Endowing Neural Language Learners with Human-like Biases: A Case Study on Dependency Length Minimization

Yuqing Zhang[†], **Tessa Verhoef**[‡*], **Gertjan van Noord**[†], **Arianna Bisazza**[†*]
[†]Center for Language and Cognition, University of Groningen, The Netherlands
{yuqing.zhang, g.j.m.van.noord, a.bisazza}@rug.nl

[‡]Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands
t.verhoef@liacs.leidenuniv.nl

Abstract

Natural languages show a tendency to minimize the linear distance between heads and their dependents in a sentence, known as dependency length minimization (DLM). Such a preference, however, has not been consistently replicated with neural agent simulations. Comparing the behavior of models with that of human learners can reveal which aspects affect the emergence of this phenomenon. In this work, we investigate the minimal conditions that may lead neural learners to develop a DLM preference. We add three factors to the standard neural-agent language learning and communication framework to make the simulation more realistic, namely: (i) the presence of noise during listening, (ii) context-sensitivity of word use through non-uniform conditional word distributions, and (iii) incremental sentence processing, or the extent to which an utterance's meaning can be guessed before hearing it entirely. While no preference appears in production, we show that the proposed factors can contribute to a small but significant learning advantage of DLM for listeners of verb-initial languages.

Keywords: neural-network based simulations, language universals, dependency length minimization, artificial language learning

1. Introduction

When several word order options are available to convey a message, human language speakers prefer the order that reduces the overall length of syntactic dependencies (Arnold et al., 2000; Gildea and Temperley, 2010; Futrell et al., 2020). The origins of such DLM preference remain subject of debate (Culbertson and Adger, 2014; Fedzechkina et al., 2018). Is this preference acquired by abstracting statistics from linguistic input, shaped by cognitive biases of human information processing, or arising from the pressure to communicate efficiently? A fruitful approach to studying the influence of human cognitive biases and processes like language learning and use in shaping linguistic structure is to simulate them computationally (De Boer, 2006; Steels, 1997). Recent advances in machine learning and computational linguistics have yielded powerful (neural-network based) artificial learners that can deal surprisingly well with the complexity of human languages and can be used to set up increasingly realistic simulations. Artificial neural networks rely on statistical learning for acquiring representations and assume minimal inductive biases specialized for language. By comparing the behavior of models with that of human learners, we can gain insights into the types of linguistic knowledge that can be statistically learned, and which aspects are shaped by additional cognitive constraints or communication pressures. These comparisons may also reveal that neural network models lack important human-like biases, which in turn can explain the limited abilities of current language models to generalize in linguistically sound ways (Chaabouni et al., 2021; Portelance et al., 2021; Ren et al., 2020; Warstadt and Bowman, 2022).

Previous simulation work in the context of DLM, however, has produced contradictory patterns: in production tasks involving miniature artificial languages, LSTMs were found to prefer shorter-distance dependencies (Chaabouni et al., 2019) while pre-trained Transformers like BART and T5 were not (Zhao, 2022). These two studies, however, are not comparable to each other due to different architectures, (pre-)training regimes, and miniature languages. Moreover, these studies adopted linearized meaning representations which could be implicitly biased towards shorter dependencies.

In this work, we investigate the minimal conditions that lead neural learners to develop a preference for shorter dependencies, while using a more principled setup. To this end, we follow an artificial language learning paradigm and avoid any pretraining of the networks to rule out that the observed preferences are inherited by statistical properties of real-language training corpora. Secondly, we use miniature languages that are directly inspired by an artificial language learning experiment with human subjects (Fedzechkina et al., 2018), which provides clear expectations on the results of our simulations. Lastly, we represent meaning in a way

[*] Shared senior authorship.

that is not biased toward any specific linear order, following recent work on the simulation of language universals (Lian et al., 2023).

In the standard version of this controlled setup, we find that RNN-based learners do *not* show any DLM preference in production, nor do they show any ease of learning the shorter-dependency languages compared to their longer-dependency counterparts.

We proceeded to consider three additional factors that may trigger DLM in the neural learners while making the simulation more realistic, namely: (i) the presence of noise during listening (Gibson et al., 2013; Brochhagen et al., 2017; Futrell and Levy, 2017), (ii) the strongly non-uniform nature of word distributions and selectional preferences in real languages (Katz and Fodor, 1963; McRae et al., 1998), and (iii) incremental utterance processing (Futrell and Levy, 2017; Kamide et al., 2003), i.e. the extent to which the meaning of utterances can be guessed before hearing them entirely. The results still fail to display a DLM preference in production. However, we find evidence that the proposed factors contribute to a small but significant learning advantage of shorter dependencies for listening agents under noisy conditions, as well as a higher incremental listening accuracy regardless of the presence of noise.

Our findings offer new insights into the essential elements that contribute to the development of DLM preferences in purely statistical, neural-network based learners. More generally, we demonstrate the importance of making simulations more realistic by considering factors, such as noise, complex input distributions, and incremental sentence processing, that are typically overlooked in neural agent emergent communication models [1].

2. Background

Dependency length minimization While natural languages exhibit a wide range of variations, certain patterns occur more frequently than expected by chance in world languages. Many such patterns result from a trade-off between the need to reliably exchange information and the cost of language production and processing (Kemp and Regier, 2012; Gibson et al., 2019; Kirby et al., 2015). DLM, or dependency locality, is one of the statistical language universals that has been hypothesized to result from processing efficiency during incremental production and comprehension and communicative efficiency in sending and receiving utterances. Hawkins (1994) proposes that reducing the linear order between related constituents minimizes the search time required for an incremental language parser to determine the correct head of a phrase. Another proposal (Gibson, 1998) is that long dependencies exert extra pressure on working memory in both language parsing and generation by requiring speakers to keep word representations in the working memory for a longer time. Whether it is due to a memory constraint or search-time constraint, DLM is regarded as inherent to the underlying language processing mechanism which is commonly assumed to be universal rather than dependent on a given speaker's language. Specifically, DLM refers to the tendency of natural languages to minimize the linear distance between words linked in the grammatical head-dependent relationships (Futrell et al., 2015). For example, users of verb-initial languages tend to place short postverbal constituents before long ones (Wasow, 2002). In contrast, users of verb-final languages (i.e., languages that place the verb after its dependents) typically prefer long preverbal constituents before short ones (Yamashita and Chang, 2001). The universality of DLM has been established by both psycho-linguistic studies (Gibson, 1998; Grodner and Gibson, 2005) and corpus-based analyses (Cancho et al., 2004; Liu, 2008; Futrell et al., 2015). The latter study, in particular, found that the observed word orders of human languages have a shorter dependency length than random baselines even though the level of DLM optimization varies considerably (Futrell et al., 2015). The origin of DLM has also been investigated with the artificial language learning paradigm (Fedzechkina et al., 2018, 2020; Zhao, 2022). These studies found that subjects exposed to novel miniature languages systematically restructured the input toward shorter dependency lengths, regardless of the order preferences found in their native language. The artificial languages used in our work (Figure 3) are directly inspired from Fedzechkina et al. (2018).

