Cogn Comput (2011) 3:539–553
DOI 10.1007/s12559-011-9113-4

BLISS: an Artificial Language for Learnability Studies

Sahar Pirmoradian · Alessandro Treves

Received: 27 April 2011 / Accepted: 11 October 2011 / Published online: 20 October 2011

Abstract To explore neurocognitive mechanisms underlying the human language faculty, cognitive scientists use artificial languages to control more precisely the language learning environment and to study selected aspects of natural languages. Artificial languages applied in cognitive studies are usually designed ad hoc, to only probe a specific hypothesis, and they include a miniature grammar and a very small vocabulary. The aim of the present study is the construction of an artificial language incorporating both syntax and semantics, BLISS. Of intermediate complexity, BLISS mimics natural languages by having a vocabulary, syntax, and some semantics, as defined by a degree of non-syntactic statistical dependence between words. We quantify, using information theoretical measures, dependencies between words in BLISS sentences as well as differences between the distinct models we introduce for semantics. While modeling English syntax in its basic version, BLISS can be easily varied in its internal parametric structure, thus allowing studies of the relative learnability of different parameter sets.

Keywords Artificial Language · Information Theory · Language Acquisition

Introduction

In cognitive science, artificial language learning (ALL) has become a prevailing paradigm for studying language acquisition [1–4]. ALL involves training human adults [5] and sometimes infants [6] on artificial languages with particular structural constrains, and then testing their acquisition of structural information about the miniature languages they were briefly exposed to. The complexity of natural languages makes it difficult to isolate factors critical in language learning; consequently, artificial languages are used, instead, to control more precisely the environment of language learning and study selected aspects of natural languages.

ALL seems to be a valid tool for investigating natural language acquisition and processing. Several studies of event-related potential [7] and neuroimaging [4] have shown similar brain activation in both artificial language and natural language processing.

In addition, similar ALL paradigms have been administered to synthetic agents, such as artificial intelligence expert systems or connectionist networks [8], which could not be possibly confronted with the daunting complexity of a natural language.

Artificial languages applied in both cognitive and computational studies, however, are usually designed on an ad hoc basis, and they typically include so few of the elements of a natural language, e.g., so few words and syntactic categories, as to make generalizations and the test of wide-ranging hypotheses entirely dependent on one's theoretical assumptions. In some studies, toy grammars are used to generate structured sequences without any reference to even the most basic structural patterns of natural languages (the controversial language *universals*) [9, 10]; while in others, which may include basic syntactic categories and phrase structure rules [5, 11], the rules and categories are so few as to raise questions as to how the properties investigated scale up in larger and more complex systems, which, e.g., require the interaction of properties hitherto studied separately.

S. Pirmoradian (✉) · A. Treves
Cognitive Neuroscience Sector, SISSA International School for Advanced Studies, Trieste, Italy
e-mail: pirmorad@sissa.it

The aim of the present study is the construction of an artificial language, called *BLISS*, for Basic Language Incorporating Syntax and Semantics, which mimics natural languages by possessing a vocabulary, syntax, and semantics. Importantly, the degree of complexity of the language is designed having the size limitations of synthetic agents in mind, so as to allow for the use of equivalent corpora with human subjects and with computers, while aiming for reasonable linguistic plausibility.

BLISS is generated by different language models. The term *language model* usually refers, in natural language processing contexts, to a model for predicting the next word given previous words [12]. In this study, the number of degrees of freedom, in this sense, is effectively reduced by using the grammar, which groups words into lexical classes and thus restricts the domain of the next word.

BLISS is generated by a context-free grammar of limited complexity with about 40 production rules, with probabilities that were drawn from the *Wall Street Journal* (WSJ) corpus. It contains function words, inflectional suffixes, and some embedding structure. These grammatical features were introduced to enable experiments to investigate the ability for abstract pattern acquisition [13, 6], the special role of function words [14], the role of suffixes [15], and especially hierarchical structures [16, 17] in humans.

The BLISS vocabulary contains about 150 words, which belong to different lexical categories such as noun, verb, adjective, etc., and which were selected from the *Shakespeare* corpus. There are several studies investigating category learning in humans [18], and BLISS is intended to facilitate, e.g., the analysis of the representation of distinct lexical categories.

Semantics is defined in BLISS as the statistical dependence of each word on other words in the same sentence, as determined solely by imposing constraints on word choice during sentence generation. We have applied different methods of weighing the preceding words so to determine which words come next. This should allow using BLISS to study at least rudimentary aspects of the emergence of semantic categories.

In contrast to small artificial languages that mostly focus on a single issue, BLISS provides us with several features of natural languages, including an interaction between syntax and semantics, function words, some nested structures, and at the same time, BLISS can generate in no time corpora with representative sampling, avoiding at least in principle the problem of limited sampling of linguistic structures.

The structure of this paper is as follows: the next section introduces the language components of BLISS, namely, grammar, lexicon, and semantics. After that, we describe the way we fine tune the syntactic, lexical, and semantics components by deriving statistics from real corpora, thus defining different language models. Then, the methods used to compare the corpora generated by different language models are explained. Finally, we report the results obtained from the comparison between language models.

Basic Architecture of the Artificial Language

The architecture of BLISS is composed of a formal grammar, lexicon, and semantics. A formal grammar is a set of production rules for generating sentences from the vocabulary of the language. The lexicon of a language is its vocabulary, including its words and expressions. Though denoting a range of ideas, the term "semantics", in this study, focuses on the relation between words. In this section, we describe how these three components were designed in BLISS.

Grammar

The grammar assigned to BLISS is a Probabilistic Context-Free Grammar (PCFG). A PCFG is a 5-tuple $\mathsf{G} = (\mathsf{N}, \Sigma, \mathsf{P}, \mathsf{S}, \mathsf{D})$. In this tuple, N is a set of non-terminal symbols that indicate phrasal categories (e.g., verb phrase) and lexical categories (e.g., verb). Σ is a set of terminal symbols which refer to the vocabulary (e.g., *sword*). P is a set of production rules, each of the form $A \rightarrow \beta$ where A is a non-terminal and β is a string of symbols from the infinite set of strings $(\Sigma \cup N)^*$. S is a designated start symbol, and D is a function which assigns probabilities to each rule in P, expressing the probability that the given non-terminal A will be expanded to the sequence β [19].

In the Chomsky hierarchy [20], a context-free grammar, as opposed to a regular grammar, can have center-embedded recursion (of the form $\mathsf{P}(A \rightarrow \alpha A \beta)$, such as relative clauses within a sentence. Context-free languages with center-embedded recursion cannot be generated by a finite state automaton [20]. On the other hand, in the BLISS grammar, for simplicity, the only recursion structure is *that-complement-clause* preceded by embedding verbs such as *think, believe, know*, etc. Obviously, the grammar could easily be extended to have more complex syntactic structure.

The BLISS PCFG in its current implementation (v1.1) consists of 39 non-terminal symbols, 150 terminal symbols, and 40 production rules with their assigned probabilities (See Table 1 for samples and Table 4 in "Appendix" for the full grammar). Non-terminal symbols, represented by <.> , include singular and plural noun phrases (NP1/NP+), verb phrases (VP1/VP+), determiner phrases (DP1/DP+), prepositional phrases (PP), nouns (Noun1/Noun+), verbs (Verb1/Verb+), adjectives (Adj), prepositions

Table 1 Samples of the BLISS PCFG and of its lexicon

No.		Rule		Probability
1	<S>	→	<DP1> <VP1>	0.50
2		→	<DP+> <VP+>	0.50
3	<VP1>	→	<Verb1>	0.85
4		→	<Neg1> <Verb1>	0.15
5	<Verb1>	→	<TransVerb> <DP>	0.37
6		→	<IntransVerb1>	0.41
7		→	<EmbedVerb1> <S>	0.07
8		→	…	…
9	<Noun1>	→	*Sword*	0.13
		→	*Dog*	0.05
		→	…	…
10	<TransVerb1>	→	*Kills*	0.03
		→	*Guards*	0.02
		→	…	…

(Prep), and a complementizer (ThatClz); each of which was subcategorized as listed in Table 2. The procedure of selecting terminal symbols is discussed in section Lexicon. The way we enumerated the production rules is indicated in Table 1.

Further, inflectional suffixes used in the grammar are *s* plural (e.g., *swords*), *s* third person singular present (e.g., *kills*), and *n't* negative (e.g., *don't*).

