Language Type Frequency and Learnability

from a Connectionist Perspective

Ezra van Everbroeck
Department of Linguistics
University of California, San Diego
La Jolla, CA 92093-0108
USA
ezra@ling.ucsd.edu

Abstract

To investigate a possible connection between language type frequency and learnability, we systematically compared how neural network models learn all possible combinations of three linguistic strategies for encoding grammatical relations: word order, nominal case marking, and verbal affixes. Other variable linguistic dimensions included accusative and ergative marking systems, the consistency of genitive marking, and the complexity of the grammars used to generate the artificial languages. The results of our simulations mesh well with some of the typological tendencies observed among the languages of the world: e.g. Subject-before-Object languages are more frequent than their Object-before-Subject counterparts; ergative languages are less common than accusative languages; and SOV languages almost always appear with a nominal case marking system. In general, the networks were able to learn the attested language types, but typically had severe problems learning the unattested types. However, the simulation results do not explain why some language types are more frequent than others.

Keywords: word order, case marking, head marking, frequency, learnability, connectionism

1. Introduction

In this paper, we present and analyze the correspondences between the learnability and frequency of different strategies for encoding the two main arguments of a clause, i.e. the "Subject" and "Object".[1] Cross-linguistically, there are three common strategies for disambiguating between these two arguments: word order, case marking on noun phrases (dependent-marking), and encoding of the argument relations on the predicate (head-marking). This three-way distinction, which goes back to Sapir (1921), can be illustrated by the contrast between English, in which disambiguation is achieved by word order, Latin, in which the arguments are distinguished by case affixes attached to the respective noun phrases, and Yimas, in which who did what to whom is encoded by affixes on the verb. Examples of these three strategies are given in (1) through (3).

(1) a. The bull killed the matador. (English)
b. The matador killed the bull.

(2) a. *canis* *occidit* *catum* (Latin)
dog-NOM killed cat-ACC
"The dog killed the cat."

b. *canem* *occidit* *catus*
dog-ACC killed cat-NOM
"The cat killed the dog."

(3) a. *payum* *narmaŋ* *na-mpu-tay* (Yimas, Foley 1991: 193)
man-PL woman-SG 3SG-3PL-see
"The men saw the woman."

b. *payum* *narmaŋ* *pu-n-tay*
man-PL woman-SG 3PL-3SG-see
"The woman saw the men."

The fact that languages can encode functional information along these different lines has often been described as an opposition between syntax (as in modern English) and morphology (be it nominal marking as in Latin, or verb marking as in Yimas). Three empirical findings about

these types of encoding are of interest: first, consistent simultaneous application of all three types of coding in a single language is not attested. Second, there also does not appear to be a single language which completely fails to use morphological or syntactic means to tell apart the two main arguments.[2] Finally, some combinations of features occur with greater frequency than others. For example, some word orders appear to be at least 30 times more common than certain others (see below).

The first two empirical findings have an intuitively obvious explanation: efficient communication would be difficult if a sentence like *taljun kovents horan* could mean with equal probability that *the farmer killed the baker* or that *the baker killed the farmer* — hence, we can expect that languages will typically have some strategy for telling Subjects and Objects apart. The temporal dimension of language and the respect for conversational maxims should also force speakers to avoid coding the same information redundantly, because in the time which it takes to produce and process the redundant forms, more important information could have been passed on from the speaker to the hearer. Still, plausible though these explanations may seem to linguists, they are inadequate in that they are currently no more than intuitions, without much precision or experimental support. They also entirely fail to elucidate the third observation, namely why certain language types should be so much more frequent than others. Finally, they are of limited use for answering the question whether the complete absence of certain language types is motivated by general (communicative) principles — i.e. they are outside of the human language potential (cf. Comrie 1993) — or whether these absences are likely to be historical accidents, for example, as the result of all such languages having become extinct due to non-linguistic factors.

Typologists have long tried to account for the distribution of various linguistic properties which we observe in the world today. Maslova (2000) has recently suggested that the current data may still largely reflect the distribution of linguistic properties in the presumably much smaller set of languages spoken thousands of years ago. If this is the case, we have to be careful about reading much into the correlations of features that we find nowadays (but see Dryer 2000). Hawkins (1983, 1994), on the other hand, has long argued that there is an underlying performance principle shaping the forms natural languages take. He finds the strongest explanations for language universals in the perceptual and processing capacities which all human being share. Specifically, language users aim to maximize the left-to-right immediate constituent-to-word ratios of the phrasal categories they construct (see below). Another factor which may account for the current distribution of linguistic properties is the language acquisition capacities of children (e.g. Jespersen 1922; but see Croft 2000 for the limitations of child-based theories of language change). For example, Bybee Hooper (1980) has described how the ways in which morphological patterns are acquired by children are sometimes strikingly similar to how the morphologies of languages change with time.

Our work relates to the explanation of universals by both performance and language acquisition strategies because it explores the effects of language type learnability on the observed frequencies of different strategies for encoding Subjects and Objects — from completely unattested to very common. In particular, we test under which conditions a small set of language data is sufficient for a simple learning mechanism to develop a suitable parsing strategy for more sentences generated by the same grammar. We intend to show that there are many correspondences between how our computer models perform on the various language types and how well these same types are attested in the real world.

The learning mechanism we have chosen to use for our experiments are connectionist models. The last few decades have seen a considerable amount of successful research into using artificial neural networks for a wide variety of language-learning tasks, from issues in phonology and morphology over syntax to pragmatics (Sejnowski & Rosenberg 1987; Gasser 1993; Smith 1995; Elman 1990, 1992; Katz & Dorfman 1992).[3] Artificial neural networks also boast a higher degree of cognitive plausibility than most other learning mechanisms: although the popular backpropagation algorithm used in many simulations (including the ones presented below) is biologically unrealistic, there is growing evidence from neuroscientific studies that the more general class of error-driven learning algorithms which backpropagation belongs to is indeed represented in the brain (e.g. Fletcher *et al.* 2001). Finally, the use of neural network simulations for the current task of comparing the learnability of various language types provides us with a relatively easy way of extending these models to look both at the trajectory of acquisition of specific languages (cf. the U-shaped curve in the acquisition of the English past tense, Rumelhart & McClelland 1986), as well as how languages may evolve over time (Hare & Elman 1995).

This paper uses neural network models to compare how three different strategies for encoding two main arguments — linear word order, dependent-marking (namely, case-marking), and head-marking — affect the learnability of the language types which implement them. Given the desire to look at a wide spectrum of language types, both attested and unattested, it is necessary to compare all of them in a systematic manner by only varying how the languages use morphological markers and word order to encode grammatical relations. In order to keep our simulations manageable, both in terms of number and complexity, we have had to limit the artificial languages used in two important areas: first, the languages do not have any semantics associated with the words in their lexica; second, all noun phrases in the languages are expressed

as lexical nouns — there are no personal pronouns or null forms in any of the sentences. These limitations obviously entail that our models of language types (e.g., SOV with accusative case-marking) cannot be expected to map one-on-one onto any particular natural language which is of a similar type (e.g., Japanese); rather, the models we use should be viewed as abstractions of the relevant types. Nonetheless, we feel that the limitations are also a strength of the models because it allows us to compare the relative difficulty of disambiguating Subject and Object when only minimal information is available. As a result, we can expect to find more pronounced differences between language types than if we included helpful semantic or contextual information. Studying the hardest cases, i.e. pushing the limits of learnability, is also a useful first step. The modeling provides us with a baseline that can be later used to compare with richer models which include, for example, pronouns or rich semantic information.[4] The task of the networks is to learn how each language type encodes its Subjects and Objects. The experiments allow us to compare how easy it is to learn the different strategies, and the results can give us insight into the patterns of language types established by typology.

So far, we have discussed the two main assumptions of our research: the first one is that connectionist models are appropriate tools for studying language acquisition; the second that the different Subject and Object encoding strategies can be investigated in simplified artificial languages. The central hypothesis which we investigate in this paper connects the two assumptions by positing a link between language type learnability and frequency: languages in which the Subject/Object distinction is easy to learn would remain stable over time, whereas those in which it is harder to acquire would be more likely to evolve until they too happen to reach a combination of linguistic properties which is learnt well by human children. After sufficient time has passed, we could expect that — all other things being equal — the frequency

of language types would be a straightforward reflection of their learnability (cf. Christiansen & Devlin 1997). The main mechanism at work here is that of "imperfect learning". The learners fail to acquire the more infrequent and/or unusual forms provided to them in their language input and end up replacing them with forms which follow the main patterns in the language more closely. That such a learning mechanism can indeed lead to linguistically plausible results has been demonstrated by connectionist models of the historical changes in the past tense system of Old English (Hare & Elman 1995), as well as the nominal system of Vulgar Latin (Polinsky & Van Everbroeck 2000).