Neural-network based simulations An increasing number of researchers adopt neural networks to explore the extent to which abstract linguistic knowledge can be acquired via statistical learning, or whether specific inductive biases are needed for linguistic patterns to emerge. Some of these studies were specifically focused on DLM: Chaabouni et al. (2019) trained sequence-to-sequence LSTM agents to communicate about paths in a simple grid world using miniature languages. Their experiments show higher learning speed for the local-dependency languages, which results in increased production of local utterances across generations of a simulated iterated learning procedure. Since their meaning representation was itself a sequence, it cannot be ruled out that the RNN sequence-to-sequence agents were actually biased towards languages having less reordering w.r.t. the input

[1] All code and data are available at: `https://github.com/yuqing0304/DLM_exp`

meaning sequence. In a different study, Futrell and Levy (2019) found that RNNs trained on English texts prefer ordering short constituents before long ones regarding several DLM-related linguistic phenomena (heavy NP shift, particle shift, dative alternative, and genitive alternative). Zhao (2022) did similar tests using pre-trained transformers (i.e. BART (Lewis et al., 2020) and GPT2 (Radford et al., 2019)) and found these models acquire human-like word order preferences that are consistent with the DLM principle. However, the same models did not show a similar preference in production tasks involving an artificial semi-English language, which might be due to the total lack of memory constraints in the transformer architecture (through the self-attention mechanism). In summary, the essential elements contributing to the emergence of DLM preferences in neural learners remain largely unknown.

3. Experiment

We follow the Neural-agent Language Learning and Communication framework (NeLLCom), which was recently developed by Lian et al. (2023) for the replication of language universals with neural learners. In this framework, neural-network agents are trained to exchange messages in a simplified world of agent-patient-action triplets (e.g. *Tom-Jerry-chase*) using pre-defined miniature languages. Listening is defined as the process of converting a sequence of symbols (utterance: u) into an unordered set of items (meaning: m), whereas speaking is the reverse process of converting a meaning into an utterance. After a listening/speaking training phase based on supervised learning (SL), pairs of agents communicate with each other using the learnt language while maximizing communication success via reinforcement learning (RL). Specifically, both agents' goal is to maximize the listener's ability to reconstruct the intended meaning m, given a speaker-generated utterance $\hat{u}$. Testing is performed on meanings not observed during any training phase.

To study DLM in this work, we expose the agents to various flexible-order case-marking miniature languages where long- and short-dependency utterances occur with different distributions (illustrated in Figure 3 and explained in detail in Section 3.2). If the learners have an intrinsic bias towards DLM, we expect to observe (i) a higher learning accuracy for the language that has overall shorter dependencies, and/or (ii) a speaker's preference to produce short-dependency utterances during testing.

3.1. Agent Architectures

The listening network has a *sequence-to-linear structure* (see Figure 1). The input utterance is passed to the listening agents' RNN encoder[2] and consumed word by word until the EOS token. The listener then passes the last hidden state to n parallel linear layers, one for each aspect of the meaning. Finally, each of the n elements is generated through a softmax layer. The speaking network has a mirrored *linear-to-sequence* architecture (see Figure 2).

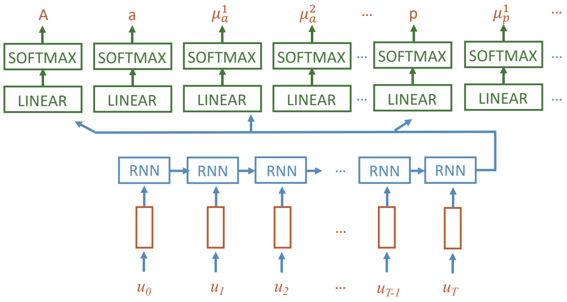


Figure 1: RNN-based listening agent

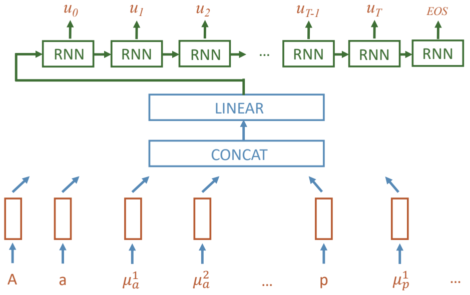


Figure 2: RNN-based speaking agent

To experiment with different dependency lengths, we expand Lian et al. (2023)'s original meaning space composed of action-agent-patient triplets ($\{A, a, p\}$, $n = 3$) by adding optional modifier phrases to agent and patient: μ_a and μ_p, respectively. Each modifier phrase consists of three items corresponding, respectively, to adposition, adjective, and inanimate noun in Fedzechkina et al. (2018)'s experiment (e.g. *by frozen river*). Thus, the resulting meaning space has $n = 9$.

3.2. Miniature Languages

Following Fedzechkina et al. (2018), we design (i) a **verb-initial** language that has flexible order (either verb-subject-object (VSO) or verb-object-subject (VOS)), and nouns that can be modified with postnominal prepositional phrases (e.g., *Jerry*

[2] Following Lian et al. (2023), we use Gated Recurrent Units, or GRU (Cho et al., 2014) in all experiments. In the preliminary phase of our experiments, we observe similar results obtained from LSTM and GRU models.

