

Why verb-initial languages are not so frequent

CogSci Summer School 2002, Sofia

André Grüning*

Max Planck Institute for Mathematics in the Sciences

Inselstr. 22–26

D-04103 Leipzig, Germany

May 31, 2002

Abstract

In our simulations with simple recurrent networks we demonstrate that small artificial languages are learnt differently depending on their basic word order. We show that verb-initial languages are difficult to learn, reflecting the lower frequency of verb-initial natural languages.

We try to go beyond mere simulations proposing two objective mathematical measures to explain our results.

1 Introduction

Most natural languages can be assigned a *basic word order*, in which verb **V**, (non-pronominal) subject **S** and a possible (non-pronominal) direct object **O** appear in simple declarative clauses. English e.g. is an **SVO**-language, while Welsh is **VSO** and Japanese **SOV**. The six possible orders of **S**, **V**, **O** are not equally frequent in the world's languages (table 1) [10].

We want to follow the connectionist approach assuming that complex (linguistic) behavior can be explained better by sub-symbolic computation using neural networks rather than symbolic rules.

While traditional linguists ascribe the similarity of natural languages to some innate hard-wired *universal grammar* (UG), the connectionist belief is that rule-like behavior emerges from the cooperation of many simple neurons.

There has been a lot of connectionist work to

show that UG rules need not be hard-wired but emerge in a natural way in trained neural networks, e.g. [4]. These network simulations are paralleled to human linguistic behavior: Networks learn a particular rule better (worse) which is a hint why this rule is (not) preferred for natural language, too. This is done for basic word order in section 2.

In fact, one only shows that a rule is learnt better/worse by one particular network type with one particular learning rule, and one feels the need for a deeper explanation of the results, making the connection to natural language stronger. This does not mean that we want to go back to formal grammar. We rather think in terms of dynamical systems [8, 1].

Processing language means translating hierarchically structured data into a time series and vice-versa. We believe that natural measures of complexity for times series can be found that are relevant for natural languages. In the best case these measures would assign a low complexity to rules

order	freq.	ΔI	states
SOV	45%	0.218	10
SVO	42%	0	10
VSO	9%	0.817	15
VOS	3%	0.820	15
OVS	< 1%	0	8
OSV	< 1%	0.193	10

Table 1: Frequency of natural languages, information loss ΔI and the number of states in the minimal FSA.

*e-mail: gruening@mis.mpg.de

that are frequent in natural languages. Some steps in this direction are undertaken in section 3.

2 Simulations

Corpus Our corpus of sentence templates is constructed from a lexicon of 6 nouns and 10 verbs (tables 2, 3) with labels that resemble their real world counterparts.

The verb is commonly assumed to select its arguments (subject and object) [3, 9]. Even though a particular subject may restrict the possible objects and vice-versa, here as a first approximation the verb selects its arguments independently.

To build a sentence template a verb is chosen, then its arguments are selected conforming to the subcategorizing properties.¹ 182 different templates can be generated, each with a certain probability (table 4).

A corpus of templates is constructed drawing 10000 sentence templates according to their probability and is then output in the six possible basic word orders to give six differently ordered corpora (SVO, SOV, ...). To each sentence an end-of-sentence marker is added.

Simple recurrent networks (SRN) have an explicit short-term memory of one time step, but develop during training a short-term memory that implicitly extends further back in time. The networks consist of input and output layer and one hidden layer. The corpus is presented to the network word by word using unary coding in a word prediction

¹Wherever more than one choice is possible, all alternatives have equal probability. Optionally transitive verbs take an object in half the cases.

label	property
<i>book</i>	-
<i>dog</i>	a
<i>house</i>	-
<i>man</i>	h
<i>mouse</i>	a
<i>woman</i>	h

Table 2: Nouns in the lexicon. Properties: a = animal, h = human, - = none

label	transitive	subj.	obj.
<i>break</i>	optional	-	-
<i>call</i>	optional	h	a ∨ h
<i>chase</i>	yes	a ∨ h	a ∨ h
<i>cry</i>	no	h	-
<i>destroy</i>	yes	-	-
<i>eat</i>	yes	h	a
<i>kill</i>	yes	a ∨ h	a ∨ h
<i>move</i>	optional	a ∨ h	-
<i>run</i>	no	a ∨ h	-
<i>see</i>	yes	a ∨ h	-

Table 3: Verbs in the lexicon. Required argument properties: a = animal, h = human, - = none

task [5]. Thus there are 17 input and output nodes (16 words and end-of-sentence). To check if learning is successful, the mean square error (MSE) between the networks' output and target activations is computed.