Lexicon

The BLISS lexicon contains about 150 lexical words, including singular and plural forms. The words were extracted from the Shakespeare corpus, which includes about 1,000,000 word *tokens*, as multiple occurrence of about 27,000 word *types* (vocabulary words). The BLISS lexicon contains different category of words such as nouns, verbs, adjectives (content words), and prepositions, articles, demonstratives, complementizer (function words), as listed in Table 2. The content words (except for *proper* nouns) were selected from a set of high-frequency words in the Shakespeare corpus. The selection was unique, regardless of either singular or plural form. For *common* nouns, after extracting frequent nouns from Shakespeare, we chose among them some which were shared with a database [21], in which a set of feature norms were collected for 541 living and non-living basic-level concepts. Knowing the feature norms of BLISS nouns enables us to derive distributional statistics, when desirable, from these norms, such as their pairwise correlation. Further, BLISS common nouns were chosen so as to be categorized into animates (e.g., dog, horse, etc.), buildings (e.g., church, house, etc.), and objects (e.g., sword, crown, etc.).

As proper nouns, we selected four singular proper nouns (Zarathustra, Ahuramazda, Ahriman, Yasmin) and two plural proper nouns (Magi, Greeks), inspired by Friedrich Nietzsche's *Thus Spake Zarathustra*.

Semantics

In BLISS, semantics is defined as the statistical dependence of each word on other words in the same sentence, purely determined by imposing constraints on word choice during sentence generation. After the grammar chooses a legal lexical category by rewrite rules, the words of the chosen category compete for the appearance in a sentence on the basis of their relative frequency and joint probabilities, which were extracted from the Shakespeare corpus. How we calculate these probabilities will be explained later.

Table 2 Lexical categories with examples

Lexical category	Subcategories	Non-terminal symbols	Words	No. of words
Verbs	Intransitive	<IntransVerb1>	*Comes, dies*, …	8
	Monotransitive	<TransVerb1>	*Loves, guards*, …	20
	Ditransitive	<DitransVerb1>	*Gives, brings*	2
	Embedding	<EmbedVerb1>	*Wishes, believes*, …	7
Nouns	Common	<Noun1>	*Sword, dog*, …	18
	Proper	<PropNoun1>	*Zarathustra, Ahuramazda*, …	4+2 plural
Determiners	Article	<Det1>	*The, a*	2
	Demonstrative	<Dems1>	*This, that*	2
Prepositions	Following nouns	<PP-n>	*Of, in, with, on*	4
	Following ditransitive verbs	<PP-vb>	*To*	1
	Following verbs	<PP-to>	*For, in, with*	3
Complementizers		<ThatClz>	*That*	1
Adjectives		<Adj>	*Great, noble*, …	18

The last column indicates the number of words of the corresponding subcategory that are used in the BLISS grammar

Tuning the Components to Produce Full Language Models

After designing a general BLISS grammar and vocabulary of intermediate complexity, we aim to make it as natural as possible. Therefore, we adjust the transition probabilities of grammar rules, from non-terminals to either non-terminal or terminal symbols, as well as the joint frequency between pairs of words, to real corpora. We choose two corpora, the WSJ and Shakespeare. In this section, the procedure followed for this adjustment is explained. Additionally, we elaborate on our definition of semantics and the difference between the four language models we introduce.

Extraction of Statistics from Real Corpora

In the BLISS grammar, the probabilities of transitions from non-terminal to non-terminal symbols, i.e., between phrasal and lexical categories, were adjusted to the statistics derived from the WSJ corpus[1], a syntactically annotated corpus, with about 1,000,000 word tokens and 38,000 word types. Using tgrep2 software, we derived the frequency of all lexical categories which were used in the BLISS grammar and a probability was assigned to each category according to its derived frequency. Moreover, the transition probability of non-terminal symbols to function words, i.e., prepositions, determiners, and auxiliary verbs, were adjusted to WSJ as well.

On the other hand, the transition probabilities from lexical categories to content words were extracted from the Shakespeare corpus, to give it a less prosaic feel. These probabilities of words are called *prior probabilities* in this paper, i.e., prior to the choice of other words in the same sentence. The same nouns and verbs as well as their probabilities were used for singular and plural categories (for example, see probabilities in the rules starting with <N1> and <*N*+> in Table 4). Proper nouns were set as equiprobable in BLISS (see <PropN1> and <PropN+> in Table 4).

Beside the frequency of single words (1st order statistics), the joint frequency of word-pairs—2nd order statistics—was also extracted from the Shakespeare corpus. The joint frequency $f(w_i, w_j)$ was calculated for each pair of content words (nouns, verbs, adjectives); not for function words and not for Proper Nouns. $f(w_i, w_j)$ is the number of Shakespeare sentences in which the content words (w_i, w_j) orderly appear together consecutively or non-consecutively. Because of the poetic style in Shakespeare, having very long sentences with many short phrases, we counted the occurrence of a word-pair within a window of size 5 in a sentence.

[1] Treebank-3 release of the Penn Treebank project.

Different Semantics Models

After the extraction of the probability of grammar rules, of words, and of the joint frequencies from real corpora, we construct alternative language models based on the different selection algorithms applied for choosing content words. After a legal lexical category is selected by the grammar, the words of the chosen category compete for the appearance in a sentence based on their *selection probability*. The *selection probability* of content word w_n at position n is calculated as a weighted sum of the prior probability of the word (the probability encoded in the grammar G) and its *semantics probability*, which is the conditional probability of the word given the preceding words $w_1 \cdots w_i \cdots w_{n-1}$ in the same sentence (denoted as the history h in Eq.1). The probability function P which is applied as the selection probability of the next word in a sentence is

$$P(w_n|h, G) = (1 - g) * P_{\text{prior}}(w_n|G) + g * P_s(w_n|h), \quad (1)$$

where

$$P_s(w_n|h) = \frac{1}{C}(c_1 P(w_n|w_1) + \cdots + c_{n-1} P(w_n|w_{n-1})), \quad (2)$$

and g is a *semantics coefficient*, which is set to zero when the selection of a word only depends on the grammar, with no dependence on other words in the sentence. Semantics can then be switched on and off by changing the parameter g.

In Eq. 1, P_{prior} and P_s denote prior probability and semantics probability, respectively. The c_i's are *dependence coefficients*, and $C = \sum_{i=1}^{n-1} c_i$; if $c_i \neq 0$, the word w_i is called a *semantically effective word*. The conditional probabilities $P(w_n|w_i)$ between content words were adjusted to the Shakespeare corpus in the manner explained in the previous section (and they are zero when either or both are function words).

Based on different choices of parameters in the semantics probability function (Eq. 2), 4 different language models, 3 with semantics and one without semantics, were defined: the *Exponential*, *Subject–Verb*, *Verb–Subject*, and *No-Semantics* models.

Exponential Language Model

In the Exponential model, all preceding content words in a sentence affect the next word; dependence coefficients in the semantics probability function (Eq. 2) are exponential functions of the distance between preceding words and the current word, $c_i = e^{\lambda(i-n)}$; that is, words at the beginning of a sentence are less effective than the words which are close to the word currently being selected. Thus, the semantics probability function looks like

$$P_s(w_n|h) = \frac{1}{C}\left(e^{\lambda(1-n)}P(w_n|w_1) + \cdots + e^{\lambda(-1)}P(w_n|w_{n-1})\right), \quad (3)$$

where $\lambda>0$ is a *temporal decay coefficient*. As Fig. 1a illustrates, if the sentence *The bloody sword kills* has been generated so far and, according to the grammar, a word from plural noun category (Fig. 1b) must be selected, all words in this lexical category compete to be chosen. The *selection probability* of each word, e.g. crowns, dogs, rocks, …, is measured as the sum of their prior probability (Fig. 1b), and their semantics probability (Eq. 3). In Fig. 1a, as $\lambda = 1/3$, since the distance between *kill* and the new word is 1, the dependence coefficient of the effective word *kill* is $e^{-(1/3)}$. Note that, in the example, *rocks* are likely to have been the instrument rather than the object of several Shakespearian killings. Our semantics models, however, are unfortunately insensitive to such subtle distinctions; hence a certain quirkiness of the resulting sentences.