What would it take for there to be a strong correlation between learnability by a neural network and frequency of language types? First, our neural networks would have to learn the attested language types. Second, these same neural networks should have problems learning the unattested types.[5] Third, the networks should consistently learn the more frequent types faster and/or better. We will see below that this third prediction is not borne out by the simulation results: for the attested language types, we usually did not find a correlation between their observed frequency and how well the models learn them. Consequently, our simulations provide strong experimental evidence that the frequency of language types is not determined solely (or even mainly) by how easily they can be learned — instead, historical factors like the ones described in Diamond (1999) have almost certainly played a more important role. Nonetheless, we also want to argue that learnability issues can not be ignored altogether: many of the connectionist models suggest that the absence of certain combinations of linguistic features are probably linked to learnability constraints: the language types which are unattested typically seem to be so for a reason. In those cases where the simulation results diverge from what is known about the world's languages, the models suggest that we should look for other linguistic

properties — currently not yet modeled — which turn an otherwise hard to learn language type into a learnable one.

The structure of this paper is as follows: in section 2, we will survey typological literature dealing with word order patterns, as well as the opposition between marking on verbs and noun phrases. Section 3 briefly describes the key features of the implementation of the connectionist model. Section 4 presents the experimental results of the three sets of simulations: the first one uses very simple languages in which the sentences contain only a verb, a Subject, and (possibly) an Object; the second set also includes locative phrases and possessives; and the last set adds relative clauses to the grammar. Sections 5 and 6 provide the general discussion and conclusion.

2. Language data

There are three main areas of typological research which are important for the current investigation into possible language types: basic word order, head and dependent marking, and word order correlations. We will now go over the relevant data from each.

2.1. *Free and fixed word order languages*

Careful study of a large number of natural languages has shown that very few of them are limited to sentences with a single word order of Subject, Object and Verb, and some of them appear to allow all six possible orders (VSO, SVO, SOV, OSV, OVS, VOS). Even in English, a prime example of a rigid (SVO) language (MacWhinney & Bates 1989), it is possible to use the inverted VS construction presentationally or in stage directions (Lambrecht 1994; Birner & Ward 1998). Word order splits motivated by tense, animacy, definiteness and the like are otherwise well attested across languages (e.g. Tsunoda 1981; Polinskaja 1989; Dryer 1992). Exhaustively exploring this space of possible language types would entail looking at all possible combinations of one or more word orders, as well as varying the frequencies of all the word orders (Steele 1978; Siewierska 1998). Even with the computational resources available nowadays, it would be practically infeasible to investigate such a large space. In addition, it is not obvious that the differences between all these possible types would have much linguistic relevance.

If we use criteria like "basic" or "dominant" word order (Steele 1978; Hawkins 1983; Siewierska 1988, 1996; Dryer 1998), we find that there are certain orders of Subject, Object and Verb which cross-linguistically occur with much greater frequency than others. Of the six possible orders only two are really common, and two others appear hardly at all. Using data from various sources we can construct the frequency hierarchy shown in Table 1 (Tomlin 1986; Dryer 1989; Siewierska 1998).[6]

[approximate location of Table 1]

Probably the most direct criticism of the hypothesis that all languages should have a basic word order of Subject, Verb and Object has been proposed by Mithun (1987). Using data from Cayuga, Ngandi and Coos, three highly polysynthetic languages with obligatory pronominal affixes on all verbs, she has shown that none of these languages has a fixed order like SVO or SOV. It is even impossible to use a simple frequency criterion because the percentage of clauses with a full lexical Subject and Object noun phrases is only 1% to 3%. And for those sentences in which two full nominals occur, she has found that there is no preferred reading: for example, the Cayuga sentence below is ambiguous as to who beat whom without more context (Mithun 1987: 286).

(4) *Khyotro:wę́:* *Ohswe:kę́'* *ahǫwati:kwé:ni'*
Buffalo Six Nations they-beat-them
"Buffalo beat Six Nations." or "Six Nations beat Buffalo."

Although all three languages also allow basically any order of constituents, it turns out that there is not as much freedom as there appears to be at first sight. Though all of the orders are attested (and considered grammatical), they differ in the way they present their information: pragmatic factors are always at work, and it is usually the constituent in the clause which is taken to be newsworthy which is fronted "because it represents significant new information, because it introduces a new topic, or because it points out a significant contrast" (Mithun 1987: 304).

2.2. Head-marking and dependent-marking

Another line of typological research has been concerned with preferences for marking. Nichols (1986) has shown that systematic patterns of marking can be found among the world's languages in how they use Head-Marking (H-marking) and Dependent-Marking (D-marking). Nichols proposes a four-way split here: languages can be predominantly Head-marking, Dependent-

marking, double-marking (i.e. have markers on both the head and the dependent), or split-marking (i.e. have some H-marking mechanisms, and some D-marking ones).

The results of analyzing a sample of 60 languages show that though most languages in the sample are mostly either H-marking or D-marking, there is apparently not a single language which uses a single type exclusively. However, the splits are not random at all: H-marking is favored at the clausal level (e.g. negation words, or person/number/gender agreement on verbs), while D-marking occurs more often at the phrasal level (i.e. case markers typically attach to nouns, not to the clause as a whole). Nichols also found that D-marking languages are not only the most frequent, but they also have the widest geographical distribution. Finally, Nichols reported that there are statistically significant correlations between the use of certain marking patterns in a language, and their dominant word order. Specifically, "head-marking morphology favors verb-initial order, while dependent-marking morphology disfavors it" (81). D-marking, however, is strongly correlated with verb-medial languages. Though Nichols hints at a functional motivation for these data, she does not provide any experimental evidence to support it. Hence, one of the main goals of our modeling project has been exactly to put her hypothesis to a connectionist test, and to find out with more precision how useful it is, for example, for a verb-medial language to have D-marking as opposed to a verb-initial or verb-final one.

2.3. *Word order correlations*

Ever since Greenberg's (1963) paper describing some "universals of grammar", there has been a steady interest in cross-linguistic word order generalizations (e.g. Venneman 1975; Hawkins 1983; Dryer 1992). The field has seen increasing sophistication, mostly through the use of better sampling methods and larger language samples (Harris and Campbell 1995; Rijkhoff & Bakker 1998). The common thread in this research paradigm has been the search for correlations

between various constituent pairs, like Noun — Adposition, or Noun — Genitive. The finding has been that languages with prepositions tend to have the Genitive follow the Noun; while postpositional languages commonly occur with the Genitive preceding its head noun.

In Table 2, we have summarized Dryer's (1988) data for correlations between the position of the verb in relation to the Subject and Object on the one hand, and the orders of four other word pairs on the other. It shows the number of language families which combine a certain basic word order with specific orders of Adposition — Noun, Noun — Genitive, Noun — Relative Clause, and Noun — Adjective. This last pair shows a typical pattern for a non-correlated pair.

[approximate location of Table 2]

It is not hard to see that SVO and VSO pattern together for the Adpositions and Relative Clauses, but that there is no such uniform behavior for the order of Noun — Genitive; for the latter, VSO and SOV are each strongly correlated with one particular order, but SVO languages do not display a similar preference for one or the other. In some cases, like English, an SVO language will appear with both orders of Noun and Genitive: compare *God's wrath* and *wrath of God*.

There is an obvious pattern to all these correlations: languages either tend to feature the phrasal head consistently preceding the modifier (i.e. head-initial), or vice versa in the head-final cases. Where the null hypothesis would have predicted a random distribution of these operators and operands, we instead find strong tendencies towards the extreme positions on the scale: the pairs tend to be "harmonic" with one another in which element precedes the other one. This observation has received several different explanations in the history of language typology,

though the one which is of special interest here is the performance-based explanation proposed by Hawkins (1988, 1993, 1994, 1999; as an alternative, compare the Branching Direction Theory in Dryer 1992). Hawkins expects to find the strongest explanations for language universals in the perceptual and processing capacities which all humans share. For example, avoiding garden path constructions with sentence-initial relative clauses (Hawkins 1993), and postposing of heavy sentential complements in SOV languages (Hawkins 1988) can both be interpreted as an attempt to reduce the processing load for the hearer: during parsing, humans prefer to maximize the left-to-right immediate constituent-to-word ratios of the phrasal categories that they construct. The construction of such categories is in turn made possible when we hear words which uniquely predict the existence of such a phrase.[7] In order to make the early analysis of immediate constituents feasible, the heads tend to appear at the periphery of their phrases. As it is presumably easier to parse them if they consistently appear at the same side of various phrases, we end up with a situation in which the heads are, in the unmarked case, consistently left-peripheral (head-initial) or consistently right-peripheral (head-final) within their phrasal categories.