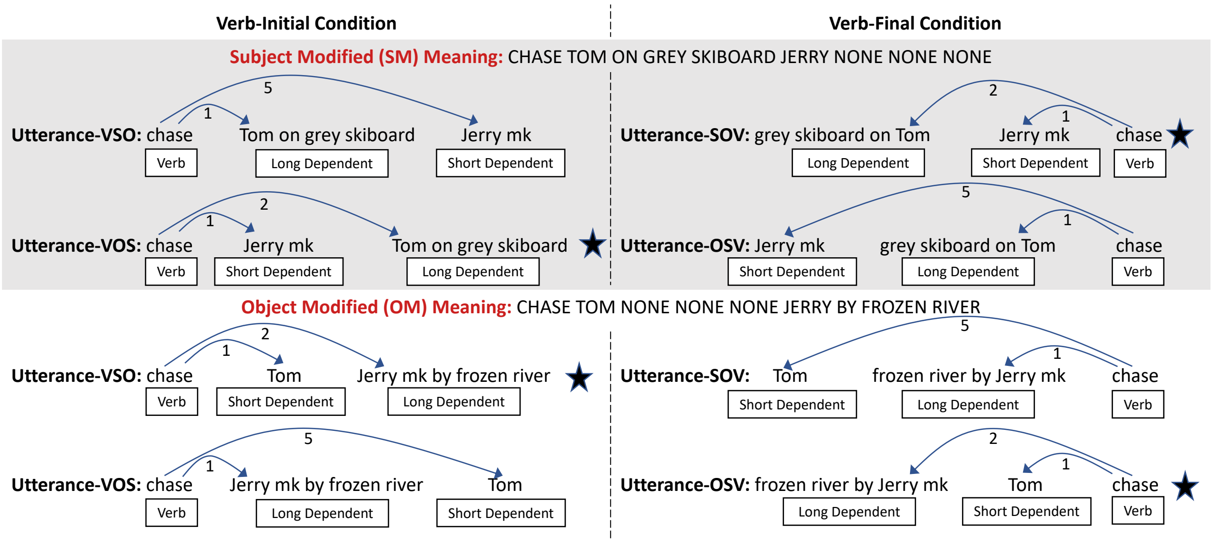

Verb-Initial Condition
Verb-Final Condition
Subject Modified (SM) Meaning: CHASE TOM ON GREY SKIBOARD JERRY NONE NONE NONE
Utterance-VSO: chase Tom on grey skiboard Jerry mk
Verb
Long Dependent
Short Dependent
1
5
Utterance-VOS: chase Jerry mk Tom on grey skiboard
1
2
Utterance-SOV: grey skiboard on Tom Jerry mk chase
2
1
Utterance-OSV: Jerry mk grey skiboard on Tom chase
5
1
Object Modified (OM) Meaning: CHASE TOM NONE NONE NONE JERRY BY FROZEN RIVER
Utterance-VSO: chase Tom Jerry mk by frozen river
1
2
Utterance-VOS: chase Jerry mk by frozen river Tom
1
5
Utterance-SOV: Tom frozen river by Jerry mk chase
5
1
Utterance-OSV: frozen river by Jerry mk Tom chase
2
1

Figure 3: The miniature languages used in our simulations, inspired from Fedzechkina et al. (2018). Case marking ('mk') occurs after all objects. Shaded parts contain utterances expressing the same meaning. Curved arrows represent grammatical dependencies between the verb and its two dependents. Numbers represent dependency lengths, measured in words. Since the relative positions of other constituents stay the same for each condition, we operationalize DLM as minimizing the linear distance between the verb and the head of subject and the head of the direct object. Total dependency length is minimized (starred) by placing short-before-long dependents in the verb-initial language (left), and vice versa by placing long-before-short dependents in the verb-final language (right). For simplicity, case markers are not counted when calculating dependency length.

by frozen river for the meaning describing the scene JERRY BY FROZEN RIVER; (ii) a **verb-final** language that also has flexible order (SOV/OSV), and nouns that can be modified with prenominal prepositional phrases (e.g., *frozen river by Jerry* for the meaning JERRY BY FROZEN RIVER). The ordering of the adposition (e.g., *by*) relative to its dependent (e.g., *frozen river*) and head (e.g., *Jerry*) follows the typologically frequent distribution (e.g., *frozen river by Jerry* for JERRY BY FROZEN RIVER in the verb-final language) (Fedzechkina et al., 2018, 2020; Zhao, 2022).

The meanings are descriptions of scenes that have only one long constituent (i.e. only subject or object has adpositional-phrase modification). For all languages, half of the meanings are subject-modified (SM), while the other half are object-modified (OM). The subject and object are never the same within a single meaning. For each meaning, there are two possible utterance orderings (SO or OS). For all *local* languages, i.e., *skewed_local* and *uniform_local*, utterances are generated by placing short dependents closer to the head verb (see Figure 3, starred utterances). Similarly, for *long* languages, i.e., *skewed_long* and *uniform_long*, utterances are generated by placing long dependents closer to the head verb (see Figure 3, non-starred). For languages mixed with both local and long dependency utterances, i.e., *skewed* and *uniform*, we sample half of the subject-modified meanings and then generate local dependency utterances, then generate long dependency utterances for the other half of the subject-modified meanings. We repeat the same steps for object-modified meanings. See Figure 3 for an illustration of both languages.

The two languages share a common lexicon consisting of six transitive verbs, six animate nouns, four adpositions, three adjectives, and three inanimate nouns, leading to vocabulary size |V| of 6+6+4+3+3+1(marker)=23.

3.3. Noisy Communication

Real-life communication between speakers and listeners is often subject to various sources of noise and errors that occur during transmission, for instance due to external factors or limited listener attention (Gibson et al., 2013; Brochhagen et al., 2017). While this aspect is not often considered when designing neural network agent-based simulations (but see Ueda and Washio (2021); Chaabouni et al. (2022) for exceptions), it may play an important role in the emergence of DLM and other universals. Notably, Futrell and Levy (2017) found that a model of sentence processing in which the context is noisy predicts that short dependency sentences will be easier to process.

To simulate noise, we adopt a **word dropout** technique with dropout rate δ with which randomly chosen parts of the input are masked to the listener.

$$\delta \sim \text{Bernoulli}(p), \quad p \in \{0, 0.1, 0.2\} \tag{1}$$

Specifically, words that are sampled by a Bernoulli distribution (with probability p ranging from 0 to 0.2 in our setup) get their word embedding replaced by that of a dummy MASK token which is never updated during training (Gal and Ghahramani, 2016; Sennrich et al., 2016).

While probabilistic dropout is commonly used to avoid overfitting when training neural networks, we apply it here to both training *and* testing to simulate what listeners will encounter during their learning and communication. This approach allows us to more accurately replicate real-world communication scenarios, where listeners must rely on various cues and context to comprehend the speaker's message under imperfect transmission conditions.

3.4. Conditional Word Distributions

Artificial language learning experiments with human participants or neural agents typically assume meaning spaces where all items are uniformly distributed (Fedzechkina et al., 2016). In reality, however, human languages exhibit highly skewed word frequency distributions resembling Zipfian patterns. Moreover, certain word combinations tend to co-occur much more frequently than others, an especially well-known tendency for verbs and their arguments, called selectional preference (Katz and Fodor, 1963). More generally, Futrell et al. (2019) demonstrated that syntactic heads and their dependents are characterized by word pairs with especially high mutual information.

Incorporating more realistic input distributions in artificial language learning simulations may be crucial to ensure generalizability of findings to real languages (Hupkes et al., 2019). In our context particularly, endowing head-dependent pairs with realistic statistical properties may be needed to simulate the emergence of DLM. In fact, while human subjects have a pre-existing notion of head-dependent relations and are likely to recognize such relations in the novel artificial language provided in the lab, our neural learners have no prior experience of language or the world and simply perceive sequences of random symbols at the beginning of training.