Subject–Verb Language Model

In the Subject–Verb model, we presume the subject and verb of a sentence to quite strongly influence which word will come next. Therefore, the semantics probability is only affected by the subject and verb of a clause. The dependence coefficients of other words, other than the subject and verb of the clause, are zero:

$$P_s(w_n|h) = \frac{1}{C}\left(c_{\text{subject}}P(w_n|w_{\text{subject}}) + c_{\text{verb}}P(w_n|w_{\text{verb}})\right), \quad (4)$$

where $c_{\text{subject}} + c_{\text{verb}} = 1$. If there is still no verb in a sentence, $c_{\text{subject}} = 1$. Moreover, in case of a *that-complement-clause*, the subject and verb of the specific clause (either the main clause or the complement clause) which the currently selected word belongs to are those considered as effective words. In Fig. 2, only the subject and verb of the sentence *The bloody sword kills*, i.e., *sword* and *kills* respectively, are effective, and in this example $c_{\text{subject}} = 0.2$, and $c_{\text{verb}} = 0.8$. As illustrated, the selection probabilities of the competing words (Fig. 2a) are different from the prior probabilities of these words (Fig. 2b).

When the subject and verb of a sentence are chosen, not only they influence the choice of successive words, but also they change the words previously produced in the same sentence. In the example, the word *bloody* might be another adjective, which replaces it as another word which could appear with the subject *sword* and the verb *kill*.

Verb–Subject Language Model

In the Verb–Subject model, only the subject and the verb affect the semantics probability of coming words, same as in the Subject–Verb model, and with the same probability. However, in the Verb–Subject model, first a verb is chosen and then a noun that is likely to appear with the chosen verb is selected as the subject. Figure 3 illustrates the difference between these two models. As shown in Fig. 3a, first the verb *kills* was chosen, then a subject among the possible nouns, e.g., sword, dog, rock, …; whereas in the Subject–Verb model (Fig. 3b), first the subject was chosen, i.e., *sword*, next the verb. As depicted in this figure, the selection probability of either noun (a) or verb (b) is different from their prior probability in (c) and (d), respectively.

No-Semantics Language Model

In the No-Semantics model, the semantics is switched off, i.e., $g = 0$ in Eq. 1. As Fig. 4a shows, preceding words do not have any effect on the selection probability of currently

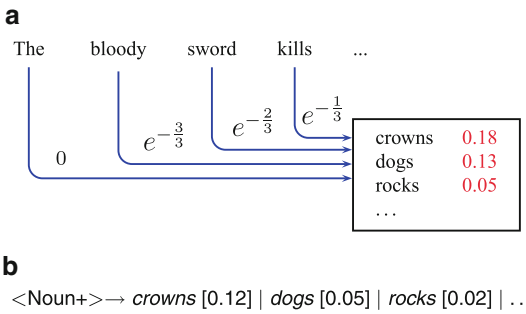

Fig. 1 Example of the Exponential model. **a** The words of plural nouns compete for the appearance in the currently being generated sentence *The bloody sword kills*. Their prior probabilities represented in **b** are modified by the effect of the preceding content words, with an exponential factor, and turned into the probabilities written in the *box*

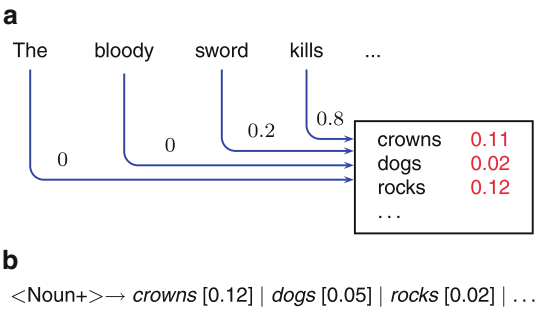

Fig. 2 Example of the Subject–Verb model. **a** The words of plural nouns compete for the appearance in the currently being generated sentence *The bloody sword kills*. Their prior probabilities represented in **b** are influenced by only the subject (*sword*) and the verb (*kills*) of the sentence and are changed to the probabilities written in the *box*. The subject and verb of a sentence constraint the choice of not only the successive words but also the words which were generated already. that is, the word *bloody* might replace another adjective less probable to appear with *sword* and *kills*

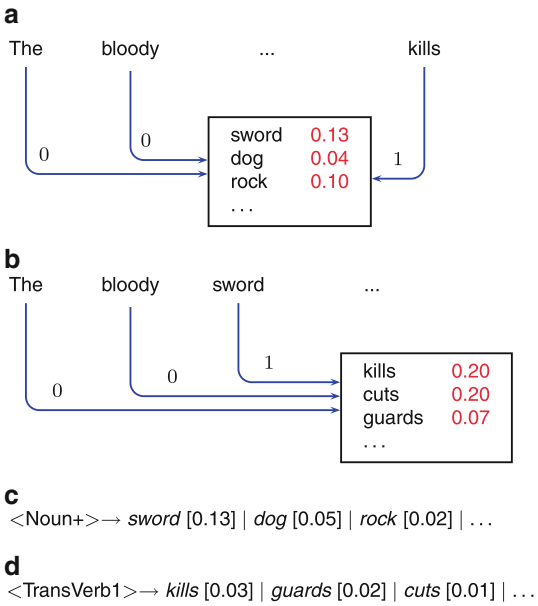

Fig. 3 Example of the Verb–Subject versus. the Subject–Verb model. **a** In the Verb–Subject model, first the verb, i.e., *kills*, is chosen by its prior probability, *0.03* in **d**. Then, the words of singular nouns compete with the probabilities that are influenced by the verb, thus their prior probabilities **c** will change to the probabilities written in the *box*. **b** In the Subject–Verb model, first the subject, i.e., *sword*, is chosen by its prior probability, *0.13* in **c**. Then, the words of singular transitive verbs compete with the probabilities which are influenced by the subject, so their prior probabilities **d** will change to the probabilities written in the *box*. In both of the models, the subject and verb influence the choice of not only the following words but also the preceding words, e.g., *bloody*

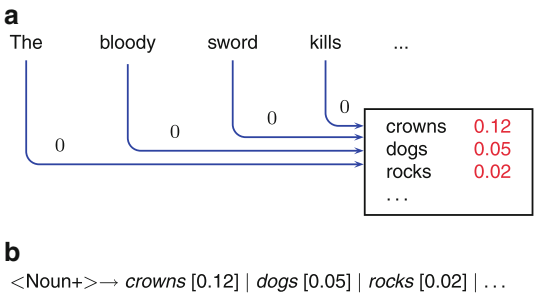

Fig. 4 Example of the No-Semantics model. **a** The words of plural nouns compete by their prior probabilities in **b**, not influenced by the preceding words

selected words, which is the same as the prior probability of these words (Fig. 4b).

Generating Corpora without Grammar

To measure the effect of grammar in BLISS, as a control, we produced corpora without following a grammar, even though adopting, in one variant, the semantics models explained earlier.

Unigram Assuming the entire independence between words of a sentence, we used the so-called Unigram language model, in which the selection probability of a word is its relative frequency in a corpus generated by the No-Semantics model; that is, disregarding the preceding words, h, and the grammar, G, the selection probability of the word w_n is the count of the word, $C(w_n)$, divided by the total number of tokens N in the No-Semantics corpus:

$$P(w_n|h,G) = P(w_n) = \frac{C(w_n)}{N}. \tag{5}$$

Exp-NoSyntax To measure merely the effect of semantics, in the absence of grammar, we defined an Exp-No-Syntax model in which the semantics effect (in Eq. 1) is $g = 1$ and the dependence coefficient is defined by the same function as in the Exponential model (Eq. 3):

$$\begin{aligned} &P(w_n|h,G) \\ &= P_s(w_n|h) \\ &= \frac{1}{C}\left(e^{\lambda(1-n)}P(w_n|w_1) + \cdots + e^{\lambda(-1)}P(w_n|w_{n-1})\right). \end{aligned} \tag{6}$$

A Further Control Model

Equiprobable As a control for the No-Semantics model, we also introduced the Equiprobable model, which is the same as the No-Semantics one, except for the prior probability of words. Words that belong to the same lexical category are equiprobable, not following the probabilities derived from Shakespeare. For instance, the probability of plural demonstratives (<Dem+> in the last row of Table 4) changes from 0.60 and 0.40, for *these* and *those*, respectively, to 0.50 and 0.50.

