The interactions of rational pragmatic agents provide a framework for understanding efficient language structure and use

Author 1
bpeloqui@stanford.edu
Department of Psychology
Stanford University

Author 2
author1@university.edu
Department of Psychology
Some University

Abstract

We discuss a framework for studying the distributional properties of linguistic systems as emerging from in-the-moment interactions of speakers and listeners. Our work takes Zipfian notions of lexicon-level efficiency as a starting point, connecting these ideas to Gricean notions of conversational-level efficiency. To do so, we begin by deriving an objective function for measuring the communicative efficiency of linguistic systems and then examining the behavior of this objective in a series of simulations focusing on the communicative function of ambiguity in language.

Keywords: Communicative efficiency, Rational Speech Act theory, computational modeling, information theory, agent-based simulation

Introduction

Why do natural languages look they way they do? While Zipf (1935) presented his "Principle of Least Effort" as a domain general framework for understanding human behavior, he took language as his central case-study. He proposed that distributional properties found in natural language were evidence of speaker-listener effort minimization. In his own words, "we are arguing that people do in fact act with a maximum economy of effort, and that therefore in the process of speaking-listening they will automatically minimize the expenditure of effort." Evidence for this claim was largely derived at the level of the lexicon. Zipf argued that the particular relationship between a word's frequency and its rank, length, and denotation size could be explained as emergent phenomena from speaker-listener effort minimization.

Zipf articulated what is now considered a *functionalist* approach to language science – analyzing language structure and use in terms of efficiency. Such an approach might reframe our opening question as follows – how does having property x make using language ℓ more or less useful for communication? This reframing provides an opportunity to study both language *structure* (as Zipf primarily did) or *use* in terms of efficiency. For example, Regier et al. (201X) showed that languages appear organize semantic domains (form-meaning mappings) in maximally efficient ways. Likewise, Piantadosi et al. (2012) argued that lexical ambiguity is a useful property of any communication system when communication is contextualized. We will return to Piantadosi's argument later in the paper.

As the opening question indicates, language has both *structure* and *users*. Either of these dimensions are available to analysis. For example, both Regier (CITATION) and Piantadosi, et al (2012) assumed that language users were approximately Bayesian. In particular, Piantadosi needed to assume that language users could *use* context during conversation to derive their particular argument.

// Maybe don't include par below... Other work has focused more directly on efficient use of linguistic properties. For example, Levy & Jaeger (2007) showed speakers will choose to include an optional language items when oncoming material is high-surprisal. They argued this behavior indicates an efficient strategy under an information-theoretic analysis, following from Shannon's Source Coding Theorem. Such efficient use of redundancy has also been described at the morphological level (CITATION) and across multiple languages (CITATION). Similar theories have described efficient production behavior at the level of discourse.

There is little coincidence that Zipf introduced his Principle of Least Effort with an analysis of language. Language-structure and -use provide a particularly fascinating case-study under an efficiency-based analysis as speaker- and listener-effort is asymmetric. Simply put, in the act of communicating with another person, what is effortful as a speaker is likely different from what is effortful as a listener. For example, Zipf noted that purely from the standpoint of speaker effort, what Zipf called "Speaker's Economy," an optimal language $\ell^*_{speaker}$ would tend toward a vocabulary of just a single, low-cost word. Given such a language, the full set of potential meanings would be conveyed using only that word, i.e. $\ell^*_{speaker}$ would be fully ambiguous and all possible meanings would need to be disambiguated by a listener. Conversely, from the standpoint of listener effort, what Zipf called "Auditor's Economy," an optimal language $\ell^*_{listener}$ would bijectively map all possible meanings to distinct words, eliminating a listener's need to disambiguate. Under this analysis, speaker effort is related to production cost and listener effort to understanding (disambiguation) cost. Clearly natural languages fall between the two extremes of $\ell^*_{speaker}$ and $\ell^*_{listener}$. Zipf proposed that the particular lexicon-level properties he observed emerged from the competition of these forces – the pressure to jointly minimize speaker and listener effort.

Implicit in Zipf's (1935) argument is that aggregate, distributional-level effects emerge from optimization during local interactions between speaker-listeners. However, Zipf did not formalize "local interaction" beyond the example of a optimal speaker and listener languages. In that example, he described a reference game setting; similar to those described by Wittgenstein (CITATION). In this scenario, speakers and listeners are aware of a set of objects M, which will refer to as *meanings* and are knowledgeable about the set of possible

signals U (*utterances*) that could be used by a speaker to refer to a given meaning. Utterances may differ in their cost, the number of meanings to which they refer, and the objects may also vary in the degree to which they need to be described in a given context. Subsequent functionalist projects examining efficient language structure (Piantadosi, Regier) have also adopted the reference game setting as the primary unit of analysis.