2.4. *Summary*

Given the goal of keeping the artificial languages used in the computer simulations similar in their properties to the known data about the languages of the world, it should come as no surprise that the findings from each of the three areas of typological research have contributed to the way the connectionist simulations were created. The data on basic word order are reflected in the use of the six fixed word order types, as well as a free word order type. It should be mentioned, though, that all the fixed word order languages only use a single word order with no variants. The main reason for this apparent simplification is that it explores the limits of a particular word

order when no alternatives are available. This in turn makes it much easier to compare the implications of each word order for the acquisition of the relevant language. With respect to the free word order language in the simulations, it should be pointed out that the "free" order in the simulations was truly random, with each order having exactly as much chance of appearing as any other one. There is a good practical reason for this limitation of the models: in the absence of a theory of pragmatics which is compatible with connectionist modeling, it is hard to imagine how one would go about accounting for pragmatic effects with a neural network. Nonetheless, it turns out that the free word order models, despite their obvious flaws, are still good enough to exhibit some of the same preferences found in natural languages. In particular, they demonstrate why "free" word order languages favor Dependent-marking over Head-marking. Finally, each fixed word order type in the model is further divided into subtypes with (i) both Head- and Dependent-marking, (ii) one of the two types of marking, or (iii) neither marking. For the locative phrases, the possessives, and relative clauses, only D-marking was implemented as an option. The models were made to reflect the well-known word order correlations. For example, a fixed VSO word order was modeled with prepositions, postnominal possessives, and postnominal relative clauses. Models learning SOV sentences, on the other hand, saw postpositions, prenominal possessives, and prenominal relative clauses. A certain simplification was assumed with respect to SVO models: they were in all respects like English, in featuring prepositions, postnominal relative clauses, and both prenominal and postnominal possessives. Thus, SVO languages that show features exclusive of a verb-initial type (as in Romance or Bantu) were not modeled. The next section is concerned with the actual implementation of the models.

3. The model

Cognitive scientists have been developing computational models of linguistic phenomena more than a decade now (e.g. Rumelhart & McClelland 1986), but there have been few simulations which have looked at cross-linguistic phenomena. Kirby (1997) investigated a possible scenario for the emergence of the accessibility hierarchy in relative clauses (Keenan & Comrie 1977) by simulating generations of language speakers who were in contact with one another. He found that if there were certain costs and benefits associated with producing and parsing relative clauses, the speakers would eventually assume languages which obey the accessibility hierarchy. Of more relevance to the current simulations, Christiansen & Devlin (1997) used connectionist models to assess the importance for a language of being consistently head-final or head-initial. They showed that the models had more problems learning inconsistent language types, and that the latter were also more infrequent among the languages of the world — a similar explicit link between frequency and learnability is also the central hypothesis behind the simulations presented below. The present work is different from Christiansen & Devlin in that it varies more and different parameters: it assumes the consistency of head-dependent pairs but considers various possibilities for word order, presence of H- and D-marking, as well the accusative versus ergative opposition. Given the large number of possible language types already generated by these parameters, the addition of the one explored by Christiansen & Devlin (1997) would have made the project intractable.

3.1. *The artificial language*

Before we describe the key features of the way in which the neural network models were set up, we would like to elaborate on the use of an artificial language[8]. The use of such a language for *any* experiment with supposedly linguistic implications is not immediately obvious — even more

so when this artificial language does not do justice to a number of basic facts about natural languages. For example, words in human languages invariably have meanings (be it lexical or grammatical), whereas the ones used in this experiment did not. Nonetheless, there are several reasons which lead us to believe that an artificial language rather than one or more natural ones is the appropriate choice for the simulations presented below:

Not all possibilities are represented by natural languages. Certain conceivable combinations of word order type and morphological marking do not seem to occur in the real world. Hence, if one limits oneself to natural languages, one can never explore how a connectionist network would deal with unattested types. One of the advantages of models like these is exactly that they can distinguish between language types which are unattested because they are in some way unlearnable, and language types which are only unattested because of historical accidents. Also, the option of only using an artificial language for the possibilities that are not represented in the real world (versus natural languages for the ones that are) is not desirable because it would have made it considerably more difficult to compare the results for the different language types.

Natural languages are hard to generate. Human languages are so complex that no one has been able to generate more than a limited fragment of them automatically. To the best of my knowledge, no complete phrase-structure grammar (context-free or context-sensitive) exists for English, let alone for all the other language types which had to be simulated.

Natural languages differ along many dimensions. Given that it has been one of the goals of the experiments to find out whether, for example, a fixed word order SOV language with H-marking suffixes is easier to learn than a similar language without such suffixes, we did not want

other parameters to influence the comparison. In real languages, the fact that one features vowel harmony while another one does not might conceivably have some effect on how information about Subject and Object is encoded in these languages. There is a trade-off therefore in the degree to which one can model specific natural languages, and the desire to look at more abstract language types. If one wants to do the latter, one needs perfect control over all the parameters, and only the use of an artificial language gives one such control.

The artificial language which has been developed for the simulations is completely computer-generated, though most of its phonology, morphology, and syntax are inspired by the data available about natural languages. In short, a context-free grammar was written with symbol rewriting phrase-structure rules. These rules then generated sentences of the language by replacing all the symbols for, e.g. noun and verb, with words from the lexicon. The original sentences were characterized by a fixed word order, Dependent-marking (on the nouns), and Head-marking (on the verbs); so, the information about which NP was the Subject and which the Object was encoded in three different ways. For the fixed word order language SOV, this version will be referred to mnemonically in this paper as SOV/HD — i.e. SOV order, H-marking and D-marking. A language like SOV/HD was then used to generate the SOV/HX (no D-marking), SOV/XD (no H-marking), and SOV/XX (no D- or H-marking) languages by simply removing the appropriate markers. The other fixed word order languages were generated in the same way. The four free word order languages, which will be referred to as XXX/HD, XXX/HX, XXX/XD, and XXX/XX, were created by random scrambling of the constituents of one of the fixed word order languages. Table 3 summarizes the three dimensions of coding.

[approximate location of Table 3]

An example sentence in each of the four SOV languages is presented in Table 4 — notice the placement of the Head-marking and Dependent-marking suffixes. One might wonder why all the affixation was done via suffixes, rather than prefixes or a mix of both. Especially for the Head-initial language, VSO, a prefixing strategy may appear more appropriate. However, there is a clear cross-linguistic tendency in favor of suffixation (Greenberg 1963), so (mostly) suffixing VSO languages are also attested. Moreover, the way the model was implemented practically eliminates the possible effects — whatever they may be — of prefixation as opposed to suffixation. As we will see in the discussion of the input layer of the network below, each word in a sentence was presented to the network in its entirety. So, unlike human readers or hearers, who receive their linguistic input (roughly) one letter or phoneme at a time, the simulations were shown complete words, one at a time. Hence, either prefixes or suffixes would be equally salient for the networks.

[approximate location of Table 4]

3.2. *The network*

Figure 1 gives a schematic representation of the network used for the experiments. It consists of four layers, with the two hidden layers each having a context layer to store information about the previous words in the sentence (Elman 1990). Each full line between two layers means that they are fully interconnected in that direction (i.e. each unit in the first layer has a connection with each unit in the second layer); a dashed line indicates that each unit in the first layer just copies its activation value to the corresponding unit in the second layer. Within every layer except for the input layer all units also have connections to every other unit.

