that $p_i < p'_i < \frac{1}{e}$. This is possible since the messages do not cover the whole logical space. This alternative lexicalization achieves higher expected utility than the one we started with, because $-p'_i\log(p'_i) > -p_i\log(p_i)$. Hence any optimal lexicon will cover the whole logical space. We therefore need to maximize $-\sum_{i=1}^{N} p_i\log(p_i)$ with the constraint $\sum_{i=1}^{N} p_i = 1$, which we know is achieved by taking $p_1 = \ldots = p_N = \frac{1}{N}$. □

3.2 Necessity for non-independence in the optimal lexicalization

(Based on previous notes by Benjamin Spector and an argument from Keny Chatain)

Proposition For any lexicon, there exists a lexicon with disjoint atoms that achieves better or equal utility.

Proof. Let e_i be the events denoted by the messages m_i. Without loss of generality, we assume that e_i are ordered in decreasing order of probability (so m_1 is the least informative message). Let $f_i = e_i \wedge (\forall j > i, \neg e_j)$ We will show that if the same messages denoted the events f_i, they would form a language with at least as great utility.

Let prob-$e^*(w)$ be the probability of the most informative message for world w if m_i refers to e_i and prob-$f^*(w)$ if they refer to f_i. If no message can be used in w, we will assume prob-$e^*(w) =$ prob-$f^*(w) = 1$.

Given a world w, let $I = \{1 \leq i \leq N, e_i(w) \text{ is true}\}$. If I is empty, then none of the e_i are true in w. By definition, none of the f_i will be true either. Hence prob-$e^*(w) =$ prob-$e^*(w) = 1$.

If I is non-empty, let opt $= \max I$. Since the e_i are decreasing in probability, m_{opt} is necessarily the most informative message or one of the most informative messages if there are ties. Therefore, $\mathsf{P}(e_{\text{opt}}) =$ prob-$e^*(w)$. Trivially, by definition, f_i implies e_i and $\mathsf{P}(f_i) \leq \mathsf{P}(e_i)$ for all i. Furthermore, f_{opt} is true in w (because $\forall i > \text{opt}, e_i$ is false). So:

$$\text{prob-}f^*(w) \leq \mathsf{P}(f_{\text{opt}}) \leq \mathsf{P}(e_{\text{opt}}) = \text{prob-}e^*(w)$$

In both cases, we have prob-$f^*(w) \leq$ prob-$e^*(w)$. The proposition follows by taking expectations. □

Nevertheless, we shall examine the optimal lexicalization under the constraint of independence although we know that it is not overall optimal. Apart from providing an interesting mathematical exercise, this is motivated by the fact that, empirically, our common lexicon consists of messages that are more often than not independent. This is also true of primate communication, where the most common messages are warnings for different predators and are therefore independent.

3.3 Independent messages

(Based on previous notes by Emmanuel Chemla)

We start by computing by hand the expected utility for $N = 3$ in order to get an idea as to the generalization. Let $p_1 > p_2 > p_3$.
a_3, being the most informative atom, is used whenever it is true. a_2 is used when it is true and when the atoms more informative than itself – that is a_3 – are false. The same goes for a_1. So, the expected utility $U_{N,S}$ is given by:
$U_{3,1}(p_1, p_2, p_3) = -p_3 \log(p_3) - (1-p_3)p_2 \log(p_2) - (1-p_3)(1-p_2)p_1 \log(p_1)$
For whatever choices of p_2 and p_3, the maximum for the third term is obtained by taking $p_1 = \frac{1}{e}$.
Let $\hat{p}_1 = \frac{1}{e}$

$$\begin{aligned} U_{3,1}(\hat{p}_1, p_2, p_3) &= -p_3 \log(p_3) - (1-p_3)p_2 \log(p_2) + \frac{1}{e}(1-p_2)(1-p_3) \\ &= -p_3 \log(p_3) - (1-p_3)(p_2 \log(p_2) - \frac{1}{e}(1-p_2)) \end{aligned}$$

The function $p_2 \mapsto -p_2 \log(p_2) + \frac{1}{e}(1-p_2)$ reaches its maximum at $\hat{p}_2 = \exp(-(1+\hat{p}_1))$

$$\begin{aligned} U_{3,1}(\hat{p}_1, \hat{p}_2, p_3) &= -p_3 \log(p_3) - (1-p_3)(-\hat{p}_2(1+e^{-1}) - \frac{1}{e}(1-\hat{p}_2)) \\ &= -p_3 \log(p_3) + (1-p_3)(\hat{p}_2 + \hat{p}_1) \end{aligned}$$

The function $p_3 \mapsto -p_3 \log(p_3) + (1-p_3)(\hat{p}_2 + \hat{p}_1)$ reaches its maximum at $\hat{p}_3 = \exp(-(1+\hat{p}_1+\hat{p}_2))$
Thankfully $\hat{p}_1 > \hat{p}_2 > \hat{p}_3$.