A half century after Zipf, the linguist Lawrence Horn (1984), highlighted the importance of Zipf's principles for explaining conversation-level phenomena. Horn suggested a direct link between Zipf's Speaker and Listener economy and, what was at the time, more recent work on conversational pragmatics by Grice (1975). Horn highlighted that the interaction of Zipf's forces were "largely responsible for generating Grice's conversational maxims and the schema for pragmatic inference derived therefrom." Put differently, system-level efficiency we see in languages is deeply related to local-level efficiency during the in-the-moment interactions of speakers and listeners.

In this work we present a step toward formalizing this connection highlighted by Horn – formalizing the theoretical connection between Zipf and Grice, and in doing so, deriving aggregate, distributional level properties from local, in-the-moment interactions of speaker-listeners. To gain traction on this project we adopt a simulation-based approach, modeling communication as reference games in which we can vary the types of agents and languages present. This framework allows us to investigate a range of functionalist theories broadly under the categories of efficient language design and language use.

As a proof of concept, we focus on an analysis of the communicative function of ambiguity in language following work by Piantadosi et al. (2012). We adopt this particular linguistic phenomena to demonstrate the extent to which we can explore both the design- and use- based functionalist ideas. We focus on the question of efficient language structure and efficient language use. In the former, we ask "when is it desirable to have a language with lexical ambiguity?" In the latter we ask "under what circumstances do speakers and listeners use ambiguity efficiently?"

We begin with a high-level introduction to the modelling framework introducing the basic ingredients we will need to represent language as repeated reference games. Following this introduction we derive a simple objective function for measuring the efficiency of linguistic systems in this setting. Subsequently, we move on to two case-studies examining the questions posed above, framing results in terms of our efficiency measure.

Exploring efficient language- design and use in rational pragmatic agents

Simulation set-up.

Reference games We take as our unit of analysis the basic referential communication game similar to those described by WittgensteIn (CITATION). Both speakers and listeners are aware of a set of objects M, which will refer to as *meanings* and are knowledgeable about the set of possible signals U (*utterances*) that could be used to refer to a given meaning. Utterances may have different relative costs, operationalized via a prior over utterances $p(u)$. Similarly, meanings differ in the degree to which they need to be talked about. This is operationalized as a prior over meanings $p(m)$. Note that the prior over meanings are analogous to the *need probabilities* assumed in previous work (Regier, CITATION). We consider a set of contexts C which describe different need probability distributions over our set of meanings $p(m|c)$.

Languages A language ℓ defines the set of semantic mappings between utterance and meanings. For example, in a world with three utterances $U = \{u_1, u_2, u_3\}$and three meanings $M = \{m_1, m_2, m_3\}$ the boolean matrix

	m_1	m_2	m_3
u_1	1	1	0
u_2	0	1	0
u_3	0	0	1

describes the literal semantics of ℓ. E.g. the language describes semantic mappings $[[u_1]] = \{m_1, m_2\}, [[u_2]] = \{m_2\}, [[u_3]] = \{m_3\}$.

Speakers and listeners A *speaker agent* defines a conditional distribution over utterances, mapping from intended meanings M to utterances U using a particular semantic mapping given by ℓ. That is, a speaker defines $P_{speaker}(u|m;\ell)$. We will use $S(u|m;\ell)$ as short-hand throughout. A *listener agent* defines a conditional distribution over meanings, mapping from utterances U to meanings M using a particular semantic mapping given by ℓ. We will use $L(m|u;\ell)$ as shorthand. Note that both speakers and listeners can induce joint distributions over the set of all signaling events E, although, importantly, these distributions may differ:

$$P_{speaker}(u,m;\ell) = S(u|m;\ell)p(m)$$

$$P_{listener}(u,m;\ell) = L(m|u;\ell)p(u)$$

In general, we would like to consider the efficiency of a system ℓ in terms of these joint distributions.