By introducing verb-subject/object selectional preferences in our language design, we test whether shorter dependencies may strengthen the listener's ability to recover missing bits of a sentence by relying on the presence of frequently co-occurring words in the nearby context. Concretely, we construct a **skewed** and a **uniform** version of the verb-initial and verb-final languages. In the skewed version, we account for the tendency of certain animate nouns to co-occur more frequently with specific verbs, as well as their varying probabilities of taking agent or patient roles given each transitive verb. Thus, animate nouns have different probabilities of being agent or patient *given* a particular verb. To generate the skewed version of each language, we use the following discrete distributions which roughly correspond to six equally spaced sample points from the Zipf distribution (Zipf, 1949), as shown in Equation 2 (N stands for the number of elements; k is the rank counting from 1; s is the exponent parameter). We set the number of elements N as 6 and the exponent parameter s as 1.5: $P_{\text{agent}} = [0.55, 0.19, 0.11, 0.07, 0.05, 0.04]$ for the probabilities of six randomly shuffled animate nouns being the agent given each verb, and the same distribution for the probabilities of six randomly shuffled animate nouns being patient given each verb.

$$f(k; s, N) = \frac{1}{H_{N,s}} \frac{1}{k^s} \tag{2}$$

By contrast, the uniform versions assign an equal probability to all animate nouns for being agents or patients with each verb, i.e., $P_{\text{agent}} = [1/6, 1/6, 1/6, 1/6, 1/6, 1/6]$, and similarly for P_{patient}.

For simplicity, all meaning items, except for subject and object, occur with uniform probability.

3.5. Data and Model Training

We train the agents on six different types of language (*skewed_local, skewed, skewed_long, uniform_local, uniform, uniform_long*) to study the interplay between DLM and the above-mentioned factors. To study speakers' production preferences after learning and communication, we focus on the *skewed* and *uniform* languages, which have 50% long- and 50% short-dependency utterances. A total of 1418 meaning-utterance pairs are generated for each language. The data generation process is explained in detail in Appendix A.

To evaluate the agents' ability to convey new meanings, we split the dataset into 66.7% for training and 33.3% for testing purposes.

The speaker network has an 8-dimensional embedding layer followed by a 64-dim. GRU layer. The listener has an 8-dim. embedding followed by a 64-dim. GRU layer. Training used a batch size of 32 and was limited to maximum 60 epochs. Learning rate was set to 0.01. Experiments are repeated with 40 random seeds for each language type.

3.6. Evaluation Metrics

Speaking Accuracy measures whether a generated utterance fully matches the one in the dataset (1) or not (0). For languages that combine short and long dependencies and therefore have two acceptable utterances to express the same meaning, we count 1 if any of the two gold utterances is matched, otherwise 0[3].

Listening agents are evaluated by **meaning item-level accuracy**, which is the proportion of the 9 meaning items ($\{A, a, p, \mu_a^1, \mu_a^2, \mu_a^3, \mu_p^1, \mu_p^2, \mu_p^3\}$) that are correctly predicted by the listener, averaged over all test utterances.

To analyse the experimental measures, we fit Bayesian multilevel linear models using the `brms` (Bürkner, 2017) package in R (R Core Team, 2023). Our data involves complex structures where multiple observations of accuracy measures are nested within various experimental conditions and these models provide a robust framework for analyzing such data. We investigate the impact of epochs (centered at the final epoch), conditional word distributions (categorical: skew or uniform), noise levels (categorical: dropout 0, 0.1, or 0.2), and dependency conditions (categorical: local, mixed, or long) on both speaking and listening accuracy. Values for epoch were centered so that the intercept is focused on agents' learning performance at the end, thereby ensuring that measured effects represent relationships between variables concerning the final stable stage. Therefore, in our statistical analysis, the reference point of epoch 0 represents the final point of the epochs instead of the first epoch. Importantly, we accounted for all possible interaction terms among these factors to capture relationships within the data. Default priors of the `brms` package were used for all models. All models included four chains with 4,000 iterations each and a total of 2,000 post-warm-up samples.

Additionally, we measure **incremental listening accuracy** by letting trained listeners process varying-length prefixes of the utterances. This is inspired by the fact that human listeners process incoming input sentences incrementally without delay and unfold parts of meaning when the input is still incomplete (Kamide et al., 2003; Altmann and Mirković, 2009; Futrell and Levy, 2017). For instance, for length 2 in the utterance *'chase Tom behind white door Jerry mk'*, we feed the prefix *'chase Tom'* to a listening agent, which encodes it into a hidden state via RNN and uses this to predict the full intended meaning of the utterance. Incremental listening accuracy is then calculated similarly to item-level accuracy, but for each prefix length. Note that we only apply incremental accuracy at test time. Using this metric as a reward to optimize listening agents (Rita et al., 2020) is an interesting avenue for future work.

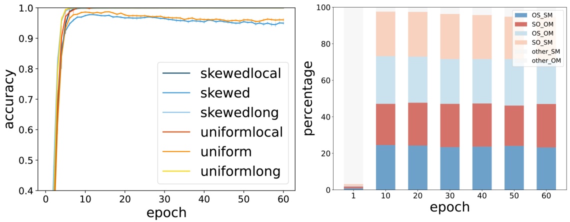
All plotted accuracies are computed over the unseen test set and averaged over all random seeds.

4. Supervised Learning Results

This section presents the results of training speaking and listening agents on languages with varying proportions of short/long dependencies, and their interplay with the presence of noise and non-uniform conditional word distributions.

4.1. Speaking Agents

During SL, long-dependency languages are learned equally well at the end as local-dependency languages (see Figure 4 (left) for the verb-initial languages, mostly similar results were found for the verb-final languages (see Appendix B for some differences).



(a) Acc initial

(b) %Order skewed initial

Figure 4: (a) Speaking accuracy as a function of training epoch for verb-initial languages. (b) Production patterns of speaking agents after learning the verb-initial *skewed* language. Color denotes word order (blue: OS, red: SO). Shading denotes dependency length (dark: local, light: long). Utterances not belonging to these categories are colored in grey.

Languages with mixed dependency utterances were found to be harder to learn for speaking agents. Specifically, statistical analyses showed significantly lower speaking accuracies in verb-initial languages with mixed dependency utterances compared to those with local dependency utterances (*b* = -.042, Bayesian 95 % Credible Interval [-.058, -.026]). Similarly, for verb-final languages, languages with mixed dependency utterances led to lower speaking accuracies compared to their counterparts with local dependency utterances (*b* = -.039, 95 % CI [-.055, -.023]). These differences possibly result from the high variation and entropy of languages with mixed dependency utterances.

Production preferences of speaking agents at the end of training (Figure 4, right) show that RNN learners preserve the distribution of long/short dependencies found in the training data, thus not displaying any DLM in a purely SL setting. This is in contrast with Chaabouni et al. (2019)'s results with sequence-to-sequence agents, but aligns well with previous studies showing that neural learners exhibit strong probability-matching behavior after SL (Lian et al., 2021, 2023).