Generalization

Let $1 > p_1 \geq \ldots \geq p_N > 0$. In this whole document, we will consider probabilities different from 0 and 1.

$$U_{N,1}(p_1, \ldots, p_N) = -\sum_{k=1}^{N} \prod_{i=k+1}^{N} (1 - p_i) p_k \log(p_k)$$

Similarly, at the k-th step we factorize the last term of $U_{N,1}(p_1, \ldots, p_n)$ by a function of p_k. The derivative of that function will be $-(1 + \log(p_k) + \sum_{i=1}^{k-1} p_i)$, so the optimal lexicalization is obtained by taking $\hat{p}_k = \exp(-(1 + \sum_{i=1}^{k-1} \hat{p}_i))$

Actually, the way in which we proceeded amounts to the same as setting all the partial derivatives over the p_i's equal to zero. It is also possible to get the same result without computing by hand and directly from the general formula of the expected utility for S spots in the case of independent messages:

$$U_{N,S}(p_1, \ldots, p_N) = -\sum_{i=1}^{N} \mathsf{P}(\#\{\text{true atoms among } a_{i+1}, \ldots, a_N\} < S) p_i \log(p_i)$$

For some $1 \leq k \leq N$, this can be written as:

$$\begin{aligned}
U_{N,S}(p_1, \ldots, p_N) = &- \sum_{i=k+1}^{N} \mathsf{P}(\#\{\text{TA among } a_{i+1}, \ldots, a_N\} < S) p_i \log(p_i) \\
&- \mathsf{P}(\#\{\text{TA among } a_{k+1}, \ldots, a_N\} < S) p_k \log(p_k) \\
&- p_k \sum_{i=1}^{k-1} \mathsf{P}(\#\{\text{TA among } a_{i+1}, \ldots, a_N \text{ but not } a_k\} < S - 1) p_i \log(p_i) \\
&- (1 - p_k) \sum_{i=1}^{k-1} \mathsf{P}(\#\{\text{TA among } a_{i+1}, \ldots, a_N \text{ but not } a_k\} < S) p_i \log(p_i)
\end{aligned}$$

Hence

$$\begin{aligned}
\frac{\partial U_{N,S}}{\partial p_k} = &-\mathsf{P}(\#\{\text{TA among } a_{k+1}, \ldots, a_N\} < S)(1 + \log(p_k)) \\
&- \sum_{i=1}^{k-1} \Big[\mathsf{P}(\#\{\text{TA among } a_{i+1}, \ldots, a_N \text{ but not } a_k\} < S - 1) \\
&- \mathsf{P}(\#\{\text{TA among } a_{i+1}, \ldots, a_N \text{ but not } a_k\} < S) \Big] p_i \log(p_i) \\
= &-\mathsf{P}(\#\{\text{TA among } a_{k+1}, \ldots, a_N\} < S)(1 + \log(p_k)) \\
&+ \sum_{i=1}^{k-1} \mathsf{P}(\#\{\text{TA among } a_{i+1}, \ldots, a_N \text{ but not } a_k\} = S - 1) p_i \log(p_i)
\end{aligned}$$

Applying the law of total probability, we have

$$\hat{p}_k = \exp\Big[-1 + \frac{1}{\sum_{\alpha=0}^{S-1} \mathsf{P}(\#\{\text{TA among } a_{k+1},\ldots,a_N\} = \alpha)} \times \sum_{\alpha=0}^{S-1} \mathsf{P}(\#\{\text{TA among } a_{k+1},\ldots,a_N\} = \alpha) \sum_{i=1}^{k-1} \mathsf{P}(\#\{\text{TA among } a_{i+1},\ldots,a_{k-1}\} = S-1-\alpha) p_i \log(p_i)\Big]$$

For $S = 1$, when all the partial derivatives are set to zero, we have

$$\hat{p}_k = \exp\Big[-1 + \sum_{i=1}^{k-1} \mathsf{P}(\text{none of } a_{i+1},\ldots,a_{k-1} \text{ is true}) \hat{p}_i \log(\hat{p}_i)\Big]$$

This allows us to show that $\hat{p}_k = \exp(-(1 + \sum_{i=1}^{k-1} \hat{p}_i))$ for every $1 \le k \le N$.