Comparison Among Model-generated Corpora

To compare corpora generated by the different semantics models, we used methods of Information Theory, with the necessary controls to have reliable measures. The effect of introducing semantics into BLISS was measured as the distance between distributions of word-pairs in corpora generated by the different language models.

Before comparing word-pair distributions, one should eliminate the effect of word frequency differences, so as to measure purely semantics, i.e., word-dependence effects. That is, only corpora which have the same or very close word frequencies can be compared to assess semantics effects, which then result in differences in pair distributions. In this section, first we detail the methods used for measuring semantics effects. Further, we describe how the word frequencies of the semantics models were adjusted to the No-Semantics.

We point out that mutual information and Kullback–Leibler divergence are objective standard measures applicable to any probability distribution, and we refer readers interested in a background in information theory to comprehensive textbooks such as [22] and [23].

Kullback–Leibler Divergence

The Kullback–Leibler (KL) divergence (or distance) is a non-symmetric measure of the difference between two distributions [22]; it measures how many extra bits are required to code a random variable having its own probability distribution P using a code based on a given ("wrong") probability distribution Q:

$$D_{KL}(P||Q) = \sum_{w} P(w) \log_2 \frac{P(w)}{Q(w)}. \quad (7)$$

This measure is applied here to compare the differences between word distributions, as well as word-pair distributions, of corpora generated by the different language models. Although non-symmetric, i.e., $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, differences in the two directions are small, for the distributions we considered, almost always in the third decimal place. We then show the average of these two KL-distance values in the "Result" section, as they all are, anyway, rounded to second decimal precision.

Markov Properties

A stochastic process has the Markov property if the present state predicts future states independent of how the current state is obtained; that is, the process is memory-less. Considering BLISS as a stochastic process, we aim to see whether it has the Markov property, and if, as expected, it does not, then to examine how much it deviates from a Markov process, that is, from a (first-order) Markov chain.

To measure how close BLISS is to a first-order Markov chain, we measure the three-way mutual information [23], defined as:

$$I(X_n; X_{n-2}, X_{n-1}) = I(X_n; X_{n-1}) + I(X_n; X_{n-2}|X_{n-1}), \quad (8)$$

where X_{n-2}, X_{n-1}, and X_n are the values of a random variable in three consecutive states, and $I(X_n; X_{n-2}, X_{n-1})$ is the amount of information the values of this random variable in the first and second states, $n-2$ and $n-1$, convey to the variable in the third state, n.

A random variable has a first-order Markov property if its distribution only depends on the distribution of the previous state, i.e., if in Eq. 8, $I(X_n;X_{n-2} \mid X_{n-1}) = 0$. We measure this quantity in the corpora generated by our models, where X_i is the random variable for position i in the sentence and takes as value the words of the vocabulary. From now on, we use $I(i;i-1)$ instead of $I(X_i; X_{i-1})$, for simplicity, when referring to a mutual information value relative to a specific position i, and $I(n;n-1)$ when referring to its average across positions.

Adjusting Word Frequencies

As explained at the beginning of this section, we need to eliminate the effect of first-order statistics, i.e., word frequencies, thus producing corpora with close word frequencies; that is, corpora must be generated with pre-assigned posterior word frequencies. This goal can be achieved if the prior word frequencies used by each language model are controlled, that is, adjusted.

We analyzed the relation between prior probability of a word w_j and its posterior probability after it is observed having a specific role r (i.e. subject, verb, …) in sentences generated by BLISS language models (See Eq. 10 in "Appendix" for details):

$$P_{\text{prior}}(w_j) = \frac{P_{\text{post}}(w_j|r) - g\sum_{h_e} P(h_e)P_s(w_j|h_e)}{\sum_{h_{ie}} P(h_{ie}) + (1-g)\sum_{h_e} P(h_e)}, \quad (9)$$

where all possible configurations of preceding words (history h) are considered as either *effective* history (h_e) or *ineffective* history (h_{ie}). A history is called effective if $P_s(w_j \mid h) \neq 0$, and is ineffective if $P_s(w_j \mid h) = 0$. Knowing $\Sigma_{h_e} P(h_{ie}) + \Sigma_{h_e} P(h_e) = 1$, we see in this equation that $P_{prior}(w_j) = P_{post}(w_j)$, if $g = 0$, as it is in the case of the No-Semantics model.

Using Eq. 9, prior probabilities of content words in the semantics models (the Subject–Verb, Verb–Subject, and Exponential) were calculated to yield the same posterior probabilities as those in the No-Semantics model. In this fashion, we generated corpora, though by the different language models, with very close word frequencies. The distances between word distributions were measured using KL-divergences (see "Results").

Results

Applying the methods described above, we have constructed three language models with semantics, as well as one language model with syntax but without semantics (No-Semantics) and one with semantics but without syntax (Exp-NoSyntax), to serve as controls, together with two straightforward controls, one with neither syntax nor semantics (Unigram) and one with syntax but no individual word frequency tuning (Equiprobable). See Table 3 for examples of sentences in the Subject–Verb model.

To quantitatively assess the dependencies between words thus introduced, we have measured KL-distances between the word-pair distributions of these models; the

Table 3 Samples of sentences generated by the Subject-Verb model

The church stands
The precious crown guards the royal sword
Crowns give Zarathustra to the sword
The sweet horse knows that a church believes that Ahriman dies
A noble sword fights for Ahuramazda
Zarathustra loves a holy church
Rocks of bloody doors cut Yasmin
That noble dagger keeps the house
Horses don't go
Foul gates enter
The dogs don't fight
A gracious dog doesn't come with the sheep
A house of Zarathustra thinks that Ahriman keeps the doors
A royal house sits
Gates with a strong sword praise the horse
The swords prove that Magi hold the wall
The sword doesn't die
A gate thinks that Ahriman guards a house in the door
Strong rocks don't sit
The dog doesn't wish that the gates in the holy house keep dogs

results are discussed in the first part of this section. In the second part, we are concerned with individual sentences, to see how much information is conveyed by the words in a sentence. Finally, we are interested in investigating the memory characteristics of BLISS.

The length of sentences and the size needed of corpora for the measures used in this section is discussed in "Appendix".

Distance Between Models: A Mild Effect of Semantics

In order to compare word-pair distributions among models, first we need to make sure that individual word frequencies are almost equal in all the main models (No-Semantics, Subject–Verb, Verb–Subject, and Exponential). This is because comparing second-order statistics (word-pair distribution) is meaningless unless the first-order statistics (individual word frequencies) are equal.

In Fig. 5a, the KL-distances between individual word frequencies are shown, after adjusting the word frequencies for the semantics models to those in the No-Semantics one. The vertices of the pyramids indicate the language models. The number in (and, approximately, the size of) the circles on the edge between each two vertices indicates the KL-distance between the word distributions of corpora generated by the corresponding models. As shown, the distances between the No-Semantics and the three semantics models are zero, to 2nd decimal precision. As a control, the No-Semantics model is at a considerable distance from the Equiprobable model, and, likewise, the Exponential model is at a large distance from the equivalent exponential model without syntax (Exp-Nosyntax). The Exp-Nosyntax model is at about the same distance, in fact, from the other two semantics models (not shown).

Given the same word frequencies across the main language models, we expect to see different word dependencies between the models, especially between the No-Semantics model and the semantics ones due to the semantics effect. We use the KL-divergence between word-pair distributions of every two models to quantify the difference in word dependencies. Therefore, the greater the KL-divergence between two word-pair distributions, the larger the difference in word dependencies between the two models.