Zipfian objective for linguistic system efficiency

Metrics for optimal linguistic systems in the reference game setting

Given a linguistic system ℓ used by a speaker S and listener L we need some measure of efficiency. For our present purposes, we'd like to tie this objective to fundamental principles of speaker and listener effort suggested by Zipf. To this end, we define an objective function, which as a function of a particular linguistic system ℓ, speaker S and listener L returns some measure of efficiency (i.e. $f(S,L,\ell) \to \mathbb{R}$). Additionally, we will take the fundamental unit of analysis as the reference game event.

Zipf proposed that the particular distributional properties found in natural language emerge as a result of competing speaker and listener pressures. We operationalize this in equation (1) – the efficiency of a linguistic system ℓ being used by speaker and listener agents S and L is sum of the expected speaker and listener effort to communicate over all possible communicative events.

$$\text{Efficiency}(S,L,\ell) = \mathbb{E}_{e\in E}[\text{speaker effort}] + \mathbb{E}_{e\in E}[\text{listener effort}] \tag{1}$$

Let speaker effort be the log probability of a particular utterance. $log_2(p(u))$ E.g. the number of bits needed to encode the utterance u.

$$\text{speaker effort} = -log_2(p(u))$$

Let listener effort be the log probability a listener disambiguates an intended meaning m given an utterance u E.g. the number of guesses a listener would need to discover the intended meaning m given an utterance u.

$$\text{listener effort} = -log_2(L(m|u;\ell))$$

Rewriting (1) we now have

$$\begin{aligned}\text{Efficiency}(S,L,\ell) = \mathbb{E}_{e\in E}[-log_2(p(u))] + \\ \mathbb{E}_{e\in E}[-log_2(L(m|u;\ell))]\end{aligned} \tag{2}$$

In general, these expectations are each taken over all possible communicative events $e \in E$ weighted by the probability of a particular event $p(e)$. Recall that is the set of all utterance, meaning pairs $< u,m >= e \in E$.

$$= \sum_{e\in E} p(e)[-log_2(p(u))] + \sum_{e\in E} p(e)[-log_2(L(m|u;\ell))] \tag{3}$$

We assume that the particular joint distribution over u,m pairs follows from a simple generative model. First, some meaning is sampled with probability $p(m)$. Our speaker attempts to convey this intended meaning to a speaker by encoding it in the utterance u via the conditional $S(u|m;\ell)$. Combining these terms leads to the *speaker's joint distribution over events* which we can write as $P_{Speaker}(u,m;\ell) = S(u|m;\ell)p(m)$.

$$\begin{aligned}= \sum_{u,m} p_{speaker}(u,m;\ell)[-log_2(p(u))] + \\ \sum_{u,m} p_{speaker}(u,m;\ell)[-log_2(L(m|u;\ell))]\end{aligned} \tag{4}$$

Simplifying we arrive at (5):

$$= \sum_{u,m} P_{speaker}(u,m;\ell)[-log_2(L(m|u;\ell)p(u))] \tag{5}$$

Note that $L(m|u;\ell)p(u)$ is the listener-based joint distribution over all communicative events ($P_{listener}(u,m;\ell)$).

$$= \sum_{u,m} P_{speaker}(u,m;\ell)[-log_2(P_{listener}(u,m;\ell))] \tag{6}$$

This is the simply the cross-entropy between the speaker an listener joint distributions.

$$\begin{aligned}&= \mathbb{E}_{P_{speaker}}[-log_2(P_{listener})] \\ &= H_{cross}(P_{speaker}, P_{listener})\end{aligned} \tag{7}$$

From a mathematical standpoint this objective is intuitive. An optimal linguistic system being used by a speaker and listener agent will induce a particular set of form-meaning mappings that are both *close* between speaker and listener agents and *peaky* in that they assign non-uniform probability mass to the valid form-meaning mappings.

Two case-studies on efficient language use and design

We now have a metric for exploring questions of language optimality in the reference game setting (the speaker-listener cross-entropy). In the following sections we examine the behavior of the objective with respect to the communicative function of ambiguity. In particular, we examine three questions. The first is a question of optimal design – given a set of possible languages L in what circumstances does the optimal languages ℓ^* contain ambiguity? The second is a question of optimal *use* – give a particular language that theoretically allows for the use of ambiguous material under what circumstances will ambiguous forms be used efficiently?

Experiment 1: Rational-pragmatic speakers prefer languages with ambiguous items

Piantadosi et al. (2012) proposed that a language with ambiguous items is strictly more efficient than a language with out, *when ambiguous items can be disambiguated with context*. Formally, they authors highlighted the fact that if we measure the efficiency of a linguistic system in terms of entropy, we have the basic result from (information theory citation) that the conditional entropy of an ensemble is strictly less that the unconditional entropy, when the information we condition on is informative ($H(X|C) < H(X)$).