Proof. By induction:
Base case It has already been shown that $\hat{p}_1 = \frac{1}{e}$.

Induction hypothesis Suppose for some $1 \le k \le N-1$ that $\hat{p}_k = \exp(-(1 + \sum_{i=1}^{k-1} \hat{p}_i))$

Induction step

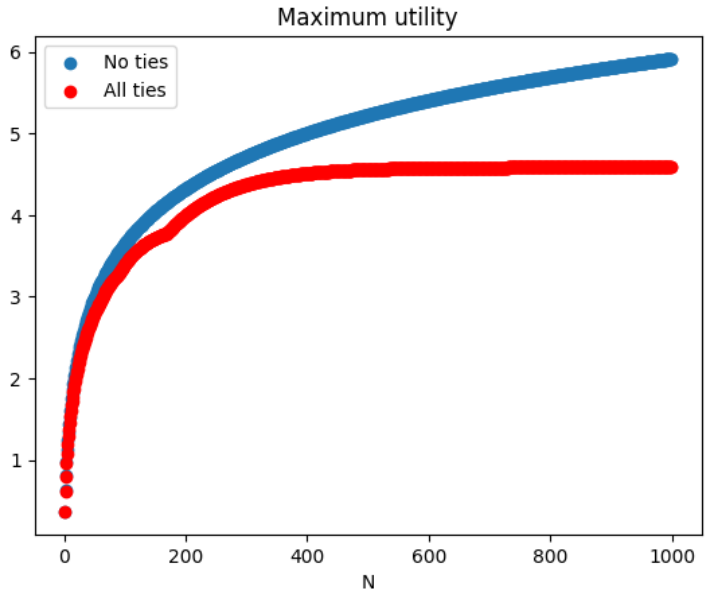
$$\begin{aligned}
\hat{p}_{k+1} &= \exp\Big[-1 + \sum_{i=1}^{k} \mathsf{P}(\text{none of } a_{i+1},\ldots,a_k \text{ is true}) \hat{p}_i \log(\hat{p}_i)\Big] \\
&= \exp\Big[-1 + \sum_{i=1}^{k-1} \mathsf{P}(\text{none of } a_{i+1},\ldots,a_k \text{ is true}) \hat{p}_i \log(\hat{p}_i) + \hat{p}_k \log(\hat{p}_k)\Big] \\
&= \exp\Big[-1 + (1-p_k) \sum_{i=1}^{k-1} \mathsf{P}(\text{none of } a_{i+1},\ldots,a_{k-1} \text{ is true}) \hat{p}_i \log(\hat{p}_i) + \hat{p}_k \log(\hat{p}_k)\Big] \\
&= \hat{p}_k \times \exp\Big[-\hat{p}_k\Big(\sum_{i=1}^{k-1} \mathsf{P}(\text{none of } a_{i+1},\ldots,a_{k-1} \text{ is true}) \hat{p}_i \log(\hat{p}_i) - \log(\hat{p}_k)\Big)\Big] \\
&= \exp(-(1 + \sum_{i=1}^{k-1} \hat{p}_i)) \times \exp(-\hat{p}_k) \\
\hat{p}_{k+1} &= \exp(-(1 + \sum_{i=1}^{k} \hat{p}_i))
\end{aligned}$$

□

This proves that the best lexicalization is incrementally optimized, as the optimal probability for a_k does not depend on the probabilities of the atoms more informative than itself. We can compare this to the case where the p_i's are constrained i.e. only events having a certain fixed probability p are lexicalized. If $p_1 = \ldots = p_N = p$ then

$$U_{N,1}(p,\ldots,p) = -p\log(p) \sum_{k=1}^{N} (1-p)^{N-k} = -\log(p)(1-(1-p)^N)$$

In either case, there's no closed-form expression of the maximum utility $U_{N,1}(\hat{p}_1, \ldots, \hat{p}_N)$. Consequently, the computations were performed using Python.:



Two conjectures can be made from the graph:

- If some but not all of the p_i's are tied, the scatterplot would lie between those for 'no ties' and 'all ties' (but we haven't tried to prove this).
- In the 'no ties' case, the maximum utility as a function of N is equivalent to $\log(N)$.