[approximate location of Figure 1]

At the input layer, the individual words in each sentence were presented to the network one word at a time, and the 288 units of the input layer were used to give the network a phonological representation of each word. Each of the phonemes of a word was translated into a 16-bit vector of ones and zeros, with each bit signifying a phonological feature, for example, whether the phoneme was [± nasal] or [± voice]. So, phonemes which are linguistically similar also received similar input representation: e.g. the representations for /t/ (0000000010100000) and /d/ (0000000010100001) were only different in the last bit, as it encoded [± voice].[9]

The units in the hidden layers allow the network to construct its own representations of the regularities in the input, and to solve problems which a network without such a layer would never be able to tackle. Though it is counter-intuitive, the architecture has a second hidden layer in order to *decrease* the number of connections in the network and thus speed up the computations (see Elman 1992).

After each input pattern has been presented, the pattern of activation in a hidden layer is copied as-is to its context layer (Elman 1990). When the next input pattern is presented, the hidden units receive information not only from their input units, but also from their context units, so the new pattern of activation over the hidden units combines information from both the current input pattern and the previous one. As new words are presented at the input layer, the recurrent context layers act as a short-term memory and allow the networks to construct a representation which contains information about all the words it has recently seen. Using this information, the network can learn to classify the very same input pattern in different ways: for example, whether a noun is the Subject or Object can depend on its position in the sentence.

Although they all function as an indicator of a single linguistic property, the 12 units in the output layer conceptually fall into three distinct groups (see Table 5). A first set of 7 units signal which kind of element the network thinks the current input word is; the options here are Subject, Object, Possessive, Location (used for both the adposition and the accompanying noun), Main Verb, Subordinate Verb, and Relative Pronoun. The second set of 4 units are used as word class detectors: Noun, Verb, Adposition, and Pronoun. The last unit signals whether the word is part of the main clause or the subordinate (relative) clause.[10]

[approximate location of Table 5]

Table 6 shows the various possible desired output patterns for the first 11 output units, along with a brief description of what each pattern corresponds to.

[approximate location of Table 6]

Before a network was trained on a corpus, all its forward connections were initialized to random values between –1 and +1; the values used were exactly the same for all networks, though, so that the results for different language types could be compared more easily (i.e. a within-subjects design). As is customary for this type of network architecture, the one-to-one copy-back connections from the hidden layers to the context layers were frozen at +1 to make sure that the memory capacity of the latter would be used. Within the context layers, the units were initialized with activation values of a slightly inhibitory –0.1 and the weights from the recurrent units to themselves were frozen at 0.

Training and testing was always done with corpora of 3,000 sentences. During training, the network saw each sentence in the training corpus 30 times.[11] On each iteration, the network was presented with a pattern of activation at the input layer representing a single word. The activations traveled through the connections between the units and produced another pattern of activation over the output units. This pattern was then compared to the correct pattern. If there were discrepancies for any unit between the output produced by the network and the desired output, the weights on the relevant connections were changed slightly using the backpropagation learning algorithm (we used the logistic activation function and set η to 0.05; Rumelhart, Hinton & Williams 1986), so that a future presentation of the same input pattern would produce an output pattern which was closer to the correct one. After training, the final performance of the network was determined, and it was also checked how well it could generalize its knowledge to the completely new sentences in the testing corpus. Naturally, a network trained on the SOV/HX language would also see SOV/HX sentences in the testing corpus.

To conclude this section, it may be useful to step through a single sentence the way the model would during training. An actual sentence from the SOV/XD corpus is given in (6) — because the words don't have meanings, the glosses and translation are obviously arbitrary and the abstract description [[[Possessive Subject] [Location Postposition] Verb] [Possessive Subject] Object Verb] is in many ways more appropriate. The example also illustrates the complexity of the sentences generated by the most complex grammars used in the simulations.

(5)

lægGutil	*GjæXæuX*	*tjElew*	*poX*	*sjEts*	*slOfsil*	*spisiskEguX*	*jigEEmt*	*bælo*
baker-GEN	dog-NOM	yard-LOC	in	bitten	butcher-GEN	cat-NOM	bird-ACC	saw

"The butcher's cat, which the baker's dog had bitten in the yard, saw a bird."

The desired output patterns for each of the words in (5) are shown in regular typeface in Table 7; the output values actually produced by the network after 30 epochs of training are

shown in italics. All the cells in which the unit's activation exceeded 0.2 are indicated by a darker background. Note how each input word always requires the activation of one output unit from the first set of seven units and one from the second set of four units. The very last output unit is supposed to be active only when the network is processing words belonging to the main clause.[12]

[approximate location of Table 7]

A comparison of the desired and actual values on the output units shows that the SOV/XD network can parse this sentence without any real problems: the correct unit is always activated for the first and second sets of output units, and no spurious units are activated. If we look at the last output unit, we can see that the network interprets the first four words as the beginning of the main clause, but then realizes its mistake when it "sees" *sjEts* as the input word. The network has learned that this is a transitive verb, so the absence of a preceding Object noun is a signal that the verb is part of a (subordinate) relative clause, and the value of the last output unit reflects this interpretation. When the main clause actually starts, the network correctly keeps the final output unit activated until it reaches the end of the sentence.

4. Results and discussion

This section presents the main results from the various experiments. We will first describe the outcome of the simulations with a simple grammar containing only Subjects, Objects and Verbs. We will look at models with both accusative and ergative case-marking (4.1). We will then turn to the experiments with the expanded grammar including possessives and locative phrases (4.2). The most complex simulations which add sentences with relative clauses are discussed in section 4.3.

4.1. *The basic grammar*

The first set of simulations used corpora of sentences which either consisted of a Subject, Object and Verb (transitive), or just a Subject and Verb (intransitive). The variable dimensions include word order (the six possible fixed word orders, and a single free word order) and type of marking (i.e. the four combinations of H- and D-marking). In 4.1.1, we present the results for the nominative/accusative marking, while 4.1.2 presents the results for the absolutive/ergative marking.

4.1.1. *Accusative marking*

In languages with nominative/accusative marking, the same case forms are used for marking the Subject noun phrases of transitive and intransitive clauses, whereas the Objects of the transitive clauses receive a different marker. In accusative languages with Head-marking, the affixes on the Verb referring to the Subject will also be the same in transitive and intransitive clauses, while a different set of affixes is used for referring to Objects. The type of H-marking and D-marking used in the connectionist models reflect these properties of accusative languages.

Because of the number of variable dimensions involved, the results cannot be presented in a single table. So, Table 8 starts by giving summary data for the three Subject-before-Object languages and the free word order language. The performance measure reported in Table 8 is the MSE, i.e. "Mean Sum of squared Errors", which gives an overall error value calculated over all the output units and all the patterns in a single corpus.[13] The "Train" values show how well the network did on the training corpus at the end of the 30 training epochs; the "Test" values illustrate how well a trained network could generalize its knowledge to a corpus with new words — the closer these values are to zero, the better.

[approximate location of Table 8]

Table 8 shows that among the three fixed word order languages *all* perform well with any kind of marking, except for SOV/XX (i.e. the fixed SOV language, with no H- or D-marking) which has MSE values which are many times higher than that of any other order/marking combination. First, the fact that the two VO languages are learnt without any problems is not hard to explain: what the network learns during training is that the first word of each sentence is always the V (VSO) or S (SVO), and the second the S (VSO) or V (SVO). The third word, when present in transitive clauses, is always the O. This parsing technique generalizes to all the sentences in the test corpora. The MSE values are slightly higher for the VSO and SVO "Test" columns in Table 8 than for the corresponding "Train" columns, but this is not surprising if we take into account that all the noun and verb stems in the test corpora were new words, with phonological forms which were completely unfamiliar to the networks. It is likely that some of the new words resembled familiar words, and in those cases where the new word was, for

example, a noun and the familiar word a verb, the network may have erred slightly in its response to the new word.

Second, for SOV, there is the noticeably worse performance of the variant without any kind of marking (i.e. SOV/XX). In general, all the "Test" values are also higher than those observed in the corresponding SVO and VSO language types. If we look at the task of SOV/XX, then it is the only network which was faced with sequences of two unmarked nouns followed by an unmarked verb. For verbs with phonological forms similar to those of inanimate nouns, the network had a relatively hard time learning to distinguish between the bare stems of the Object and verb. As soon as either the nouns (XD) or the verb (HX) received markers, this problem disappeared and training progressed as for the two other Subject-before-Object languages. On the test corpora, however, the SOV networks were all affected to some extent by the new vocabulary. In the case of SOV/XX, the task of classifying the second word in a sentence was an impossible one, because the word form was both new and unmarked. The network reacted by partly activating both the Object and Verb units, but this was obviously counted as an error.