Piantadosi et al. (2012) provided empirical evidence for this claim, showing (SAY WHAT PIANTADOSI DID). We consider their basic argument in the reference game setting examining the full-space of valid (footnote here for "valid" we have the constraint that all meanings must be expressible) languages (semantic mappings) between forms and utterances in which $|U| = |M| = 4$. A language $\ell \in L$ is "valid" so long as each possible meaning in $m \in M$ can be referred to by at least one form $u \in U$ (every column of ℓ has some non-zero assignment). For example, a one word language

that is fully ambiguous would map a single form to all possible meanings. We represent that via the boolean matrix ℓ_1 in which $[[u_1]] = \{m_1, m_2, m_3\}, [[u_2]] = [[u_3]] = \{\}$

	m_1	m_2	m_3
u_1	1	1	1
u_2	0	0	0
u_3	0	0	0

Note that $\mathbb{E}$[speaker effort] can be arbitrarily low for such a language, as it is directly proportional to $p(u_1)$. That is in the single utterance language, $\mathbb{E}$[speaker effort] $\to 0$ as $p(u_1) \to 1$. However, $\mathbb{E}$[listener effort] is likely to be high as the only available form (u_1) is full ambiguous between possible meanings.

By contrast consider a language ℓ_2:

	m_1	m_2	m_3
u_1	1	0	0
u_2	0	1	0
u_3	0	0	1

In this case $\mathbb{E}$[listener effort] $= 0$ – each form has a one-to-one mapping to a given meaning. However speaker effort will likely be greater than in ℓ_1 so long as $p(u_1) < p(u_2)$ and $p(u_1) < p(u_3)$.

Finally, consider a langauge ℓ_3:

	m_1	m_2	m_3
u_1	1	1	0
u_2	0	0	0
u_3	0	0	1

Our third language contains an ambiguous lexical item - u_1 can be used to refer to either m_1 or m_2. This language represents a trade-off between ℓ_1 and ℓ_2. From the speaker perspective ℓ_3 likely requires more effort from the speaker than ℓ_1, however it likely requires less than ℓ_2 (so long as $p(u_1)$ is cheap). From the listener perspective ℓ_3 requires more effort than ℓ_3 which requires no disambiguation, but less than ℓ_1.

As this small set of languages demonstrate, the efficiency of a particular language depends on both utterances costs $P(u)$ and need probabilities $P(M)$. E.g. If u_1 is actually a costly utterance ($p(u_1)$ is close to 0) then ℓ_1 is inefficient both for the speaker and listener. For this reason, in the following simulations, we randomly sample utterance costs $P(u)$ and need probabilities $P(m)$. Our simulations procced as follows. For $n = 100$ simulations we enumerate all valid langauges (4×4 boolean matrices) as well as a set of need probabilities ($p(m)$) and utterance costs ($p(u)$).

Given this set-up we can find the optimal language:

$$\ell^* = \text{argmin}_{\ell \in L} H_{Cross}(P_{Speaker}(u,m;\ell), P_{Listener}(u,m;\ell))$$

Critical to Piantadosi's claim, however is the fact that ambiguous langauges should be preferred *when* context is disambiguating. For that reason we compare the optimal langauge ℓ^* as the amount of contextual information increases. In our case a "context" describes a conditional distribution

Simulation set-up.

over meanings. For example, take a setting with two contexts c_1 and c_2. The first context dictates a set of need probabilites which differs from the second ($p(m|c_1) \neq p(m|c_2)$). If ambiguity is useful when context is disambiguating we should expect the likelihood that an optimal langauge ℓ^* contains an ambiguous item to increase as the number of contexts increases.

$$p(m) = \text{Dirichlet}(ones)$$

$$S(u|m;\ell) = \text{S1 rsa speaker}$$

Results We run $n = 100$ simulations in four condtions. Our first is a *single-context* condition ($|C| = 1$) – there is a only a single context describing $p(m|c_1)$. Our second condition contains two-contexts ($|C| = 2$) – we consider efficiency under both $p(m|c_1)$ as well as $p(m|c_2)$. The third and fourth condition correspond accordingly with $|C| = 3$ and $|C| = 4$, respectively. In each condition we consider a random set of utterances costs $P(u)$ and need probabilities $p(m|c)$ parametrized as Dirichlet distributions with $\alpha = 1$. Piantadosi et al. (2012) argued that ambiguity is an efficient property of contextualized langauge use. That is, it is useful to assign words multiple meanings if those words can be disambiguated in context. Formalized in our setting, this would mean that we expect $\ell*$ to be more likely to contain ambiguous mappings as the number of contexts increases.