Proof. When $S = 1$ and $p_1 \geq \ldots \geq p_{N+1}$, we have the following induction formula:

$$U_{N+1,1}(p_1, \ldots, p_{N+1}) = -p_{N+1}\log(p_{N+1}) + (1 - p_{N+1})U_{N,1}(p_1, \ldots, p_N)$$

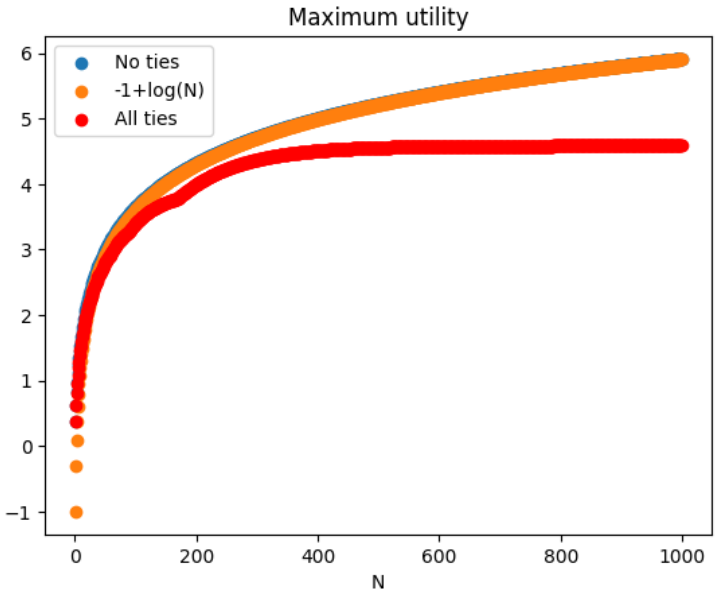
$$\frac{\partial U_{N+1,1}}{\partial p_{N+1}} = 0 \iff -1 - \log(p_{N+1}) - U_{N,1} = 0$$

Thus the maximum of $U_{N+1,1}$ is reached when $\hat{p}_{N+1} = \exp(-U_N - 1)$. (To alleviate notations, the argument is still $(\hat{p}_1, \ldots, \hat{p}_N)$ for $U_{N,1}$ and $(\hat{p}_1, \ldots, \hat{p}_{N+1})$ for $U_{N+1,1}$ but we won't write it every time.) Injecting that in the previous expression for $U_{N+1,1}$ leads to

$$\begin{aligned} U_{N+1,1} &= -\hat{p}_{N+1}(-U_{N,1} - 1) + (1 - \hat{p}_{N+1})U_{N,1} \\ &= e^{-U_{N,1}-1} + U_{N,1} \\ &= \hat{p}_{N+1} + U_{N,1} \end{aligned}$$

Therefore $U_{N,1}(\hat{p}_1, \ldots, \hat{p}_N) = \sum_{k=1}^{N} \hat{p}_k$. If we define $V_{N,1} = U_{N,1} + 1$ then $V_{N+1,1} = e^{-V_{N,1}} + V_{N,1}$ and $V_{N,1}$ is roughly like $\log(N)$. We can also assume that p_N is roughly like $\frac{1}{N}$. □

Graphically, it seems that $U_{N,1}(\hat{p}_1, \ldots, \hat{p}_N)$ is indeed roughly like $-1 + \log(N)$. The orange and blue curves are almost superimposed.



3.4 Is negation ever useful ?

(Based on previous notes by Benjamin Spector)

If we have N messages and negation, we can communicate $2N$ messages in total. Suppose all the messages have different probabilities. When we order them by informativity, we find that among the $(N+1)$ most informative messages, there is a pair of the form $\{m, \neg m\}$ (by the pigeonhole principle). As the speaker is sure to find one true message among the first $(N+1)$ ones, they will never use any message less informative than those $(N+1)$ ones. With no other constraints, such as disjointness or independence, adding negation is like adding only one message, but with a constraint on its probability which is absent if we are free to lexicalize the additional message as we want. Therefore, the benefit of adding negation is – when all messages correspond to events with different probabilities (no ties) – always less than what one gets from adding adding a new lexical message.

The previous argument does not generalize to the case where any ties are allowed but the theorem below still holds when ties are taken into account. The general proof has not been studied in the context of this internship.

Theorem For $S = 1$, the utility of a language with $N+1$ propositional atoms is at least as high as that of a language with N propositional atoms and negation.