In Fig. 5b, adopting the same pyramidal graphic scheme as in Fig. 5a, the KL-distances are shown between word-pair distributions among the same models. As illustrated, the No-Semantics model is at a considerable distance from

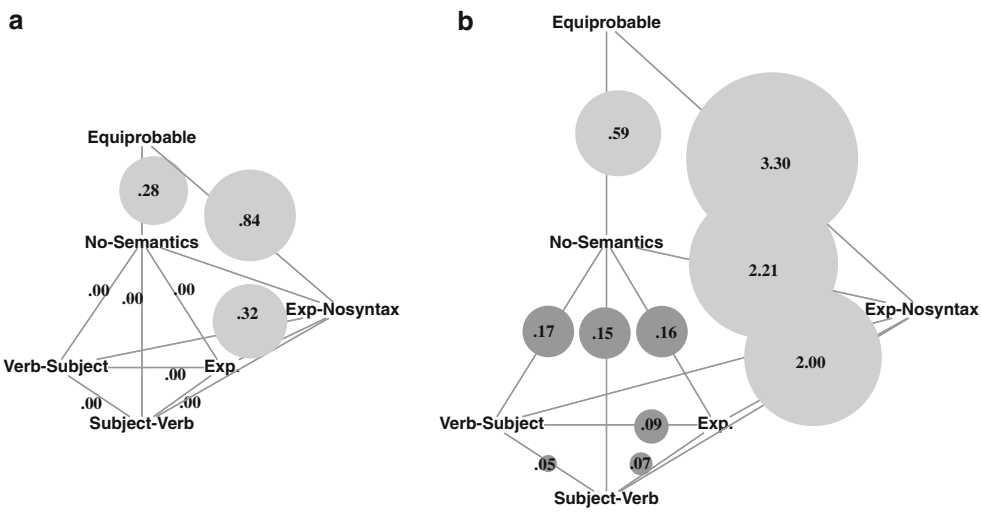

Fig. 5 KL-distances between individual word distributions **a** and word-pair distributions **b**. The *vertices* of these *pyramids* indicate the distributions of the No-Semantics, Subject–Verb, Verb–Subject and Exponential main language models, as well as the Equiprobable and Exp-Nosyntax control models. The *number* and the size of the *circles* on the edge between each two vertices indicate the KL-distance between the distributions of the corpora generated by the corresponding models

each of the semantics models, larger than the distance between each pair of these semantics models. This result quantifies semantics effects in these models. Note that the effects of syntax on word-pair distributions are much larger, as shown by the distances with the control models without syntax (the Exp-Nosyntax). Considering the Equiprobable and the No-Semantics models demonstrates the necessity for adjusting word frequencies across models. These are in fact the same model (without a semantics effect) except for their word frequencies. Hence, the distance between pairwise statistics of these two models (Fig. 5b) merely reflects the (large) effect of unequal word frequencies, and such effect has to be removed to probe the more interesting distance arising, in other comparisons, from unequal word dependencies.

Figure 6 shows that the distance between the semantics models and the No-Semantics one gets larger monotonically as we increase the parameter g which controls semantics effects (Eq. 2). g could theoretically vary between 0 (when the semantics is switched off) and 1 (when the selection probability of a word is only influenced by the preceding history, not by its prior probability); however, as illustrated, the distances are measured up to $g = 0.9$, the point where the corpora could still be produced with the same word frequency, using Eq. 9.

Mutual Information Between Words: Again a Mild Effect of Semantics

To find out how much information the words in a sentence convey to the words in the same sentence, that is, to what extent successive word choices are mutually constrained, we looked at different measures of mutual information (see Fig. 7).

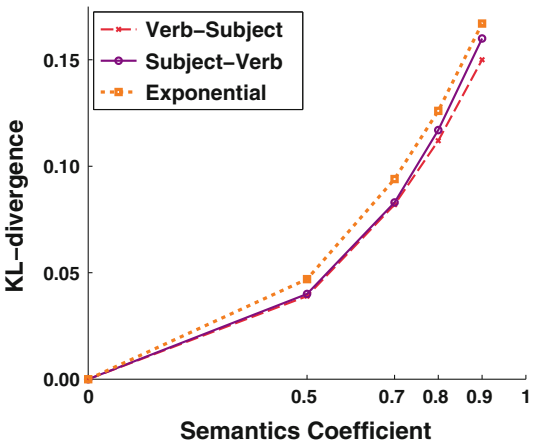

Fig. 6 KL-divergence between the semantics models and the No-Semantics model, when the semantics parameter g changes. Note that g can only reach up to $g = 0.9$, before Eq. 9 becomes inapplicable for adjusting word frequency

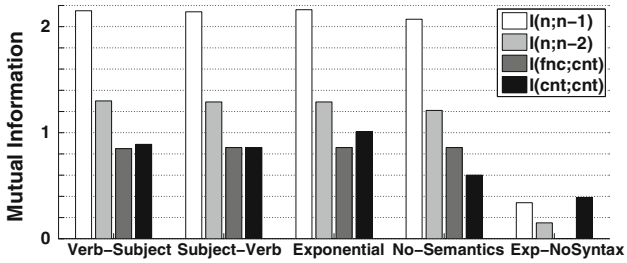

Fig. 7 The mutual information words convey to words in the same sentence. $I(n;n-1)$, in *white*, is the average mutual information which words in position $n-1$ of a sentence convey to their nearest neighbor (in position n); the average is over all positions n within sentences. $I(n;n-2)$ (in *light gray*) is the mutual information between words which are next- nearest neighbors. The *darker gray bars*, I(fnc;cnt), represent the amount of information which function words (e.g., prepositions) convey to the nearest content word (e.g., a noun) in the same sentence. I(cnt;cnt), in *black*, represents the amount of information that content words convey to the nearest content word in the same sentence. We see considerably higher values for the semantic models than for the No-Semantics one, for this last measure. Note that the mutual information between content words due to syntax and to semantics in the Exp model happens to be nearly exactly the sum of that due to syntax alone (in the No-Semantics model) with that due to Semantics alone (in the Exp-NoSyntax model)

$I(n;n-1)$ is the mutual information which words in position $n-1$ of a sentence convey to their nearest neighbor (in position n); the average is over all positions in the sentence. As shown, this measure is almost the same for the semantics models (the absolute value of their difference is about 0.01), somewhat lower in the No-Semantics model (about 4% less), and considerably lower in Exp-Nosyntax one (85% less), thus pointing at the strong effect of syntax. Likewise, we see a similar pattern across models for $I(n;n-2)$ (in light gray) which is the mutual information between words which are next-nearest neighbors. The semantics models show again the same statistics (with at most a 0.01 difference in their absolute values), the No-Semantics is about 7% lower, and the Exp-Nosyntax model significantly decreases, by 88%.

Further, in Fig. 7, the darker gray bars, I(fnc;cnt), represent the amount of information which function words (e.g., prepositions) convey to the nearest content word (e.g., a noun) in the same sentence. As shown, all the semantics models show the same I(fnc;cnt) as the No-Semantics one, because there is no semantically induced component of the joint probability between function and content words, which reflects only syntax. In the Exp-Nosyntax model, this measure is zero, because there is neither semantics, to constrain these pairs of words (there is only between content words), nor syntax, in this particular model.

The main result is seen looking at I(cnt;cnt), in black, which represents the amount of information that content

words convey to the nearest content word in the same sentence. We see considerably higher values for the semantic models than for the No-Semantics one. The Exponential model, in which all preceding words contribute, shows the heaviest dependence of word choice on previous history. More precisely, the Verb–Subject, Subject–verb, and the Exponential convey about 48%, 43%, and 68%, respectively, more information than the No-Semantics model. Thus introducing semantics, in our BLISS models, further constrains by an additional 40–70% the mutual dependence between content words, without having other appreciable statistical effects. One should note that most of the dependence between content words, however, is already imposed by the syntax.

Memory Characteristics

To see how much our language models deviate from a first-order Markov model, we calculated the triple mutual information for 7-word sentences taking the Subject–Verb model as an example (Fig. 8). The other models show very similar memory characteristics, the semantics models with the very same numbers (the absolute value of the difference hovering around 0.01) and the No-Semantics with slightly less information (by about 0.09). The random variables, in this analysis, are the words appearing in specific positions of a sentence. Each bar, in the Figure, indicates the amount of entropy of the individual random variables, $H(n)$. Its black-colored portion shows the amount of conditional mutual information, $I(n;n-2|n-1)$, whereas the gray-colored portion shows the amount of mutual information, $I(n;n-1)$ (the sum of the two is just the triple mutual information, $I(n;n-2,n-1)$). In a first-order Markov sequence, the conditional information would be zero, so the non-zero height of the black bars quantifies the non-Markovian character of BLISS.

Figure 9 shows the dependence of the word at each position on those at each of the preceding positions, by calculating the pairwise mutual information between words at different positions within each sentence. Each shade represents the amount of information conveyed by words in a particular position to the words in the following positions. For example, the (bottom) black portion of the bar in position i represents the information conveyed by words in position 1, namely $I(i;1)$. Memory effects are seen to "decay" with a time constant, so to speak, of 1–2 words, although an initial rapid drop in the degree of dependence is followed by a more sustained level, due also to the fact that the subject and verb, the important factors in this model, rarely occur in the very first and in the few first positions, respectively. Again, results are very similar for the Verb–Subject and Exponential models (not shown).