For the free word order networks, it is clear that performance on the XX and HX languages was much worse than that on any of the fixed word order languages. With XX, the only thing which the network could do was to learn each sentence by rote during training. The abysmal failure of XX to generalize to the new test corpus, however, shows the limitations of rote learning for analyzing sentences. The HX network had a slightly easier task because it could easily spot the verbs, but categorizing the nouns as either Subject or Object was still almost impossible. As soon as D-marking was added, this part of the task became feasible and the MSE values improved dramatically.

So far, then, the network results are largely in correspondence with the tendencies observed among natural languages: the Subject-before-Object fixed word order languages can be learned without major problems, except for SOV/XX — i.e. precisely what Greenberg's Universal 41 predicts: "If in a language the verb follows both the nominal subject and the nominal object as the dominant order, the language almost always has a case system" (1963: 96).[14] The simulation results also predict the absence of free word order languages without any D-marking on the nouns. And such language types are indeed very rare: Siewierska (1998: 509) mentions five free word order languages without case marking; Klamath may be an /HX example, though it has a partial case-marking system (Barker 1964). On the other hand, free word order languages with nominal case-marking abound: e.g. Polish, Serbo-Croatian, Latin, Ngandi, Coos (all XD), and Cayuga (HD). Taken together, these results mesh well with the conclusion that "[t]he so-called free word order is connected with the richness of the endings [... while ...] paucity of grammatical means is connected with the grammatically fixed word order" (Sgall 1995:56-57). We can assume that there is a principle of economy at work here, minimizing redundancy in the coding of the same information. With a fixed word order (e.g. SVO/XX), there is less need for overt marking. We will see below, though, that things are somewhat more complicated, because more complex grammars create ambiguities which cannot always be solved as easily without marking. This is quite likely related to the fact that Subject-verb agreement is more common in natural languages (Siewierska 1996) than was built into the simple artificial languages in Table 8, where such agreement is optional.

Let us now turn to the three fixed word order languages in which the Object precedes the Subject. Table 9 contains the MSE values for the networks learning these languages, again on the training and test corpora.

[approximate location of Table 9]

The results in Table 9 are interesting, because they only correspond partially with the tendencies found among natural languages. Recall from section 2.2 that VOS (8%) and especially OVS (<1%) languages are relatively uncommon, and that there are still no uncontested OSV languages. What the simulations show is that Object-before-Subject languages without D-marking on the nouns are hard to learn: both the results on the training and test corpora are considerably worse than what we have seen for the Subject-before-Object languages. Hixkaryana appears to be an OVS language with only H-marking, though the Subject can move to a pre-verbal position and it is rare for both S and O to be expressed by full nominals simultaneously. The Hixkaryana example in (6) illustrates a sentence, with both arguments expressed (Derbyshire 1985: 32).

(6) ***toto*** ***yahosɨye*** ***kamara***
man it-grabbed-him jaguar
"The jaguar grabbed the man."

The network can learn to analyze sentences like this, but generalization to new words fails. This suggests that Hixkaryana could have a limited vocabulary (unlikely) or that its other linguistic features compensate for the inherent difficulties associated with learning an OVS/HX language type. The HD networks, which did perform on a par with their counterparts in Table 8, could conceivably be considered unrealistic because of the redundant coding of the information. But this still leaves us with the XD networks: OSV/XD actually outgeneralizes the other two XD networks, although it is an unattested language type.

The finding that at least some of the Object-before-Subject languages are hard to learn requires some extra consideration. Why should this be the case? The answer may be found not in

the transitive sentences — after all, for the networks a noun with an Object marker is as prominent as one with a Subject marker — but in the degree of similarity between the word orders in the transitive and intransitive sentences of each language. Two factors seem to be involved: first, immediate adjacency of the verb to the Subject (Polinskaja 1989) boosts learning, because it allows the parser to have a single strategy for interpreting the "SV" unit in SV(O) and (O)SV, or the "VS" unit in VS(O) and (O)VS. The S(O)V and V(O)S types violate this principle because the presence of an Object in transitive clauses breaks up the Subject + verb unit. The second factor, linear order, is needed to explain why SVO and VSO are easier to learn than OSV and OVS, respectively. As shown in Table 10, there is, in terms of linear sequence of elements, a complete overlap of S and V between intransitive and transitive clauses only for SVO and VSO.[15] So, in VSO, V and S are always bound to occur in first and second position, respectively; VOS and SOV have one such element in common; and for OSV and OVS none of the three elements always appears in the same position.[16]

[approximate location of Table 10]

When asked to generalize to new words, and without the help of D-marking morphemes, the OSV and OVS networks are at a loss whether the first element of the sentence is a Subject or Object (OSV), or verb or Object (OVS). Even after they have seen the second word, the networks still have problems with the second word because they remember the first word as being ambiguous.

On the other hand, the simulation results in Table 9 also suggests that there are object-initial languages which should be as easy to learn as their corresponding subject-initial languages. In this regard, the networks fail to illuminate why object-initial languages are so rare. The main

reason for this deficiency is the absence of relevant semantic information. The network does not live in a world in which the concepts usually referred to by Subjects are more salient and prominent than Objects (for example Subjects are generally Agents: they move, have volition, and act upon inanimate and passive Objects/Patients — see Dowty 1991). So, unlike humans, the network has little reason to consider a description mimicking the flow of energy from Subject to Object more natural than the opposite one (Talmy 1988). Another factor may be the tendency for verbs and their objects to bind together, as illustrated by noun incorporation phenomena, idioms, and borrowings (Tomlin 1986). From the standpoint of verb-object cohesion, OSV (the least frequent Object-before-Subject language) and VSO (the least frequent Subject-before-Object language) fail to keep the Verb and Object adjacent. Finally, there are information-processing constraints at work as well: as Polinskaja (1989) has argued, across the board Object-before-Subject languages tend to violate the topic before focus principle.

In summary, the models find uncommon language types like OSV and OVS harder to learn. The results from the Subject-before-Object languages and the free word order language are also largely compatible with the patterns attested in natural languages. On the other hand, it appears that the networks presented here are only of limited use in explaining the infrequency of the Object-before-Subject languages. As long as our models do not include adequate semantic and pragmatic representations, certain phenomena will resist computer simulation.

4.1.2. *Ergative marking*

An obvious question one might have about the results presented so far is whether the same patterns also hold for languages which feature an absolutive/ergative case-marking. In ergative languages, the Subject of the intransitive clause and the Object of the transitive clause receive

identical (absolutive) markers, while a separate (ergative) marker is used for the Subject of the transitive clause. The example sentences in (7) are from Yidiny (Dixon 1980: 294).

(7) a. *yiŋu* *waguuja* *gali-ŋ*
this-ABS man-ABS go-PRES
"This man is going."

b. *mujaam-bu* *waguuja* *wawa-l*
mother-ERG man-ABS look at-PRES
"Mother is looking at the man."

Similarly for Head-marking, in which the same morphological element on the verb can refer to the Subject of an intransitive clause or to the Object. K'iche' is such a language as the sentences in (8) illustrate (Van Valin 1992: 19): notice that the morpheme *at* appears in the same place in both (8a) and (8b), whereas the morpheme *u* in (8b) refers to the Subject of the transitive clause.

(8) a. *X-at-war-ik*
TNS-2SG.ABS-sleep-SUFF
"You slept."

b. *K-at-u-ch'ay-o* *ri* *achi*
TNS-2SG.ABS-3ERG-hit-SUFF CLASS man
"The man hit you."

An obvious consequence of ergative marking is that it may (but does not have to) disturb the one-to-one mapping of case markers to grammatical relations. So, the question becomes how neural networks would perform on corpora with ergative case-marking and agreement. The next batch of simulations addressed this question. All parameters used for the accusative experiments were kept identical, but the grammars of the artificial languages were changed to generate ergative markers on the heads and dependents.

Table 11 contains the MSE scores for the three fixed word order languages with the Subject before the Object, and the free word order language. As one might expect, the XX values are of limited interest because these marking-less languages are basically identical to the marking-less accusative languages described above. The results are not perfectly identical, though, because the corpora contained slightly different sentences.