4 Multi-spots case ($S > 1$)

We consider the case of independent messages with $p_1 \geq \ldots \geq p_N$.
Let us recall the formula

$$U_{N,S}(p_1,\ldots,p_N) = -\sum_{i=1}^{N} \mathsf{P}(\#\{\text{true atoms among } a_{i+1},\ldots,a_N\} < S) p_i \log(p_i)$$

Each term of the sum arises from the question "when is the atom a_i being used ?". An alternative expression is to write each term considering the number of simultaneously true atoms rather than each atom separately (I denotes the set of indices for which the a_i are true):

$$U_{N,S}(p_1,\ldots,p_N) = -\sum_{\substack{I\subset[\![1,N]\!]\\ 1\leq|I|\leq S}} \prod_{i\in I} p_i \prod_{j\notin I}(1-p_j) \sum_{i\in I}\log(p_i) - \sum_{\substack{I\subset[\![1,N]\!]\\ S+1\leq|I|\leq N}} \prod_{i\in I} p_i \prod_{j\notin I}(1-p_j) \sum_{\text{S smallest among } p_{i,i\in I}} \log(p_i)$$

Adding negation:

$$U_{N\ni neg,S}(p_1,\ldots,p_N) =$$
$$-\sum_{\substack{I\subset[\![1,N]\!]\\ 1\leq|I|\leq S}} \prod_{i\in I} p_i \prod_{j\notin I}(1-p_j)\Big(\sum_{i\in I}\log(p_i)+\sum_{j\notin I}\log(1-p_j)\Big) - \sum_{\substack{I\subset[\![1,N]\!]\\ S+1\leq|I|\leq N}} \prod_{i\in I} p_i \prod_{j\notin I}(1-p_j) \sum_{\text{S smallest among } p_{i,i\in I}} \log(p_i)$$

With the expression in this form, it is straightforward to see that for fixed values of the p_i's, adding negation is always beneficial:

$$U_{N,S}(p_1,\ldots,p_N) - U_{N\ni neg,S}(p_1,\ldots,p_N) = \sum_{\substack{I\subset[\![1,N]\!]\\ 1\leq|I|\leq S}} \prod_{i\in I} p_i \prod_{j\notin I}(1-p_j) \sum_{j\notin I}\log(1-p_j) < 0$$

In order to conjecture a general proof, we start with computations by hand of the cases $S = N-1$ or $S = N-2$ for $N = 3$ and $N = 4$ without negation.

3-message case

Let $N = 3$, $S = 2$ and $p_1 \geq p_2 \geq p_3$.
$U_{3,2}(p_1,p_2,p_3) = -p_1(1-p_2p_3)\log(p_1) - p_2\log(p_2) - p_3\log(p_3)$
$\dfrac{\partial U_{3,2}}{\partial p_1} = p_2p_3(1+\log(p_1))$
$\dfrac{\partial U_{3,2}}{\partial p_2} = -(1+\log(p_2)) + p_1p_3\log(p_1)$
$\dfrac{\partial U_{3,2}}{\partial p_3} = -(1+\log(p_3)) + p_1p_2\log(p_1)$
$U_{3,2}$ is maximized with $\begin{cases}\hat{p}_1 = 1/e\\ \hat{p}_2 = \exp(-(1+\frac{1}{e}p_3))\\ \hat{p}_3 = \exp(-(1+\frac{1}{e}p_2))\end{cases}$
Let $f : x \mapsto \exp(-(1+\frac{x}{e}))$ which is a decreasing function. $\hat{p}_2$ and $\hat{p}_3$ are both solutions to $x = f\circ f(x)$. As $f\circ f$ is strictly monotone over $[0,1]$ with $f\circ f(0) > 0$ and $f\circ f(1) < 1$, it has only exactly one fixed point in $[0,1]$. Therefore $\hat{p}_1 > \hat{p}_2 = \hat{p}_3$.

That prompts us to ask if each of the smallest S $\hat{p}_i$'s can generally be expressed in terms of itself and whether those S ones are always equal.