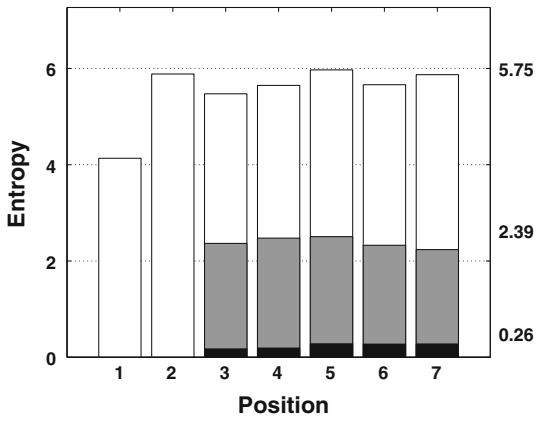

Fig. 8 Markov properties of the Subject–Verb model. The full height of each *bar* indicates the amount of entropy of individual words at each position in the sentence, $H(i)$. The *gray portion* of each *bar* shows the amount of mutual information, $I(i;i-1)$, whereas the *black portion* shows the amount of conditional mutual information, $I(i;i-2|i-1)$. Note that the sum of the *black* and *gray portions* equals the triple mutual information, $I(i;i-1,i-2)$. In a first-order Markov sequence, the conditional information (in *black*) is zero, whereas in BLISS, this measure shows some deviation from a first-order Markov model. The *numbers* on the *right vertical* side of the *box* indicate the average of each measure over positions 2:7 for $H(n)$=5.75, and over positions 3:7 for $I(n;n-1,n-2)$=2.39 and $I(n;n-2|n-1)$=0.26

Discussion

We have constructed an artificial language, BLISS, which can be used in language learning experiments. BLISS is based on the putative universal principles of natural languages: vocabulary, syntax, and semantics. To make it as natural as possible, within its limited size, the probability of grammar rules, words, and word-pairs were extracted from natural language corpora, the Wall Street Journal and the Shakespeare corpora.

The BLISS grammar is a Probabilistic Context-Free Grammar, a set of constraints which has been argued in the literature to be sufficient to represent *most* of the syntactic structure in natural languages, such as English. It also includes several phrasal and lexical categories that exist in natural languages. There are both function words (e.g., prepositions, articles, …) and content words (e.g., nouns, verbs, …) in the grammar. The relation between function and content words have been of interest to cognitive neuroscientists [14]. The grammar also contains a recursive construction, which might be used in experiments to investigate the effect of hierarchical structures in language processing [16, 17].

BLISS contains semantics that could be switched on and off, if the interaction between syntax and semantics is investigated in an experiment. Semantics is taken here to be the statistical dependence between words in the same sentence, produced by imposing additional, non-syntactic

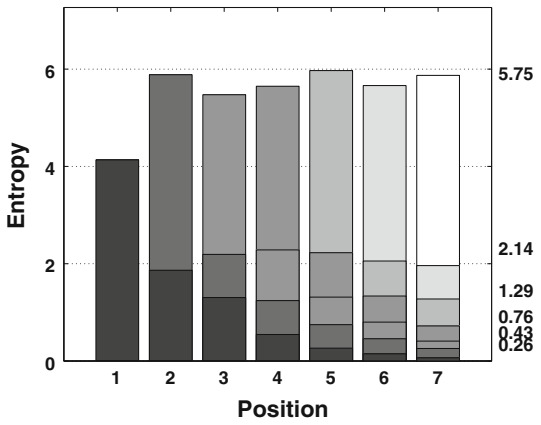

Fig. 9 Memory effects in the Subject–Verb model. The dependence of each word on words at preceding positions is quantified by calculating pairwise mutual information values between positions in the sentences. Each *shade* represents the amount of information conveyed by words in a particular position to the words in following positions. For example, *black*,in the *bar* at position i, represents the information conveyed by words in position 1, namely $I(i;1)$. The *numbers* on the *right* of the *box* indicate the average values $I(n;n-i)$ over positions i+2:7: e.g. $H(n)$=5.75, $I(n;n-1)$=2.14 , $I(n;n-2)$= 1.29, and so on

constraints on word choice during sentence generation. We have defined different models for semantics: Subject–Verb, Verb–Subject, and Exponential. In the Subject–Verb and Verb–Subject models, the importance of the thematic role was considered, as studied in several language learning experiments [24, 25].

To measure the purely distributional effect of semantics, we have used information theory, which was extensively used in natural language processing studies, from document retrieval [26] to the evolutionary basis of natural languages [27]. We find that with this limited vocabulary, the effects of semantics are small, relative to those produced by syntax.

We emphasize that in order to quantify the difference between models in terms of pairwise statistics (using either KL-divergence or mutual information), we need to have equal first-order statistics across the models. As discussed in "Appendix", this prevents us from setting the semantics coefficient at $g = 1$; in other words, we cannot increase arbitrarily the effect of semantics, by acting on that parameter, and still be able to compare quantitatively with a language model without semantics, and have all models sound natural in terms of word frequency. The only way we can see to make semantics more important, in a linguistically plausible way, is to increase the vocabulary beyond the "intermediate" size which was our BLISS target.

Grammar and semantics concur in defining the memory properties of the language. Although they cannot be characterized as first-order Markov chains, the different variants of BLISS come relatively close to that, as suggested by the fact that the mutual information between two next-nearest neighbor words, conditional to the intermediate word, is very small, and that non-conditional mutual information between more distant words decays close to exponentially. These results are shown in Figs. 8 and 9 for the Subject–Verb model, but they apply almost identically to the other semantics models. It is an interesting question for future work to understand whether such rapid decay is characteristic of artificial languages modeled after English, and whether natural languages with more articulated syntax would lead to models with longer-stretching memory. We plan to address this question within the quantitative BLISS framework, by introducing syntactic variants modeled after other natural languages.

Beside grammar and semantics, another main component of natural languages is of course phonology (the structure of sounds), which is not included in the current implementation of BLISS. There are several artificial language learning studies [28–30] which investigate the importance of phonological cues in language acquisition. For instance, Monaghan et al. [31] investigate the interaction between phonological and distributional cues in learning lexical categories. A possible variant implementation of BLISS may include replacing real words by non-words which are controlled for phonological cues while, at the same time, keeping the same distributional cues, or joint probability of words.

Using parallel programming libraries and objected-oriented programming in python, BLISS software can generate millions of sentences in a few minutes. It provides a large sample of sentences which not only mimic natural languages but are also produced under the control of the experimenter.

BLISS can be used as a training sample for language acquisition by synthetic systems, including neural networks. Also, it can be used in experiments with adults when complex sentences are needed; in addition, with humans, its words can be replaced by pseudo-words, or by visual or haptic signs. One may also modulate the complexity of a sentence, which is determined not only by the grammatical structure but also by its semantics, for example, in order to study the processing of semantically reversible and irreversible sentences [32].

An important feature of BLISS design, not elaborated here, is that it can be easily varied in its internal parametric structure. In its current version, again, BLISS is syntactically modeled after English, irrespective of the specific semantic model used. We can however keep the semantics stable, and alter the production rules, either solely in the ordering of specific elements or also in their type and relative probability, to model other natural languages. Large corpora may then be produced that

maintain a controlled and quasi-naturalistic flavor, while enabling the study of the relative learnability of distinct parameter settings [33]. Long-term, this may allow a scaled-down artificial language counterpart to the grand program of studying the evolution of language diversity [34].