[approximate location of Table 11]

The first observation about Table 11 is that the results for all the fixed word order languages are basically identical to the ones in Table 8: performance on both the training and the test corpora is very close to perfection. What this shows is that these networks can perfectly well learn to disambiguate the absolutive case marker on the Subject of intransitive clauses by paying more attention to the strict word order pattern exhibited by all the SVO and VSO sentences. If the marker appears on the first word (SVO) or the second one (VSO), then it signals the grammatical Subject, notwithstanding the possible similarity of its case marker to that of the Object. For the SOV/XX network, we again find serious problems distinguishing between the Verb and the Object on the test corpora — a clear indicator that the network had to learn certain sentences in the training corpora by heart.

The results for the free word order language are more interesting, because it is easy to see that the XXX/XD and XXX/HD networks perform much worse here than they did on the accusative languages. These networks were at a loss because they lacked the fixed word order to help them out with the ambiguous — in terms of grammatical relations — absolutive nouns. It turns out that we find similar results for the Object-before-Subject networks (see Table 12), which also exhibit some variation when compared to the numbers in Table 9.

[approximate location of Table 12]

The average result in Table 12 is considerably worse when compared with the counterpart accusative language. All the OSV languages give the networks problems in learning and generalization, as do three of the VOS languages. In each case, it turns out that the models sometimes fail to distinguish between the Object and the Subject, and being uncertain about which one the input word is, they activate both output units partially. This is hardly surprising, because the input sentence is indeed ambiguous between the two readings: it might be that the absolutive noun is the Object of the transitive clause, but it could just as well be the Subject of the intransitive clause. Without a (non-)linguistic context to help them out, these networks are at a complete loss when confronted with such sentences. The one network which stands out here is VOS/HD: though it too has some problems with Subject and Object, it still does a fairly good job overall — on the test corpus, it fails to signal only 4% of the Subjects, and 9% of the Objects. The VOS/HD network can do this, because it sees the Verb first, and the markers on the Verb tell it whether or not the sentence has an Object. If it does, the first noun which follows the Verb must be this Object, despite the fact that its absolutive case marker is inherently ambiguous.

There are only minor problems for the OVS networks with ergative marking of some sort, and the problems which occur are related to failing to recognize the Verb on the test corpora. Again, it is not hard to see why the OVS order is more compatible with ergative marking: if the absolutive word occurs in sentence-initial position (OVS), it is the Object; if it occurs in the second position (VS), then it is the Subject.

In summary, if we assume a relationship between language types which are hard to learn for a network and language types which are infrequent in the real world, then the simulations with

ergative languages make clear predictions about the word order types which should co-occur with ergativity. While VSO, SVO, SOV, OVS and some free word order languages should definitely exist — all other things being equal — the models suggest that VOS and OSV languages with ergative marking should be more uncommon than their accusative counterparts. However, VOS languages with ergative case marking are actually widely attested among the Salish languages, the Mayan languages (England 1991) and in the Austronesian family (Chung 1998; Massam 2000), so further investigation is needed in this regard.[17] Still, the fact that the networks can learn at least some of the ergative language types is a positive finding, as is the observation that it has more problems with the Object-before-Subject languages than their Subject-before-Object counterparts.

4.2. *The grammar with genitives and locative phrases*

The goal of the second set of simulations was to use a more realistic grammar for creating the languages which the networks had to learn. So far, the artificial languages only contained Subjects, Objects, and verbs, and this inventory is far too small to account for the richness observed in natural languages. As a small step towards empirical adequacy, the context-free grammar was enriched with rules which could generate locative phrases and possessives. The former always took the form of a two-word sequence, adposition + noun (marked for genitive case in D-marking language types), while the latter were always single nouns (marked for locative case in D-marking languages). It should also be noted that all the languages in the following simulations exclusively use accusative markers, and that the fixed word order languages are limited to the ones in which the Subject precedes the Object. Both restrictions were inspired by the desire to limit the number of variables, and, hence, also the number of models to be run.

The precise way in which the addition of possessives and locative phrases to the grammar was implemented depended on the fixed word order of the language involved. As we have seen in the discussion of the word order correlation pairs, languages tend to be harmonic with respect to the relative orders of their heads and dependents. So, while the Head-initial VSO type co-occurs with prepositions and possessives which follow the noun they accompany, Head-final SOV is typically combined in a language with postpositions and possessives which precede the noun. The status of SVO is not as clear, but given its tendency to pattern along with VSO, the SVO grammar was enriched with prepositions. With regard to possessives, however, rules were introduced to generate both prenominal and postnominal possessives — a situation not unlike the one found in English (compare *John's father* and *the father of John*). A final property of the grammars which was intended to be realistic was that the position of the locative phrase was just before the Verb in the SOV languages, just after the Verb in VSO languages, and in sentence-initial position for the SVO languages. We can summarize these changes by the following phrase-structure rules in which optional elements are shown in parentheses:[18]

```
SOV:   s -> [ (poss) subj ] [ (poss) obj ] ([ location postposition ]) verb
VSO:   s -> verb ([ preposition location ]) [ subj (poss) ] [ obj (poss) ]
SVO:   s -> ([ preposition location ]) [ (poss) subj ] verb [ (poss) obj ]
       s -> ([ preposition location ]) [ subj (poss) ] verb [ obj (poss) ]
```

The results obtained with the new grammars are given in Table 13.

[approximate location of Table 13]

The most striking result in Table 13 is that only languages with Dependent-marking managed to perform really well on the training corpora. The generalization results for the test corpora obviously point in the same direction. A closer analysis of which errors the networks made show that they are somewhat different for each of the various basic word orders. In the

case of VSO, the XX and HX networks ran into problems trying to distinguish Objects and possessives. With both forms appearing without any case markers, and with possessives being optional, it is easy to see why these networks became confused — compare the following sentence templates:

```
# VSO ambiguity
s -> verb subj obj                (V_trans N N)
s -> verb subj poss               (V_intrans N N)
s -> verb subj poss obj           (V_trans N N N)
s -> verb subj obj poss           (V_intrans N N N)
```

The second noun in a sentence could be either the Object or a possessive accompanying the Subject noun, and the third noun was similarly ambiguous between being the Object or a possessive accompanying this Object. With no case morphemes, the only heuristic which the network could use was the class the noun involved belonged to in the training set, but the new words presented in the test corpora did not appear to belong to any class whatsoever. As for the SVO and SOV networks, they were affected by the same kind of ambiguities as their VSO counterparts, but in addition there was also structural confusion for possessives and Subjects.

Hence, this simulation warrants two conclusions: first, if a language has a case system with distinct morphemes for Subjects, Objects, possessives and locatives, then it can be learnt easily, independent of it having a fixed or free word order; second, languages in which possessives and — to a lesser degree — locatives appear without any marking whatsoever cannot be learnt flawlessly. But neither of these conclusions is terribly surprising, because the data available from cross-linguistic studies present similar findings. In particular, possessives do not usually seem to occur without some kind of extra morphological material announcing their presence (Plank 1980). On the other hand, cross-linguistic data also provide evidence for the following language universal (Greenberg 1963: 95).[19]

> *Universal 38.* When there is a case system, the only case which ever has only zero allomorphs is the one which includes among its meanings that of the subject of the intransitive verb.

So, the nominative in accusative languages and the absolutive in ergative languages may appear without an overt marker. Combining this finding with the other one specifying that a possessive should always remain marked somehow gives us sufficient cause to check the results of such changes on the artificial language.

For this batch of simulations, the grammar was therefore changed in two ways: first, the Subject marker was never generated on the noun (the Subject H-marking was left in place); second, the Genitive marker was no longer deleted in the XX and HX languages. The results obtained with the new grammars are presented in Table 14.