[3] This corresponds to the *permissive accuracy* approach used to evaluate speakers in Lian et al. (2023).

4.2. Listening Agents

Figure 5 shows average listening accuracy under different noise conditions (word dropout 0: no noise, 0.2: maximal noise).

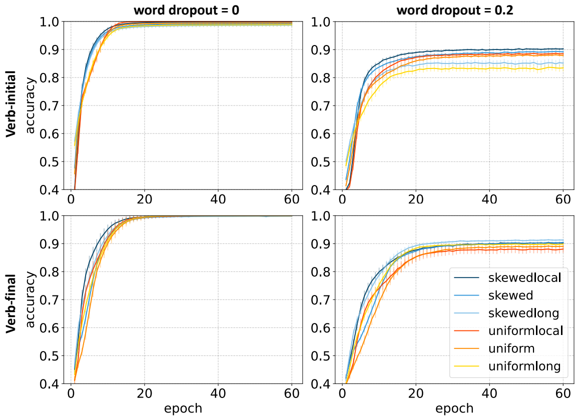


Figure 5: Listening accuracy (meaning item-level) as a function of training epoch. Color denotes word frequency distribution (blue: skewed, orange: uniform). Darkness denotes the length of dependency (dark: local, light: long). Results in this and subsequent plots are averaged over 40 random seeds.

Interplay between noise and DLM Results show that for verb-initial languages, listening accuracy is significantly lower when noise is present or when languages are of the long or mixed dependency type. Specifically, listening accuracy is substantially reduced by .050 under the noise level of 0.1 (*b* = -.050, 95 % CI [-.059, -.040]) and by .111 under the noise level of 0.2 (*b* = -.111, 95 % CI [-.120, -.101]) compared to no noise conditions. Compared to local dependency languages, listening accuracy is .022 lower for those of the long dependency type (*b* = -.022, 95 % CI [-.031, -.012]), and .010 lower when the language is of mixed dependency type (*b* = -.010, 95 % CI [-.020, -.001]). This indicates that the learning accuracy of local verb-initial languages surpasses that of long and mixed languages (*local>mixed>long*). In addition, the advantage of the local dependency language over the long dependency one becomes stronger as the learning process progresses through successive epochs or as the noise level is increased, as evidenced by the two-way interaction between epoch and long dependency language type (*b* = -.029, 95 % CI [-.044, -.014]) and interaction between the dropout level 0.1 and long dependency language (*b* = -.015, 95 % CI [-.028, -.002]). Moreover, as the learning process progresses through successive epochs, we observed a significant three-way interaction between epoch, noise, and long dependency language type. Specifically, the detrimental effect of noise on listening strengthens over time for verb-initial long-dependency languages (dropout 0.2: *b* = -.033, 95 % CI [-.055, -.012]).

Though the difference is small, it suggests that verb-initial languages with shorter dependency lengths might be easier to learn for RNN-based neural agents when noisy conditions are provided.

By contrast, in the **verb-final** languages, only noise (e.g., dropout 0.2: *b* = -.109, 95 % CI [-.116, -.103]) is a significant predictor of listening accuracy apart from epochs. While local and long languages are learned equally well without noise, a significant interaction between the noise level 0.2 and the long dependency language (*b* = .016, 95 % CI [.007, .024]), indicates that long dependency languages are learned better than local ones with noise. Interestingly, in their large-scale corpus study, Futrell et al. (2015) also reported that *real* verb-final languages tend to show less DLM compared to verb-initial languages, which instead seem to be highly optimized for efficient processing. In addition, this finding aligns with Jing et al. (2022)'s refined formulation of dependency length optimization since the DLM principle fails for certain verb-final languages.

Interplay between noise and conditional word distributions The presence of noise affects the learning accuracy of uniform languages more severely than languages with skewed conditional word distributions (see Figure 5). This holds for both verb-initial and verb-final languages. Specifically, the Bayesian models show that noise interacts with word distributions. In the absence of noise, listeners achieve a similar level of accuracy when learning uniform and skewed languages. However, with increasing levels of noise, the listening accuracy of skewed languages is higher compared to no noise (dropout 0.2, verb-initial: *b* = .023, 95 % CI [.010, .036]; dropout 0.2, verb-final: *b* = .017, 95 % CI [.008, .026]), suggesting that selectional preferences are indeed used as extra cues to recover the intended meaning from context. For example, the presence of certain verbs (e.g. *teach*) suggests that a specific animate noun (e.g. *professor*) is a more probable candidate for the agent *vs.* patient of an utterance.

Per-category listening accuracy of verb-initial languages Languages with skewed conditional word distributions are less affected by noise than uniform languages, with overall listening accuracy remaining higher for skewed languages under increasingly noisy conditions. Here we zoom in on the accuracy of specific meaning items in a verb-initial language, by plotting listening accuracy separately for each individual meaning item as a function of training epoch in Figure 6. We find that the advantage of the *skewed* languages under word dropout 0.2 concentrates on Action, agent, and

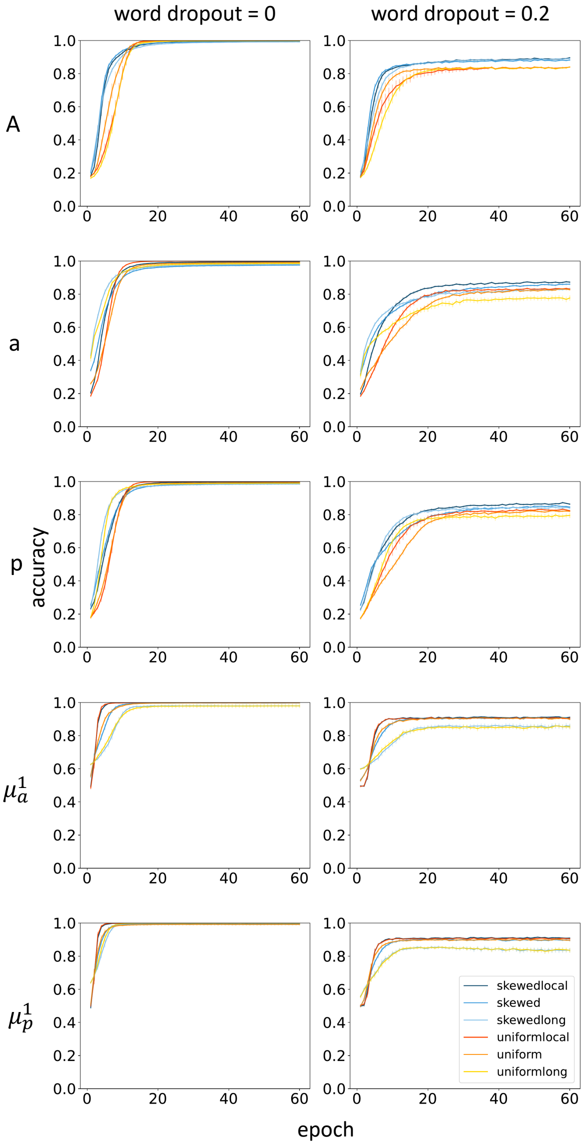


Figure 6: **Verb-initial language:** Listening accuracy computed separately for each individual meaning item, as a function of the training epoch. We see similar results for $\{\mu_a^1, \mu_a^2, \mu_a^3, \mu_p^1, \mu_p^2, \mu_p^3\}$. Therefore, only μ_a^1 and μ_p^1 are plotted here for illustration.

patient accuracy, but does not affect the modifiers.