Table 4 The full BLISS PCFG (probabilistic context-free grammar)

<S1>	→ <DP1> <VP1> [0.50] | <DP+> <VP+> [0.50]
<VP1>	→ <VR1> [0.85] | <Neg1> <VR+> [0.15]
<VP+>	→ <VR+> [0.85] | <Neg+> <VR+> [0.15]
<VR1>	→ <Vt1> <DP> [0.37] | <Vtdtv1> <DP> <PPT> [0.06] | <Vtd1> <SRd> [0.07] | <Vi1> [0.41] | <Vi1> <PPV> [0.09]
<VR+>	→ <Vt+> <DP> [0.36] | <Vtdtv+> <DP> <PPT> [0.09] | <Vtd+> <SRd> [0.05] | <Vi+> [0.35] | <Vi+> <PPV> [0.15]
<PP>	→ <Prep> <DP> [1.00]
<PPV>	→ <PrepV> <DP> [1.00]
<PPT>	→ <PrepT> <DP> [1.00]
<SRd>	→ <Conjd> <S1> [1.00]
<DP>	→ <DP1> [0.80] | <DP+> [0.20]
<DP1>	→ <Det1> <NP1> [0.60] | <PropN1> [0.40]
<DP+>	→ <Det+> <NP+> [0.20] | <NP+> [0.77] | <PropN+> [0.03]
<NP1>	→ <N1> [0.60] | <AdjP> <N1> [0.20] | <N1> <PP> [0.20]
<NP+>	→ <N+> [0.60] | <AdjP> <N+> [0.20] | <N+> <PP> [0.20]
<Det1>	→ <Art1> [0.97] | <Dem1> [0.03]
<Det+>	→ <Art+> [0.98] | <Dem+> [0.02]
<N1>	→ sword [0.13] | dagger [0.02] | crown [0.13] | dog [0.05] | horse [0.09] | dove [0.01] | calf [0.02] | deer [0.02] | worm [0.03] | sheep [0.03] | stone [0.04] | pearl [0.01] | wall [0.05] | gate [0.06] | house [0.18] | door [0.06] | rock [0.03] | church [0.04]
<N+>	→ swords [0.13] | daggers [0.02] | crowns [0.13] | dogs [0.05] | horses [0.09] | doves [0.01] | calves [0.02] | deer [0.02] | worms [0.03] | sheep [0.03] | stones [0.04] | pearls [0.01] | walls [0.05] | gates [0.06] | houses [0.18] | doors [0.06] | rocks [0.03] | churches [0.04]
<PropN1>	→ Zarathustra [0.25] | AhuraMazda [0.25] | Yasmin [0.25] | Ahriman [0.25]
<PropN+>	→ Magi [0.50] | Greeks [0.50]
<Vi1>	→ comes [0.33] | fights [0.07] | goes [0.21] | dies [0.07] | enters [0.08] | stands [0.14] | sits [0.07] | dares [0.03]
<Vi+>	→ come [0.33] | fight [0.07] | go [0.21] | die [0.07] | enter [0.08] | stand [0.14] | sit [0.07] | dare [0.03]
<Vt1>	→ praises [0.02] | loves [0.23] | follows [0.04] | cuts [0.02] | kills [0.04] | hates [0.02] | trusts [0.02] | serves [0.03] | holds [0.07] | hangs [0.03] | guards [0.02] | leaves [0.06] | needs [0.02] | wears [0.05] | marries [0.01] | gets [0.04] | finds [0.05] | takes [0.13] | banishes [0.01] | keeps [0.09]
<Vt+>	→ praise [0.02] | love [0.23] | follow [0.04] | cut [0.02] | kill [0.04] | hate [0.02] | trust [0.02] | serve [0.03] | hold [0.07] | hang [0.03] | guard [0.02] | leave [0.06] | need [0.02] | wear [0.05] | marry [0.01] | get [0.04] | find [0.05] | take [0.13] | banish [0.01] | keep [0.09]
<Vtd1>	→ wishes [0.12] | believes [0.03] | doubts [0.05] | hopes [0.12] | proves [0.12] | knows [0.32] | thinks [0.24]
<Vtd+>	→ wish [0.12] | believe [0.03] | doubt [0.05] | hope [0.12] | prove [0.12] | know [0.32] | think [0.24]
<Vtdtv1>	→ gives [0.80] | brings [0.20]
<Vtdtv+>	→ give [0.80] | bring [0.20]
<Conjd>	→ that [1.00]
<Prep>	→ of [0.60] | in [0.25] | with [0.07] | on [0.08]
<PrepT>	→ to [1.00]
<PrepV>	→ in [0.40] | with [0.30] | for [0.30]
<Neg1>	→ doesn't [1.00]
<Neg+>	→ don't [1.00]
<Art1>	→ the [0.70] | a [0.30]
<Art+>	→ the [1.00]
<AdjP>	→ great [0.16] | sweet [0.12] | noble[0.10] | poor [0.09] | long [0.08] | foul [0.04] | gentle [0.07] | bloody [0.04] | worthy [0.04] | strong [0.04] | holy [0.04] | heavy [0.03] | gracious [0.03] | royal [0.04] | mighty [0.02] | dangerous [0.02] | foolish [0.02] | precious [0.02]
<Dem1>	→ this [0.42] | that [0.58]
<Dem+>	→ these [0.60] | those [0.40]

Conclusion

We have constructed an artificial language of limited complexity, BLISS mimics natural languages by possessing a grammar of about 40 production rules, a vocabulary size of 150 words, and semantics defined by imposing dependencies between words. The effect of introducing semantics with such a limited vocabulary was quantified using methods of information theory and found to be small, but still measurable.

The code for generating BLISS is freely available from the authors. It is hoped that it will be used in a variety of ALL studies.

Acknowledgments We are grateful to Mayur Nikam and Mohammed Katran, who helped develop early versions of BLISS, and to Giuseppe Longobardi for linguistic advice and encouragement.

Appendix

Full Grammar of BLISS

All the grammar rules and the lexicon of BLISS are shown in Table 4.

Proof of Eq. 9

$$
\begin{aligned}
P_{\text{post}} &= \sum_h P(h)P(w_j|h) \\
&= \sum_h P(h)\big((1-g)P_{\text{prior}}(w_j) + gP_s(w_j|h)\big) \\
&= \sum_{h_{ie}} P(h_{ie})P_{\text{prior}}(w_j) \\
&+ \sum_{h_e} P(h_e)\big((1-g)P_{\text{prior}}(w_j) + gP_s(w_j|h_e)\big) \qquad (10) \\
&= \left(\sum_{h_{ie}} P(h_{ie}) + (1-g)\sum_{h_e} P(h_e)\right)P_{\text{prior}}(w_j) \\
&+ g\sum_{h_{ie}} P(h_{ie})P_s(w_j|h_{ie})
\end{aligned}
$$

After analyzing the relation between prior and posterior probabilities of a word in our models (Eq. 9), we have generated corpora with the desired overall word frequencies (the word frequencies of the semantics models were adjusted to the ones of the No-Semantics). Note, however, that there are some constraints regarding the possibility of generating sentences with arbitrary word frequencies. In Eq. 9, the numerator cannot be negative, because the output is a probability measure. Therefore, the possible posterior probability of a word is constrained by the parameter g and pairwise statistics in $P_s(w_j \mid h_e)$, which here has been derived from Shakespeare.

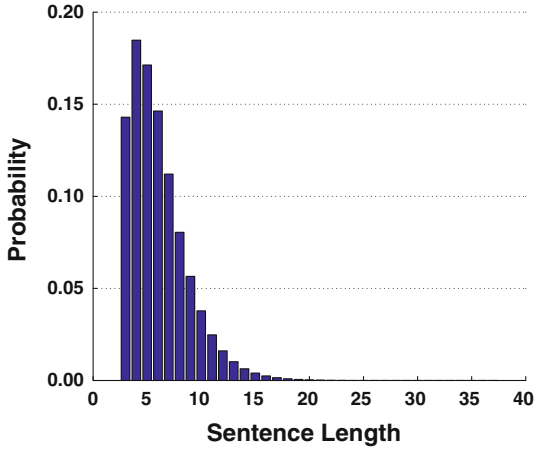

Fig. 10 Probability distribution of sentence lengths in a corpus produced by the Subject–Verb model. Other models with grammar show similar distributions

Length of BLISS Sentences

The probability distribution of the length of sentences generated by the BLISS grammar is shown in Fig. 10. To calculate mutual information values between words appearing in different positions of a sentence, we need to work with sentences of the same length. Considering the fact that the average length of the sentences is about 5 and also to obtain at least 5 distinct measures of triple mutual information values, we chose length 7. Thus, in this paper, for the results involving mutual information and triple mutual information, we used sentences of at least 7 words.

Size of Model-generated Corpora

A technical question to be addressed is how many sentences are needed, to generate sufficient statistics. To answer this question, we produced up to 40 million sentences, and calculated several of the measures used in this paper.

In Fig. 11, the mutual information between words in neighboring positions in a sentence, averaged over positions, $I(n;n-1)$, was measured for each model and with corpora of different size: 1, 10, 20, and 40 million sentences of at least 7 words. As shown, having 10 million sentences is enough for capturing pairwise statistics in the corpora, regardless of having syntax or semantics.