[approximate location of Table 14]

If we compare the new results with those of Table 13, we see that the performance of the XX and HX networks has improved dramatically, while that of the XD and HD nets has remained unaffected (but for XXX/XD — see below). In the case of SVO/HX we now even find slightly better MSE values than for the D-marking SVO/XD! It seems, then, that the two motivated changes to the grammar have the desired effect of bringing the behavior of the network closer to the patterns observed in natural languages. It turns out that a closer look at the HX languages reveals a single remaining problem for the fixed word order languages: i.e. the SOV/HX network still performs very poorly on recognizing Objects, both on the training and test corpora. The problem with identifying Objects is related to a structural ambiguity which is still present in the SOV languages without D-marking. Compare the two sentence templates below. The second noun in a clause can be either the Object or the locative noun, and the network cannot know for sure which one it is because it does not have access to the next word in the sentence. It partially

activates both the Object and Location output units, but this is still wrong as far as the error calculating algorithm is concerned.[20]

```
# SOV ambiguity
s -> subj obj verb          (N N V_trans)
s -> subj loc postp verb    (N N P V_intrans)
```

As mentioned above, Table 14 features one MSE value which is noticeably worse than the counterpart value in Table 13: for XXX/XD, the MSE score has increased from 0.0432 to 0.4281. This is the immediate result of the loss of nominative case-marking. With the earlier version of the grammar, the network was able to recognize new and unmarked verb stems in the free word order test corpora — no mean feat if one thinks about it — by applying a default strategy: if it was presented with a word which was not case-marked (i.e. the nouns) and which was not an adposition, then this word had to be the Verb of the clause. Without nominative case-marking, there are now two new unmarked phonological forms in every XXX/XD clause, the verb and the Subject. Consequently, the network can no longer apply its default strategy and therefore fails to distinguish between them. To the extent that XXX/XD languages occur in the real world, the failure of the network to account for them is a serious one. However, a case can be made that the situation is not as bleak for the simulations as one might think. The reason is that the grammars used for the artificial language are unrealistic in that there are no other markers on the verb than the ones used to co-refer to full nominals. However, there are many features which are cross-linguistically marked on the verb with a higher frequency than Subject or Object agreement markers: i.e. mood, tense, aspect and voice (Bybee 1985). If the verbs in the training and test corpora had carried markers for those features, the XXX/XD network would have been able to recognize these markers as indicators for the verb. As a consequence, it would no longer have confused the verbs with the unmarked Subjects, and the network's overall

performance would have been much better. So, the connectionist XXX/XD language type could probably be made learnable, just as "free" word order natural languages with case markers are learnable. The high MSE value obtained by the XXX/XD points out a way in which future simulations will have to be improved if they are to capture all the phenomena in natural languages.

In summary, the simulations presented in 4.2 have shown, first, that small recurrent neural networks are perfectly capable of learning to process transitive and intransitive sentences with locative phrases and possessives, provided that the latter are D-marked by an overt genitive affix. Second, we have also seen that arguments, primarily Subjects, do not necessarily have to carry an overt marker, as long as no structural ambiguities are created. Finally, the results from the free word order experiments suggest strongly that the artificial language has to be expanded to include other kinds of H-marking on the verb.

4.3. *The grammar with relative clauses*

The final set of simulations brings the grammar of the artificial language another small step closer towards the complexity of natural language. It extends the grammars from the second set further by adding rules which can generate relative clauses; the probability of any single sentence containing a relative clause was set at 50% to ensure that we would get clear results about how the introduction of relative clauses affected the models. For these models, we also kept the genitive case marker around in the XX and HX networks. To limit the number of simulations, we only investigated the Subject-before-Object accusative languages.[21]

Expanding the grammar used so far with relative clauses was not a straightforward task for there are many different types of relative clauses. Surveys of relative clause formation strategies

distinguish at least between relative clauses which either precede or follow the noun they accompany; between relative clauses with relative pronouns and those without; between relative pronouns with case and those without; and between relative clauses with resumptive pronouns and those without (Lehmann 1984; Keenan & Comrie 1977). For the purposes of the simulations, this plethora of possibilities was constrained by following two of the more frequent strategies for constructing relative clauses. First, for the SOV corpora the relative clauses were generated before the head noun, without a relative pronoun, and with the relative clause predicate occurring in a non-finite form. There are many natural languages which follow this pattern (Lehmann 1984). The example in (9) is from Tsez (Comrie & Polinsky 1999: 79).

(9) *[kid-b-er* *gagali* *tāλ-ru]* *uži* *qoqoλi-s*
girl-TH-DAT flower-ABS give-PASTPART boy-ABS laugh-PASTEV
"The boy who gave a flower to the girl laughed."

Second, the VSO and SVO corpora received relative clauses very much like those of English or French, with the relative clause following the head noun, an initial case-marked relative pronoun, and with the same kind of H-marking morphemes as on the main verb. The sentence in (10) is an example from French.

(10) *le garçon [qui avait donné une fleur à la fille] riait*
the boy who had given a flower to the girl laughed
"The boy who had given a flower to the girl laughed."

For all languages, relativization was limited to Subjects and Objects, but all possible combinations with transitive and intransitive main and subordinate verbs were implemented. The overall results for the networks learning on corpora with relative clauses are presented in Table 15.

[approximate location of Table 15]

A first general observation about Table 15 is that the MSE values have increased across the board.[22] The explanation for this overall degraded performance is that the task of the networks has become more complex, while the resources available to them (i.e. the number of hidden units and connections) have remained the same. On the other hand, networks like VSO/XD and SVO/HX demonstrate that it is still possible to learn more complex grammars than the ones discussed above. In that regard, the simulations show that at least certain language types can be modeled adequately by connectionist networks.

The SVO/XX network is plagued by a serious problem, at least on the test corpus. The difficulty is in recognizing the unmarked verb forms. It turns out that there is no confusion between the Main Verb and the Subordinate Verb, because the Clause Level output unit reveals that the network was almost always aware of which level the verb form occurred in. Instead, the network made seemingly random errors, suggesting that it was taxed beyond its representational capacities.

The MSE values for the four SVO networks are also interesting from another perspective: SVO/XD and SVO/HX score almost equally well, illustrating that the same kind of information about function and word category can be carried equally well by two different marking strategies. The combination of the two in SVO/HD does improve the MSE score, but only slightly. However, when neither type of marking is present, as in SVO/XX, the generalization performance of the network takes a steep dive as it has problems both with recognizing nouns and verbs. With English virtually being an SVO/XX language, one may wonder about the desirability of this network result. However, many English verbs *are* marked in some other way (e.g. through morphemes attached to the verb stem, modal verbs, or auxiliaries) and these

markers would presumably allow the network to recognize the Verb in each clause. The results for a better simulation of English would then be much closer to those for the SVO/HX network.

The SOV networks run into qualitatively similar, but quantitatively bigger problems as their VSO and SVO counterparts: i.e. SOV/HX performs very poorly on distinguishing Subjects, Objects and Locatives, with SOV/XD having a hard time detecting the verbs in the sentences. If neither type of marking is present, the consequences are catastrophic — the MSE value for SOV/XX is more than three times that of SOV/XD or SOV/HX — but when the network can use both, it performs reasonably well. As one might expect, the main reason why an SOV grammar with prenominal relative clauses proves harder to learn is that there are many options for structural ambiguities, especially when the nouns are unmarked. It is even the case that a sentence can begin with five different constituents:

```
# SOV ambiguity
(a) Subject of main clause:            subj obj verb
(b) Subject of relative clause:       [subj verb] subj obj verb
(c) Object of relative clause:        [obj verb] subj obj verb
(d) Locative in relative clause:      [loc postp verb] subj obj verb
(e) Subordinate verb:                 [verb] subj obj verb
```

Though eighty percent of the sentences are guaranteed to begin with a noun, the latter's correct interpretation will often depend on paying close attention to the rest of the sentence: i.e. the form of the first verb, as well as a possible prosodic pause separating the relative clause from the main clause. This means that parsing such sentences can create a relatively heavy processing load, for the connectionist networks as well as for human beings. As a consequence, it seems reasonable to posit that there is a connection between the ambiguities in sentences like those above, and the fact that more than one third of the world's SOV languages at least allow postposing of heavy constituents like relative clauses (Hawkins 1983) — this would obviously help with the disambiguation of the (a) and (b) sentence templates. It has also been mentioned

already that there is a very strong tendency for SOV languages to have a case-marking system — this makes it possible to distinguish between (b), (c) and (d). Finally, the use of participial verb forms within the relative clause could conceivably help with the correct processing of template (e) above. All of this suggests that the usual properties of relative clauses in SOV languages may not be as arbitrary as one might expect.

In summary, artificial languages with relative clauses in addition to locative phrases and possessives are learnable by neural networks. However, some combinations of properties, for example SVO/XX or SOV/HX, turn out to be much more problematic than others. In almost all cases, such degraded performance by a network can be linked to structural ambiguities which also appear to play a role in natural languages of similar types.