4-message case

Let $N = 4$, $S = 3$ and $p_1 \geq p_2 \geq p_3 \geq p_4$
$U_{4,3}(p_1, p_2, p_3, p_4) = -p_1(1 - p_2p_3p_4)\log(p_1) - p_2\log(p_2) - p_3\log(p_3) - p_4\log(p_4)$
$\frac{\partial U_{4,3}}{\partial p_1} = -(1 - p_2p_3p_4)(1 + \log(p_1)) = 0 \iff p_1 = 1/e$
$\frac{\partial U_{4,3}}{\partial p_2} = -(1 + \log(p_2)) + p_1p_3p_4\log(p_1) = -(1 + \log(p_2)) - \frac{p_3p_4}{e}$
Likewise $\frac{\partial U_{4,3}}{\partial p_3} = -(1 + \log(p_3)) - \frac{p_2p_4}{e}$ and $\frac{\partial U_{4,3}}{\partial p_4} = -(1 + \log(p_4)) - \frac{p_2p_3}{e}$
$U_{4,3}$ is maximized with $\begin{cases} \hat{p}_1 = 1/e \\ \hat{p}_2 = \exp(-(1 + \frac{1}{e}p_3p_4)) \\ \hat{p}_3 = \exp(-(1 + \frac{1}{e}p_2p_4)) \\ \hat{p}_4 = \exp(-(1 + \frac{1}{e}p_2p_3)) \end{cases}$
It seems that $\hat{p}_2, \hat{p}_3$ and $\hat{p}_4$ cannot be shown to be equal from something as simple as a cyclic system of equations as we had surmised from the $N = 3$, $S = 2$ case. However we can show pairwise equalities:
Considering p_4 a constant, $\hat{p}_2$ and $\hat{p}_3$ are both solutions to $x = f \circ f(x)$ with $f : x \mapsto \exp(-(1 + \frac{p_4x}{e}))$. $(f \circ f)'(x) =$ (a positive constant) $\times$ exp(a bunch of things) which is sufficient to say that $f \circ f$ is strictly monotone. As in the above, we have $f \circ f(0) > 0$ and $f \circ f(1) < 1$. So $\hat{p}_2(p_4) = \hat{p}_3(p_4)$. Similarly, we get $\hat{p}_2(p_3) = \hat{p}_4(p_3)$ and $\hat{p}_3(p_2) = \hat{p}_4(p_2)$, which gives $\hat{p}_2 = \hat{p}_3 = \hat{p}_4$.

Now take $N = 4$, $S = 2$ and $p_1 \geq p_2 \geq p_3 \geq p_4$
$U_{4,2}(p_1, p_2, p_3, p_4) = -p_1(p_2(1 - p_3)(1 - p_4) + (1 - p_2)p_3(1 - p_4) + (1 - p_2)(1 - p_3)p_4 + (1 - p_2)(1 - p_3)(1 - p_4))\log(p_1) - p_2(1 - p_3p_4)\log(p_2) - p_3\log(p_3) - p_4\log(p_4)$
$\frac{\partial U_{4,2}}{\partial p_1} = -(1 + \log(p_1)) \times \text{whatever} = 0 \iff p_1 = 1/e$
$\frac{\partial U_{4,2}}{\partial p_2} = -(1 - p_3p_4)(1 + \log(p_2)) + p_1\log(p_1)(p_3(1 - p_4) + p_4(1 - p_3)) = -(1 - p_3p_4)(1 + \log(p_2)) - \frac{p_3(1-p_4)+p_4(1-p_3)}{e}$
$\frac{\partial U_{4,2}}{\partial p_3} = p_1\log(p_1)(p_2(1 - p_4) + p_4(1 - p_2)) + p_2p_4\log(p_2) - (1 + \log(p_3))$
$\frac{\partial U_{4,2}}{\partial p_4} = p_1\log(p_1)(p_2(1 - p_3) + p_3(1 - p_2)) + p_2p_3\log(p_2) - (1 + \log(p_4))$
$U_{4,2}$ is maximized with $\begin{cases} \hat{p}_1 = 1/e \\ \hat{p}_2 = ?? \text{ (we need to compute } \hat{p}_3 \text{ and } \hat{p}_4 \text{ first)} \\ \hat{p}_3 = \exp(-1 + p_2p_4\log(p_2) - \frac{p_2(1-p_4)+p_4(1-p_2)}{e}) \\ \hat{p}_4 = \exp(-1 + p_2p_3\log(p_2) - \frac{p_2(1-p_3)+p_3(1-p_2)}{e}) \end{cases}$
The same kind of argument as in the previous cases shows that $\hat{p}_3 = \hat{p}_4$. Simplifying the partial derivative with respect to p_2:

$$\frac{\partial U_{4,2}}{\partial p_2} = -(1-p_3^2)(1+\log(p_2)) - \frac{2p_3(1-p_3)}{e} = 0$$

Therefore $\hat{p}_2 = \exp(-(1+\frac{2\hat{p}_3}{e(1+\hat{p}_3)})) < \hat{p}_1$. Finally, a graphical resolution ensures that $\hat{p}_3 < \hat{p}_2$.