Incremental listening accuracy As shown in Figure 7, in **verb-initial** languages, shorter-dependency languages are overall advantaged in terms of incremental accuracy. Specifically, accuracy of long and local languages is comparable until a prefix of length 3, but improves much faster for local languages after that. The general trend with increasingly long prefixes is that the accuracy starts off similar, then becomes different in the middle and eventually converges again by longer prefix or full sentences. This observation suggests that in verb-initial languages, the disambiguation of meanings related to local dependencies tends to occur earlier in the sentence. The language with mixed long and local dependency utterances (*skewed* or *uniform*), presents an intermediate level of complexity since it includes both types of utterances.

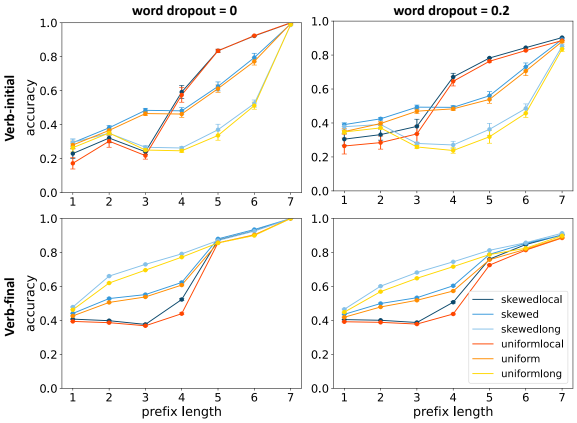


Figure 7: Incremental listening accuracy (meaning-item level) as a function of prefix length.

In **verb-final** languages, we find a reversed pattern: with increasing prefix length, the *long*-dependency utterances are advantaged in the beginning before converging around length 5. This could result from the tendency in local utterances of verb-final languages to start with a relatively long prenominal prepositional phrase, not revealing any information about the head and its dependents until later.

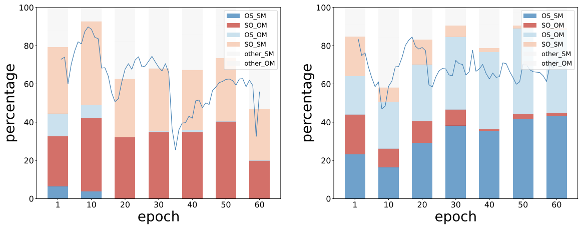
5. Communication Learning Results

We investigate here whether the slight learning advantage observed for short-dependency languages in verb-initial languages could lead to a progressively stronger DLM of our languages across repeated RL communication rounds.

The results across 40 random seeds show a consistent behavior of regularizing towards one word order, but not towards shorter dependencies. Figure 8 shows the production preferences of two representative pairs of speaking/listening agents trained on the skewed language without dropout.[4] These show two opposite regularization strategies: the first agent pair (a) quickly stops producing VOS utterances and regularizes to VSO order, whereas the second pair (ii) regularizes to VOS order. In both cases, the distribution of local vs. long dependencies remains uniform and the production patterns of other seeds all resemble either one of these two cases (cf. Appendix C).

We note that, in the language design we inherited from Fedzechkina et al. (2018), word order regu-

[4] The production patterns of other seeds and for other settings (uniform languages, with dropout) are all similar in terms of regularization strategies.



(a) regularize to SO (b) regularize to OS

Figure 8: Individual production patterns of two selected speaking agents during RL for the verb-initial *skewed* language, no dropout. Color denotes word order (blue: OS, red: SO). Shading denotes dependency length (dark: local, light: long). Utterances not belonging to these categories are colored in grey. (a): Regularization towards SO word order. (b): Regularization towards OS word order. Meaning reconstruction accuracy per epoch is shown by the blue line.

larization is in direct competition with dependency length regularization. This did not turn out to be a problem in their experiments, perhaps because human subjects do not have a strong tendency to regularize word order in the presence of case marking, as observed by Fedzechkina et al. (2017). However, neural agents were previously observed to reduce word order entropy when communicating (Lian et al., 2023), and this could explain why DLM does not emerge in our communication setup.

6. Discussion

Neural-agent simulations of artificial language learning are a promising approach to study the origins of many language universals, however, designing *realistic* simulations is essential to ensure the generalizability of findings to real languages. In this work, we focused on dependency length minimization (DLM) and suggested three important factors of human language processing and communication that should be included in the experimental setup: the presence of noise, non-uniform conditional word distributions, and the importance of incremental utterance processing. We found evidence that the proposed factors contribute to a small but significant learning advantage of shorter dependencies for listening agents of verb-initial languages. Specifically, 1) under noisy conditions, listeners learn the short-dependency verb-initial language slightly better and faster than the long-dependency one; 2) the presence of noise affects the learning accuracy of uniform languages more severely than languages with skewed conditional word distributions. 3) the verb-initial *local* language shows an advantage over the *long* one when evaluated incrementally.

In contrast, we don't see the same results for verb-final languages, where we even see a slight advantage for long dependency languages, both in SL with noisy conditions and in the beginning during incremental processing. Interestingly, as mentioned before, prior corpus studies have also shown that the DLM principle doesn't always apply to verb-final languages (Futrell et al., 2015; Jing et al., 2022).

For the speaking agents, we see no regularization behavior towards DLM after SL, nor during communication (RL). A possible reason is that in our language design, languages with a mix of long- and short-dependency utterances have both word order alternatives to describe subject-modified scenes and the same for object-modified scenes, making it impossible for neural learners to regularize both word order *and* dependency length. Under the current setting, our learners seem to prefer the former over the latter. Future work could modify the miniature language design so that both types of regularization can happen concurrently.