Figure 12 shows the result for the average triple mutual information among words, $I(n;n-2,n-1)$, which needs the largest samples among all measures we have applied. As illustrated, increasing the size of the corpus from 20 to 40 million sentences does not appreciably change results for the models with syntax (Subject–Verb, Verb–Subject, Exponential, and No-Semantics), while we see some

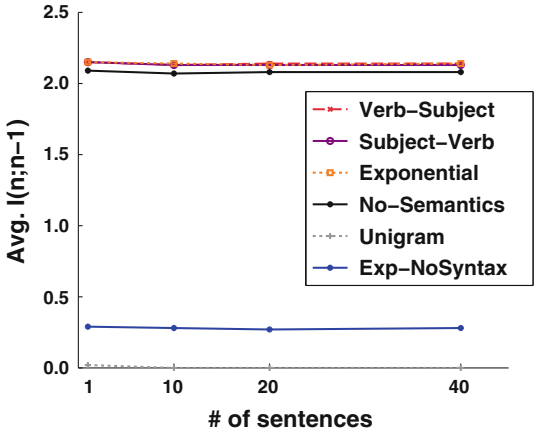

Fig. 11 Average mutual information $I(n;n-1)$, also averaged over all positions, versus number of sentences in the corpus. The mutual information between words is measured for each model and with corpora of different size: 1, 10, 20, and 40 million sentences of at least 7 words. As shown, 10 million sentences are enough for capturing pairwise statistics in the corpora, regardless of syntax or semantics

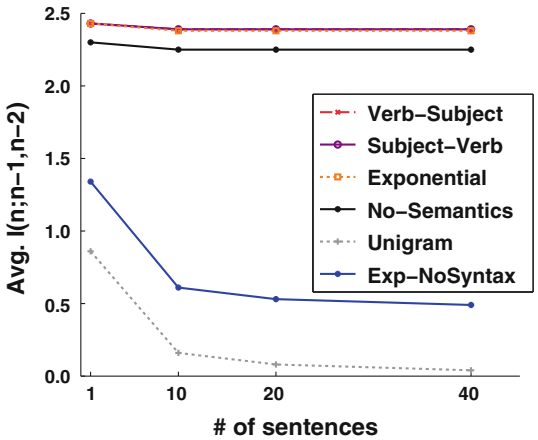

Fig. 12 Average triple mutual information $I(n;n-1,n-2)$, again averaged also across positions, versus number of sentences in the corpus. The three-way mutual information needs the largest sample among all measures we applied but, as illustrated, increasing the size from 20 to 40 million sentences does not change it considerably, for the models with syntax (Subject–Verb, Verb–Subject, Exponential, and No-Semantics). We see still some change for the models without syntax, which obviously need larger corpora, due to their larger word-to-word variability

changes for the models without syntax, which need larger samples.

In sum, corpora with 20 million sentences, of at least 7 words each, were used for the calculation of the mutual information and triple mutual information in this study.

For the KL-divergence results, we also need sufficient word-pair statistics, like for mutual information values. Hence, we have used corpora of at least 20 million sentences for the KL-divergence calculation as well, but without the 7-word constraint.

References

1. Saffran JR, Aslin RN, Newport EL. Statistical learning by 8-month-old infants. Science. 1996;274(5294):1926–28.
2. Christiansen MH. Using artificial language learning to study language evolution: exploring the emergence of word order universals. In: The evolution of language: 3rd international conference; 2000. pp. 45–8.
3. Pena M, Bonatti LL, Nespor M, Mehler J. Signal-driven computations in speech processing. Science. 2002;298(5593):604–7.
4. Petersson KM, Folia V, Hagoort P. What artificial grammar learning reveals about the neurobiology of syntax. Brain Lang. 2010;Available from: http://dx.doi.org/10.1016/j.bandl.2010.08.003
5. Friederici AD, Steinhauer K, Pfeifer E. Brain signatures of artificial language processing: evidence challenging the critical period hypothesis. P Natl Acad Sci USA. 2002;99(1):529–34.
6. Gomez R. Infant artificial language learning and language acquisition. Trends Cogn Sci. 2000;4(5):178–86.
7. Mueller JL, Oberecker R, Friederici AD. Syntactic learning by mere exposure—an ERP study in adult learners. BMC neurosci. 2009;10(1):89.
8. Kinder A, Lotz A. Connectionist models of artificial grammar learning: what type of knowledge is acquired?. Psychol Res. 2009;73(5):659–73.
9. Reber AS. Implicit learning of artificial grammars. J Verb Learn Verb Behav. 1967;6:855–63.
10. Knowlton BJ, Squire LR. Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. J Exp Psychol Learn Mem Cogn. 1996;22:169–81.
11. Opitz B, Friederici AD. Interactions of the hippocampal system and the prefrontal cortex in learning language-like rules. Neuroimage. 2003;19:1730–37.
12. Manning CD, Schuetze H. Foundations of statistical natural language processing. 1st ed. Cambridge: The MIT Press; 1999
13. Marcus GF, Vijayan S, Bandi Rao S, Vishton PM. Rule learning by seven-month-old infants. Science. 1999;283(5398):77.
14. Hochmann JR, Endress AD, Mehler J. Word frequency as a cue for identifying function words in infancy. Cognition. 2010;115(3):444–57.
15. Ullman MT, Pancheva R, Love T, Yee E, Swinney D, Hickok G. Neural correlates of lexicon and grammar: evidence from the production, reading, and judgment of inflection in aphasia. Brain Lang. 2005;93(2):185–238.
16. de Diego-Balaguer R, Fuentemilla L, Rodriguez-Fornells A. Brain dynamics sustaining rapid rule extraction from speech. J Cogn Neurosci. 2011;23(10):3105–20.
17. Bahlmann J, Schubotz RI, Friederici AD. Hierarchical artificial grammar processing engages Broca's area. Neuroimage. 2008;42:525–34.
18. Lany J, Saffran JR. From statistics to meaning. Psychol Sci. 2010;21(2):284–91.
19. Jurafsky D, Martin JH. Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition (Prentice Hall Series in Artificial Intelligence). 1st ed. Prentice Hall; 2000.
20. Chomsky N. On certain formal properties of grammars. Inform Control. 1959;2(2):137–67.
21. McRae K, Cree GS, Seidenberg MS, Mcnorgan C. Semantic feature production norms for a large set of living and nonliving things. Behav Res Methods. 2005;37(4):547–59.
22. Cover TM, Thomas JA, Wiley J, et al. Elements of information theory. vol. 306. New Jersey: Wiley; 1991.
23. MacKay DJC. Information theory, inference, and learning algorithms. Cambridge: Cambridge University Press; 2003.

24. Altmann G, Kamide Y. Incremental interpretation at verbs: restricting the domain of subsequent reference. Cognition. 1999; 73(3):247–64.
25. Bicknell K, Elman JL, Hare M, McRae K, Kutas M. Effects of event knowledge in processing verbal arguments. J Mem Lang. 2010;63(4):489–505.
26. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. 1st ed. Cambridge: Cambridge University Press; 2008.
27. Piantadosi ST, Tily H, Gibson E. Word lengths are optimized for efficient communication. P Natl Acad Sci USA. 2011 Mar; 108(9):3526–29.
28. Monaghan P, Chater N, Christiansen MH. The differential role of phonological and distributional cues in grammatical categorisation. Cognition. 2005;96(2):143–82.
29. Toro JM, Nespor M, Mehler J, Bonatti LL. Finding words and rules in a speech stream. Psychol Science. 2008;19(2):137–44.
30. Shukla M, White KS, Aslin RN. Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. P Natl Acad Sci USA. 2011;108(15):6038–43.
31. Monaghan P, Christiansen MH, Chater N. The phonological-distributional coherence hypothesis: cross-linguistic evidence in language acquisition. Cogn Psychol. 2007;55(4):259–305.
32. Richardson FM, Thomas MSC, Price CJ. Neuronal activation for semantically reversible sentences. J Cogn Neurosci. 2010;22(6): 1283–98.
33. Gruening A. Neural networks and the complexity of languages [Ph.D. dissertation]. School of Mathematics and Computer Science, University of Leipzig; 2004.
34. Longobardi G, Guardiano C. Evidence for syntax as a signal of historical relatedness. Lingua. 2009;119(11):1679–706.