Generalization for $S = N-1$

$$U_{N,S}(p_1,\ldots,p_N) = -p_1(1-\prod_{i=2}^{N} p_i)\log(p_1) - \sum_{i=2}^{N} p_i \log(p_i)$$

$$\frac{\partial U_{N,S}}{\partial p_1} = -(1-\prod_{i=2}^{N} p_i)(1+\log(p_1)) = 0 \iff \hat{p}_1 = 1/e$$

$$\frac{\partial U_{N,S}}{\partial p_{j\neq 1}} = \prod_{\substack{i=1\\ i\neq j}}^{N} p_i \log(p_1) - (1+\log(p_j)) = 0 \iff \hat{p}_j = \exp(-(1+\frac{1}{e}\prod_{\substack{i=2\\ i\neq j}}^{N} p_i))$$

Let $j,k \neq 1$ and $f : x \mapsto \exp(-(1+\frac{1}{e}x))$. Suppose $p_j < p_k$. Then $f(p_1 \times \ldots \times \cancel{p_k} \times \ldots \times p_N) < f(p_1 \times \ldots \times \cancel{p_j} \times \ldots \times p_N)$ (because f is a decreasing function) i.e. $p_k < p_j$ because the $p_{i\neq 1}$'s are such that $p_i = f(p_1 \times \ldots \times \cancel{p_i} \times \ldots \times p_N)$. Contradiction. Therefore all the $\hat{p}_{i\neq 1}$ are equal. Let $\hat{p}_N$ denote their value.

$\hat{p}_N$ is such that $\hat{p}_N = \exp(-(1+\frac{1}{e}\hat{p}_N^{N-1}))$. We wish to show that $\hat{p}_N \xrightarrow[N\to+\infty]{<} \frac{1}{e}$

Let $m = N-1$. $\hat{p}_N$ is solution to $x = \exp(-(1+\frac{x^m}{e}))$. We note that:

$$\begin{aligned}
x = \exp(-(1+\frac{x^m}{e})) &\iff (xe)^{-m} = \exp(\frac{mx^m}{e}) \\
&\iff me^{-m-1} = \frac{mx^m}{e}\exp(\frac{mx^m}{e}) \\
&\iff \frac{mx^m}{e} = W_0(me^{-m-1}) \\
&\iff x = \sqrt[m]{\frac{e}{m}W_0(me^{-m-1})}
\end{aligned}$$

where W_0 denotes the principal branch of the Lambert W function.

W is the converse function of $w \mapsto we^w$ and $\forall z \in \mathbb{C} \quad we^w = z \iff \exists k \in \mathbb{Z}, w = W_k(z)$ where the W_k's denote the branches of the Lambert W function. Here we will only be dealing with the principal branch W_0 because the argument (me^{-m-1}) is a positive real number and $k = 0$ is the only branch satisfying $W_k(x)e^{W_k(x)} = x$ for $x > 0$.

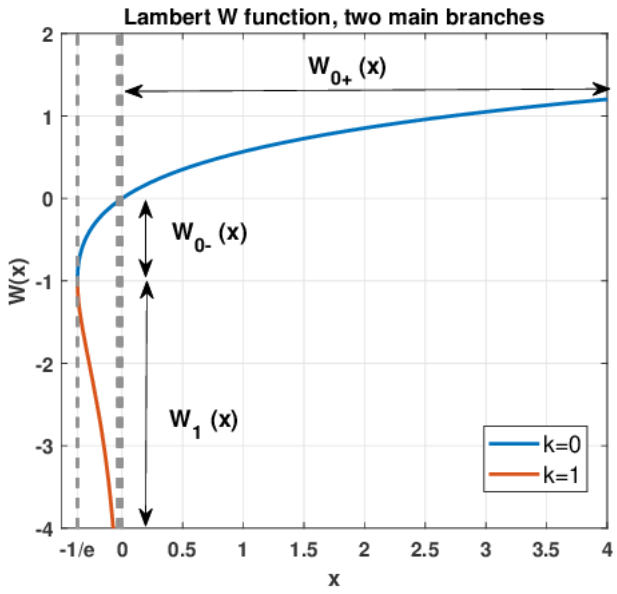


Figure 1: Two main branches of the Lambert W function. Reproduced from Nguyen et al. 2022.