Figure 1. plots the proportion of optimal languages under $H_{cross}(P_{Speaker}, P_{Listener})$ for each condition. We find that as the number of contexts increases, so does the probability that optimal language $\ell*$ contains ambiguity. For comparison we also include the optimal language under an objective that only considers speaker or listener effort. In line, with Zipf's predictions, if langauges are designed only to minimize speaker effort then optimal languages will assign all meanings to a single, low-cost utterance. Likewise, if langauges are designed only to minimize speaker listener effor then ambiguity should alwasy be avoided.

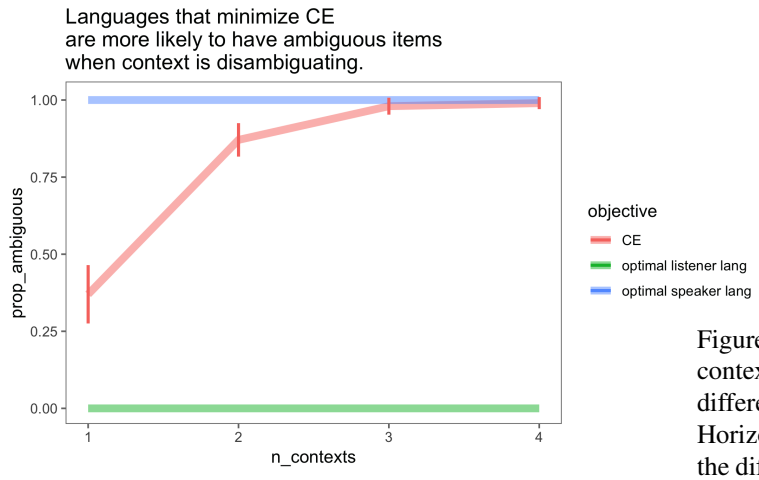

Figure 1: Optimal languages are more likely to contain ambiguous items as amount of contextual information increases. Vertical axis shows the proportion of optimal languages that contain ambiguity. Horizontal axis shows the number of contexts in each condition (1-4). Red-line represents the optimal langauge under our Zipfian cross-entropy objective while the blue and red lines show optimal languages under speaker- and listener-only consideration.

Piantadosi et al. (2012) framed their theory in terms of conditional entropy. That is, $H(X|C) < H(X)$ when C provides information about X. Put differently, when $I(X,C)$ is non-zero. In our setting this would indicate that as the amount of contextual information increases, the difference between the conditional and unconditional measures of information should increase. That is, $H_{cross}(P_{Speaker}, P_{Listener}|C_1) - H_{cross}(P_{Speaker}, P_{Listener}) < H_{cross}(P_{Speaker}, P_{Listener}|C_2) - H_{cross}(P_{Speaker}, P_{Listener})$ when C_1 contains more information than C_2. We can consider this very metric in our conditions. Figure 2 plots $H_{cross}(P_{Speaker}, P_{Listener}|C_1) - H_{cross}(P_{Speaker}, P_{Listener})$ in each of our conditions. As the number of contexts increases, the difference in efficiency when context is disambiguating increaess as well.

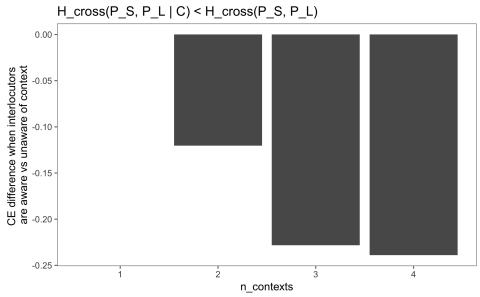

Figure 2: The gains in efficinecy increase as the amount of contextual information increases. Vertical axis shows the difference in cross-entropy for when context is informative. Horizontal axis shows each of the four conditions. Note that the difference for condition 1 is zero as thered is only a single context.

Summary Piantadosi et al. (2012) argued that it is useful to re-use low-cost linguistic forms for multiple meanings when they can be disambiguated in context. Using our speaker-listener cross-entropy measure of efficiency we showed that optimal langauges are more likely to have ambiguous items when context is informative. Further, we showed that the impact of being able to disambiguate langauge (having access to $p(m|c_i)$) is increasingly efficienct as the amount of common-ground (context) increases.