Another possible solution may lie in the implementation of memory constraints (Vogelzang et al., 2017), which have been proposed to be the main influencing factor for DLM in humans (Gibson, 1998; Futrell et al., 2020). Although RNNs must compress all processed information into a fixed-size hidden representation, their memory limitations may not be severe enough to induce a preference for placing the head verb and its dependents close to each other in our miniature languages, even when these items are useful to restore missing parts of the input in noisy conditions. Additionally, though RNNs have inherent structural sequence-processing memory biases, the precise process necessary to make such memory constraints human-like is still a debated topic. We are aware of one very relevant proposal for a more cognitively plausible, memory-limited neural architecture (Timkey and Linzen, 2023), which appeared concurrently to our work. In future work, a more explicit control of memory constraints, for instance through the use of masked self-attention in transformer-like architectures, limiting models' contextual access explicitly, or using a more cognitively plausible, memory-limited neural architecture, could shed more light on the origins of DLM and other language universals.

Acknowledgements

Zhang was partly funded by the China Scholarship Council (CSC 202207110004). We are grateful to the three anonymous reviewers whose comments and suggestions contributed to improving this work and its presentation.

7. Bibliographical References

Gerry TM Altmann and Jelena Mirković. 2009. Incrementality and prediction in human sentence processing. *Cognitive science*, 33(4):583–609.

Jennifer E Arnold, Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76(1):28–55.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Thomas Brochhagen, Michael Franke, et al. 2017. Effects of transmission perturbation in the cultural evolution of language. In *CogSci*.

Paul-Christian Bürkner. 2017. brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80:1–28.

Ramon Ferrer Cancho, Ricard V Solé, and Reinhard Köhler. 2004. Patterns in syntactic dependency networks. *Physical Review E*, 69(5):051915.

Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. 2021. Communicating artificial neural networks develop efficient color-naming systems. *Proceedings of the National Academy of Sciences*, 118(12):e2016569118.

Rahma Chaabouni, Eugene Kharitonov, Alessandro Lazaric, Emmanuel Dupoux, and Marco Baroni. 2019. Word-order biases in deep-agent emergent communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5166–5175.

Rahma Chaabouni, Florian Strub, Florent Altché, Eugene Tarassov, Corentin Tallec, Elnaz Davoodi, Kory Wallace Mathewson, Olivier Tieleman, Angeliki Lazaridou, and Bilal Piot. 2022. Emergent communication at scale. In *International conference on learning representations*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Jennifer Culbertson and David Adger. 2014. Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences*, 111(16):5842–5847.

Bart De Boer. 2006. Computer modelling as a tool for understanding language evolution. In *Evolutionary Epistemology, Language and Culture: A Non-Adaptationist, Systems Theoretical Approach*, pages 381–406. Springer.

Maryia Fedzechkina, Becky Chu, and T Florian Jaeger. 2018. Human information processing shapes language change. *Psychological science*, 29(1):72–82.

Maryia Fedzechkina, Elissa L Newport, and T Florian Jaeger. 2016. Miniature artificial language learning as a complement to typological data. *The usage-based study of language learning and multilingualism*, pages 211–232.

Maryia Fedzechkina, Elissa L. Newport, and T. Florian Jaeger. 2017. Balancing effort and information transmission during language acquisition: Evidence from word order and case marking. *Cognitive Science*, 41(2):416–446.

Masha Fedzechkina, Charles Torres, and Yiyun Zhao. 2020. Dependency length minimization: An abstract bias or an input-driven preference? In *The 26th architectures and mechanisms for language processing conference (AMLaP)*.

Richard Futrell and Roger Levy. 2017. Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 1, long papers*, pages 688–698.

Richard Futrell and Roger Levy. 2019. Do rnns learn human-like abstract word order preferences? In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 50–59.

Richard Futrell, Roger P Levy, and Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Language*, 96(2):371–412.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Richard Futrell, Peng Qian, Edward Gibson, Evelina Fedorenko, and Idan Blank. 2019. Syntactic dependencies correspond to word pairs with high mutual information. In *Proceedings of the fifth international conference on dependency* *linguistics (depling, syntaxfest 2019)*, pages 3–13.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. *Advances in neural information processing systems*, 29.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson, Leon Bergen, and Steven T Piantadosi. 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056.

Edward Gibson, Richard Futrell, Steven P Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends in cognitive sciences*, 23(5):389–407.

Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310.

Daniel Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive science*, 29(2):261–290.

John A Hawkins. 1994. *A performance theory of order and constituency*. 73. Cambridge University Press.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2019. The compositionality of neural networks: integrating symbolism and connectionism. *arXiv preprint arXiv:1908.08351*, 3.

Yingqi Jing, Damián E Blasi, and Balthasar Bickel. 2022. Dependency-length minimization and its limits: A possible role for a probabilistic version of the final-over-final condition. *Language*, 98(3).

Yuki Kamide, Gerry TM Altmann, and Sarah L Haywood. 2003. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and language*, 49(1):133–156.

Jerrold J Katz and Jerry A Fodor. 1963. The structure of a semantic theory. *language*, 39(2):170–210.

Charles Kemp and Terry Regier. 2012. Kinship categories across languages reflect general communicative principles. *Science*, 336(6084):1049–1054.

Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. 2015. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yuchen Lian, Arianna Bisazza, and Tessa Verhoef. 2021. The effect of efficient messaging and input variability on neural-agent iterated language learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10121–10129.

Yuchen Lian, Arianna Bisazza, and Tessa Verhoef. 2023. Communication Drives the Emergence of Language Universals in Neural Agents: Evidence from the Word-order/Case-marking Trade-off. *Transactions of the Association for Computational Linguistics*, 11:1033–1047.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Ken McRae, Michael J Spivey-Knowlton, and Michael K Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in online sentence comprehension. *Journal of Memory and Language*, 38(3):283–312.

Eva Portelance, Michael C Frank, Dan Jurafsky, Alessandro Sordoni, and Romain Laroche. 2021. The emergence of the shape bias results from communicative efficiency. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 607–623, Punta Cana, Dominican Republic.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Yi Ren, Shangmin Guo, Matthieu Labeau, Shay B Cohen, and Simon Kirby. 2020. Compositional languages emerge in a neural iterated learning model. In *8th International Conference on Learning Representations*, Addis Ababa, Ethiopia.

Mathieu Rita, Rahma Chaabouni, and Emmanuel Dupoux. 2020. “LazImpa”: Lazy and impatient neural agents learn to communicate efficiently. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 335–343, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376.

Luc Steels. 1997. The synthetic modeling of language origins. *Evolution of communication*, 1(1):1–34.

William Timkey and Tal Linzen. 2023. A language model with limited memory capacity captures interference in human sentence processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8705–8720.

Ryo Ueda and Koki Washio. 2021. On the relationship between Zipf's law of abbreviation and interfering noise in emergent languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 60–70, Online. Association for Computational Linguistics.

Margreet Vogelzang, Anne C Mills, David Reitter, Jacolien Van Rij, Petra Hendriks, and Hedderik Van Rijn. 2017. Toward cognitively constrained models of language processing: a review. *Frontiers in Communication*, 2:11.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. *Algebraic Structures in Natural Language*, pages 17–60.