For each of the six corpora 100 SRN are initialized and trained for 100 epochs.² The whole experiment is repeated for nets with hidden layer sizes between 5 and 20.

Results As the generic case the averaged MSE for the 100 networks with 9 hidden neurons are presented in figure 1 up to epoch 20³

We observe that the verb-initial languages are learnt much worse than the subject- or object-initial ones, which are learnt almost equally well.

Examining the output activations more closely, we note that the networks fail to learn to look back in time for more than one time step *accu-*

²Here the learning rate was 0.2. The nets turned out not to be very sensitive to variation in learning rate or a momentum different from 0.

³There are no qualitative changes after 20 epochs. For the sake of clarity confidence intervals have been left out. The statements in this section are with confidence of 95% or better.

sentence	probability
<i>man eat mouse</i>	1/40
<i>house break</i>	1/120

Table 4: Two sentences templates from the corpus in SVO order and their probabilities.

4 Conclusion

A corpus with six possible different word orders was fed into SRNs as a word prediction task. We computed conditional entropies and the minimal FSA for each language.

The computer simulations, entropies and the FSA demonstrate that verb-initial languages are more difficult to learn and have a higher complexity than argument-initial ones. The precise order within these two groups varied, but the figures are more similar for the members within each group than for any language outside the respective group.

Leaving aside the object-first languages (OSV, OVS, VOS) for a moment, the simulations as well as the additional considerations about entropy and FSA reflect the frequency distribution of word orders in the world's languages. Our theoretical considerations back-up the simulations and yield a measure of complexity that is independent of the particular network type and learning rule used.

Why did our approach fail for the object-first languages? From the construction of our corpus it is clear that object and subject are treated almost symmetrically. Thus results are expected to be similar when subject and object are exchanged.

But why are object first language so rare in the real world? Our setup is such that only syntactic phenomena can be captured. Invoking pragmatics

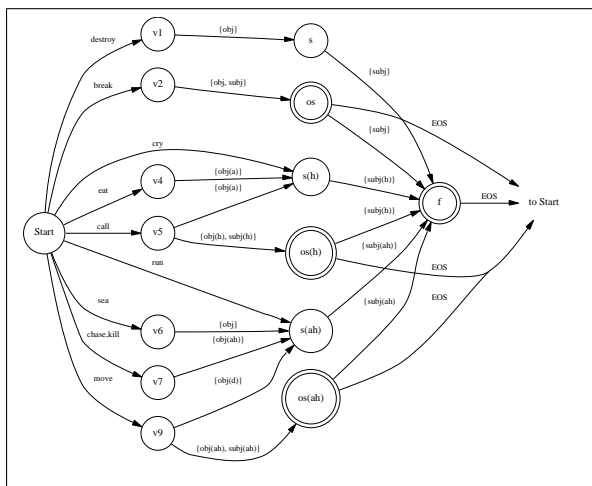


Figure 3: Minimal Automaton for VOS

we argue here that in natural languages it is useful to include subjects, which often give the topic, in the beginning of a sentence to enable the early use of contextual information, breaking the symmetry between subject and object.

References

- [1] Mikael Bodén, Janet Wiles, Bradley Tonkes, and Alan Blair. Learning to predict a context-free language: Analysis of dynamics in recurrent hidden units. In D. Willshaw and A. Murry, editors, *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN 99)*, pages 359–364, 1999.
- [2] Mike Casey. The dynamics of discrete-time computation, with application to recurrent neural networks and finite state machine extraction. *Neural Computation*, 8:1135–1178, 1996.
- [3] Vivan J. Cook and Mark Newson. *Chomsky's Universal Grammar*. Blackwell, Oxford, 2nd edition, 1996.
- [4] Michelle R. Ellefson and Morton H. Christiansen. Subjacency constraints without universal grammar: Evidence from artificial language learning and connectionist modeling. In *The Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, 2000.
- [5] Jeffrey L Elman. Finding structure in time. *Cognitive Science*, 14:179 – 211, 1990.
- [6] Stanford Goldman. *Information Theory*. Prentice Hall, New York, 1953.
- [7] John E. Hopcroft and Jerrey D. Ullmann. *Introduction to Automata Theory, languages and computation*. Addison-Wesley, Mass., 1979.
- [8] Christopher Moore. Dynamical recognizers: Real-time language recognition by analog computers. *Theoretical Computer Science*, 201, 1998.
- [9] Carl Pollard and Ivan A. Sag. *Head-driven phrase structure grammar*. Univ. of Chicago Pr., 1994.
- [10] R.S. Tomlin. *Basic Word Order: Functional Principles*. Croom Helm, London, 1986.