Experiment 2: Rational-pragmatic speakers use ambiguity efficiently

Our first experiment made a fairly direct test of the communicative function of ambiguity proposed by Piantadosi et al. (2012). Sampling a set of need probabilities and utterance costs we explored the space of possible languages, examining the properties of the language $\ell*$ which minimized our objective. Results indicated that ambiguity is an efficient property when context is informative. This experiment, however, assumed that our speaker and listener agents could disambiguate items context. That is, in our four conditions both agents had access to these conditional distributions $p(m|c_1), \ldots, p(m|c_4)$. More often than not in day-to-day language use, speakers and listeners may not have perfect knowledge of the particular set of need probabilities $p(m)$ being used. Put differently, if ambiguity is only efficient when it can be disambiguated in context, speakers should avoid using ambiguous items if they know their listener may not be able to disambiguate the item.

In experiment two we examine speaker behavior in a scenario in which the listener does not know the exact set of need probabilites a priori. That is, the listener has uncertainty over $p(m)$. Over the course of a discourse D the listener tries to infer the context along with the particular intended meaning of a given utterance. That is, we consider a listener model $L(c,m|u;D)$ where c is the particular context (or *topic*) being discussed and D is the set of previous utterances the speaker has made. Likewise we consider a speaker model $S(u|m,c,D)$

who chooses an utterance based on an intended meaning m, the particular context of conversation c, while also considering the history of their utterance D. If speakers are behaving efficiently they should only use ambiguous items when they can be disambiguated in context. That is, if the listener agents is able to correctly infer the topic of conversation as the conversation progresses $D \rightarrow$ inf then the speaker should be more likely to use ambiguous items later in discourse.

Simulation set-up

$\ell =$ current language known to both speaker and listener

$D =$ History of previous utterances by speaker

$S(u|m;\ell) =$ S1 rsa speaker

$L(m|u;\ell) =$ L0 rsa listener

$P(u) =$ utterance costs in which $p(u_1) < p(u_2) < p(u_3) < p(u_4)$

$c =$ current topic, which specifies need probabilities

$P(m|c) =$ set of need probabilites specified by c

Results Fig3

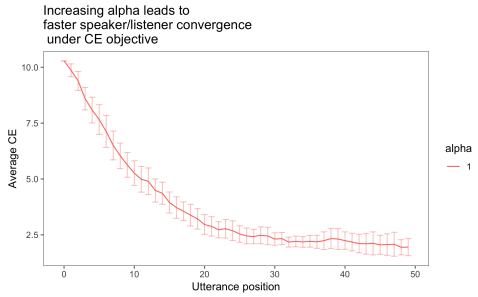

Figure 3: R plot

Fig4

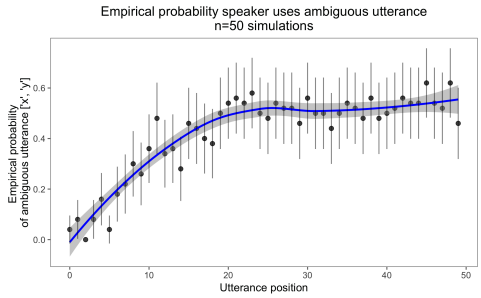

Figure 4: R plot

Summary

Experiment 3: "More" pragmatic leads to more efficiency

Simulation set-up

Results Fig5

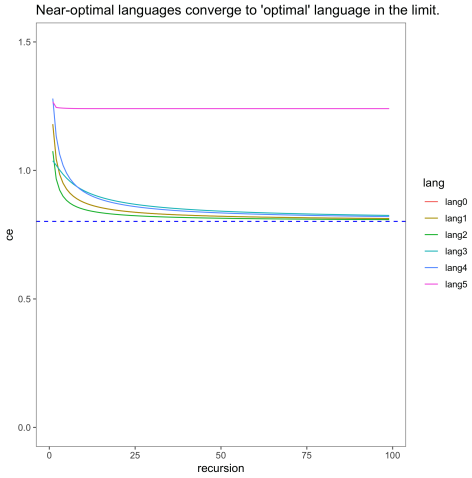

Figure 5: R plot

Summary

Discussion

Conclusion

Acknowledgements

Place acknowledgments (including funding information) in a section at the end of the paper.

References