Thomas Wasow. 2002. *Postverbal behavior*. CSLI Stanford.

Hiroko Yamashita and Franklin Chang. 2001. “long before short” preference in the production of a head-final language. *Cognition*, 81(2):B45–B55.

Yiyun Zhao. 2022. *How to Probe Linguistic Knowledge and Bias*. Ph.D. thesis, The University of Arizona.

George Kingsley Zipf. 1949. Human behaviour and the principle of least-effort. *Addison-Wesley*, 24.

A. Data Generation Details

For data generation, we first generate meanings containing all combinations of adpositions, adjectives, and inanimate nouns. Then for the skewed language type, we combine all previously generated meanings with the most probable subject. For the remaining subject candidates, we sample the previously generated meanings and then combine these meanings with subject candidates based on the probability ratio. Then we repeat this step to add the meaning category of the patient. Generating data by this simple method can guarantee that items in categories other than Action, agent, and patient occur with equal probability and the conditional frequencies of Action, agent, and patient strictly obey the selected discrete skewed distribution. For the uniform version, we generate all possible combinations of verbs, subjects, objects, and their modifiers. Then randomly sample an equal number of meanings.

B. Differences in learning curves for speakers of verb-initial and verb-final languages

Though mostly similar results were found for verb-initial languages and verb-final languages, we observe a larger increase in accuracy as the training progresses with additional epochs for both local and uniform verb-final languages (see Figure 9), as evidenced by the significant interaction between epoch and skewed distribution (*b* = -.046, 95 % CI [-.074, -.018]), and the interaction between epoch and long dependency language type (*b* = -.042, 95 % CI [-.071, -.013]). No significant interactions have been found for verb-initial languages. This suggests that for verb-final languages the learning curves for both local and uniform languages are significantly steeper than long and skewed languages.

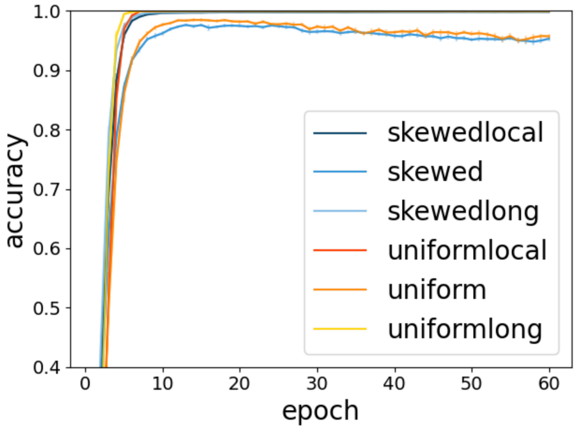


Figure 9: Speaking accuracy as a function of training epoch for verb-final languages.

C. Individual Production Patterns

This appendix includes the production patterns of 40 speaking agents (random seeds) during communication learning for the verb-initial *skewed* language, no dropout (Figure 10). Around half regularize towards SO word order and another half regularize towards OS word order. The production patterns of other seeds and for other settings (verb-final, uniform languages, with dropout) are all similar in terms of regularization strategies.

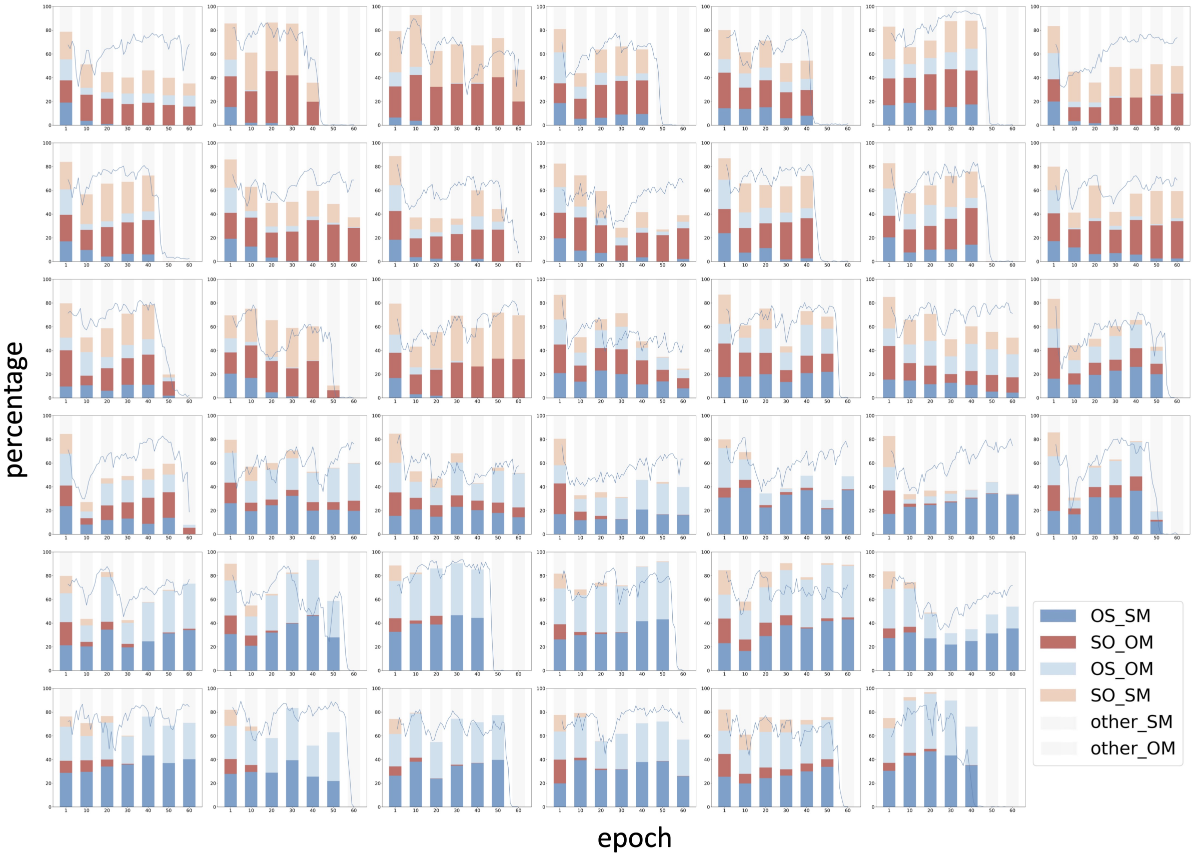


Figure 10: Production patterns of different speaking agents (random seeds) during communication learning for the verb-initial *skewed* language, no dropout. Color denotes word order (blue: OS, red: SO). Shading denotes dependency length (dark: local, light: long). Utterances not belonging to these categories are colored in grey. Meaning reconstruction accuracy per epoch is shown by the blue line. Around half regularize towards SO word order and another half regularize towards OS word order. Subplots are organized manually to highlight groupings of comparable trajectories.