There is no convenient closed-form expression for any of the branches but we need to find a lower bound for $W_0(me^{-m-1})$, at least for great values of m. We shall be accepting the following theorem:

Lagrange inversion theorem

If z is defined as a function of w by the relation $f(w) = z$ where f is analytic at point a and $f'(a) \neq 0$, then $w = g(z)$ with g being defined by a power series $g(z) = a + \sum_{n=1}^{+\infty} g_n \frac{(z - f(a))^n}{n!}$ where $g_n = \lim_{w \to a} \frac{\mathrm{d}^{n-1}}{\mathrm{d}w^{n-1}} [(\frac{w - a}{f(w) - f(a)})^n]$

Applying this to $f : w \mapsto we^w$, the Taylor series of W_0 around $a = 0$ is given by

$$\begin{aligned} W_0(z) &= \sum_{n=1}^{+\infty} \frac{z^n}{n!} (\lim_{w \to 0} \frac{\mathrm{d}^{n-1} e^{-nw}}{\mathrm{d}w^{n-1}}) \\ &= \sum_{n=1}^{+\infty} \frac{z^n}{n!} (\lim_{w \to 0} (-n)^{n-1} e^{-nw}) \\ &= \sum_{n=1}^{+\infty} (-n)^{n-1} \frac{z^n}{n!} \end{aligned}$$

In order to find a lower bound, it is sufficient to stop at the 3rd order: $W_0(z) = z - z^2 + O(z^3)$.
In particular $\forall x > 0 \quad W_0(x) \geq x - x^2$

Therefore

$$\begin{aligned}
\hat{p}_N &= \sqrt[m]{\frac{e}{m} W_0(me^{-m-1})} \\
&\geq \sqrt[m]{\frac{e}{m}(me^{-m-1} - (me^{-m-1})^2)} \\
&= \sqrt[m]{e^{-m}(1 - \frac{m}{e^{m+1}})} \\
&= \frac{1}{e} \sqrt[N-1]{1 - \frac{N-1}{e^N}} \\
&\xrightarrow[N \to +\infty]{<} \frac{1}{e}
\end{aligned}$$

Interpretation of the result: when the number of spots is such that only one message may not be uttered, the optimal $\hat{p}_i$'s all converge to $\frac{1}{e}$ as the size of the lexicon increases. It seems intuitively consistent that the case where "almost anything can be uttered" converges to the limiting case where "anything can be uttered" i.e. $S \geq N$ (then we have $\forall i \geq 1 \quad \hat{p}_i = \frac{1}{e}$).

5 Conclusion

In this report, the case $S = 1$ has been thoroughly addressed, and we have shown that the best lexicalization for independent messages requires an incremental optimization of the probabilities and offers a surprisingly simple expression for the utility of the corresponding language. Due to time constraints, we have not addressed the general case of non-independent messages when $S > 1$. Even for independent messages, we cannot yet propose a generalization of the above result for $S = N - k$ when k is other than 1, but we may conjecture that the result $\hat{p}_N \xrightarrow[N \to +\infty]{<} \frac{1}{e}$ can be extended to $S = N - k$ for any integer k. A possible line of reasoning for future continuation of this study could make use of the induction formula $U_{N+1,S} = (1 - p_{N+1})U_{N,S} + p_{N+1}(-\log(p_{N+1}) + U_{N,S-1})$. It would also be interesting to investigate whether negation becomes useful when $S > 1$ and, if so, at what stage.

References

Bergen, Leon, Roger Levy, and Noah Goodman (May 10, 2016). "Pragmatic reasoning through semantic inference". In: Semantics and Pragmatics 9. DOI: 10.3765/sp.9.20.

Enguehard, Émile and Benjamin Spector (June 4, 2021). "Explaining gaps in the logical lexicon of natural languages: A decision-theoretic perspective on the square of Aristotle". In: Semantics and Pragmatics 14, 5:1–31. DOI: 10.3765/sp.14.5.

Goodman, Noah D. and Andreas Stuhlmüller (2013). "Knowledge and Implicature: Modeling Language Understanding as Social Cognition". In: Topics in Cognitive Science 5.1, pp. 173–184. DOI: 10.1111/tops.12007.

Nguyen, Hieu et al. (2022). "Solar PV modeling with Lambert W function: An exponential cone programming approach". In: IEEE Kansas Power and Energy Conference (KPEC).