4.4. *Summary*

In this part, we have presented results from three sets of neural network simulations. The first one used corpora containing simple transitive and intransitive sentences with a subject, verb, and sometimes an object. All six possible fixed word orders as well as a completely free word order were simulated, and both the accusative and ergative case schemes were modeled with somewhat different results. The second set extended the grammar used to generate the artificial language with locative phrases (i.e. prepositions for VSO and SVO; postpositions for SOV) and possessives (prenominal in SOV; postnominal for VSO; both for SVO). It was found that the possessives had to be case-marked even in otherwise non D-marking languages. The final set of simulations expanded the grammar even further with the introduction of relative clauses whose linguistic properties were based on two types attested in natural languages (e.g. prenominal in SOV; postnominal for VSO, SVO).

5. General Discussion

This section addresses three major issues. First, we consider the evidence for a link between language type frequency and learnability. Second, we argue that the connectionist models described in this paper have been successful exactly because they suffer from the same processing and memory limitations which also affect young children. Finally, we discuss a number of deficiencies with the current simulations and propose ways in which to extend the models to overcome these problems.

5.1. *Frequency and learnability*

This paper set out to compare how different strategies for encoding grammatical relations would fare when a connectionist network had to learn them. The central hypothesis was that the simulations might display a preference for the same language types which have been found to be common among the world's languages. In particular, we claimed that a strong link between language type frequency and learnability required, first, that the neural network models be able to learn the attested language types; second, that they should fail to learn at least some of the unattested types; and, third, that the more common language types should be learnt better. We have seen that the first two requirements were largely met by our simulations: not only were the models able to learn almost all of the attested language types, independent of their word order or case marking type, but they also experienced obvious problems with many unattested or very rare types (e.g. SOV/XX, OSV/HX, XXX/XX). With respect to the third requirement, we have found that there is no consistent one-to-one mapping between learnability by our models and attested frequency. This finding does not come as a surprise, given the large number of other factors which can affect this mapping: on the one hand, the learnability of a language is obviously influenced by more than three linguistic parameters; on the other hand, the frequency

of language types must have been influenced by geographical boundaries and historical events (cf. Diamond 1999). Nonetheless, we have observed a number of interesting correspondences between how common certain linguistic properties are and how well the networks learning them performed.

With respect to basic word order frequencies, we have seen in the first set of simulations that there are two major linguistic observations which are reflected in how the networks learn the language types:

1. Object-before-Subject languages are far less frequent than Subject-before-Object languages.
2. Free word order languages are less frequent than fixed word order languages.

The models in general also have a much harder time learning the Object-before-Subject languages (generalization 1), as well as the free word order languages (generalization 2), even when they feature the same kind of H-marking or D-marking as their fixed word order Subject-before-Object counterparts.

With respect to the properties of case systems, there are several relevant linguistic generalizations which are reflected in the simulations:

3. Free word order languages typically occur with case systems (Siewierska 1998).
4. SOV languages typically occur with case systems (Greenberg's Universal 41).
5. If a language with a case system has no overt morpheme for a form, then this form can be the Subject of the intransitive clause (Greenberg's Universal 38).
6. The ergative marking system is not as wide-spread as the accusative one (Van Valin 1992; Siewierska 1996).

First, the XXX networks which were learning languages in which the sentences had Subject and Object D-marking on the nouns generally did a much better job than the ones which had to do without (generalization 3). Similarly, in all of the SOV simulations, SOV/XD outperformed SOV/XX and SOV/HX, often by a very wide margin (generalization 4). Third, the simulations in

which the nominative case marker was left unexpressed did not perform noticeably poorer on recognizing the Subject (generalization 5). Fourth, the results of the networks show that ergative marking does not perform as well as its accusative counterpart, both for language types with a fixed word order, as well as for the free word order type (generalization 6). This result is also compatible with the fact that there are very few, if any, languages which are entirely ergative whereas there are numerous languages which are consistently accusative.

With respect to the frequency of Head-marking versus Dependent-marking, there are two typological observations which are compatible with the simulations:

7. D-marking is more frequent than H-marking and has wider geographical distribution (Nichols 1986).
8. H-marking is especially common in V-initial languages.

First, the models with access to D-marking information scored consistently better than the ones without such information. For almost all of the simulations with possessives and relative clauses, the networks which could use D-marking scored better than the ones who could not. In the few cases in which this was not true, the differences were very small (generalization 7). Second, models learning V-initial languages generally benefited more from H-marking than their counterparts learning the other types. Still, we have also seen that otherwise D-marking languages could benefit from some H-marking to help locate the verb in a clause (generalization 8).

There are also several more general points concerning the trade-offs between syntax and morphology.

9. Word order and morphology can compensate for each other.
10. H-marking and D-marking can compensate for each other.
11. Redundant coding of information is not necessarily useful.

Generalization 9 turned out to be true in almost all cases, with the free word order language with D-marking (and H-marking) learning its task to nearly the same level of performance as the fixed word order languages. With respect to generalization 10, we have seen that for some of the networks, especially the VSO models learning sentences with relative clauses, there was no significant difference in performance between the XD and HX networks. In general, however, it does seem that D-marking is a safer bet than H-marking for encoding grammatical relations. Concerning generalization 11, at least for some of the networks, it has been shown that the HD model did not produce significantly better results than the HX or XD models. Among natural languages, the combination of a fixed word order, D-marking and H-marking in one and the same sentence is also very rare.

Other results worth noting about how the networks learned their tasks are given in generalizations 12 through 16:

12. Rote learning does not pay off when one has to parse new sentences (see 4.1.1).
13. Possessives benefit greatly from carrying a case-marker of some sort (see 4.2).
14. Default strategies can be developed for parsing sentences (see 4.2).
15. Frequent structural ambiguities in the input lead to impaired learning (see 4.3).
16. SOV language with relative clauses benefit from postposing them (see 4.3).

All of these results taken together suggest to us that the two main assumptions behind this paper are appropriate: encoding strategies for grammatical relations can be studied in artificial languages; and connectionist models are a useful tool for studying phenomena cross-linguistic phenomena. With respect to the hypothesis that there is a connection between language type learnability and frequency, we have observed that there are a number of interesting correlations. Consequently, we hope to have demonstrated that network simulations like ours can be used to

ground relatively vague functional notions about, for example, when information becomes redundant or when constructions are marked, in more solid experimental results.

5.2. *Connectionism and the child*

The positive correlations between the simulation results, on the one hand, and cross-linguistic variation, on the other, allow us to consider the possibility that the way our connectionist networks learn their tasks may be quite similar to the way children acquire their native languages. The models, like very young children, are subject to memory and processing limitations which affect how they can learn a language: both can only keep short sentences in active memory, both get confused when faced with ambiguous input, and both typically fail to reinterpret previously parsed words when new information is presented later on in a sentence (Bates *et al.* 1984; Trueswell *et al.* 1999). Finally, neither has access to the rich stores of semantic and pragmatic information which adults possess for their languages.[23] Because of these limitations, both network models and children appear to be influenced heavily by frequency effects, surface structure phenomena, transparency between constructions, and word order patterns (Akhtar 1999; Thal & Flores 2001).

Slobin & Bever (1982) conducted a language acquisition study which stands out in its relevance for the network results presented here. They did a comparative investigation into how four different languages are acquired, and their experiments targeted the links between grammatical relations on one side and the benefits of word order versus inflectional morphology on the other. Despite the areal bias of their sample (English, Italian, Turkish and Serbo-Croatian are all spoken in the same area of the world), Slobin & Bever found that there are significant

differences between them with respect to how fast children learn to understand *who did what to whom*. The experiments involved children in age groups from 2;0 to 4;4, who were asked to act out (with toy animals) an instruction containing two animate nouns and a reversible verb, e.g. *The squirrel the dog scratches.* For the English and Italian children, the variable parameters were word order (VNN, NVN, NNV) and prosody (stress on first or second noun). The Turkish and Serbo-Croatian children were presented with sentences varying again in word order, but this time the second dimension was the presence of (un)ambiguous nominal case markers. The overall results of the children on sentences which were grammatical in their languages are given in Table 16 (Slobin & Bever 1982: 241).

[approximate location of Table 16]

Though space considerations prevent us from discussing these results in depth, there are at least two phenomena in need of explanation: (i) why do the youngest Turkish infants perform so much better than their peers?; (ii) why do the older Serbo-Croatian children perform so badly when compared to their peers? Slobin & Bever (1982) find the answers to these two questions in how each of the languages use different cues for encoding their grammatical relations. Turkish, being a head-final language (SOV), allows a fair amount of word order flexibility at the level of the root clause (Ozkaragoz 1987; Kural 1997; Hoffman 1998). However, it is the only one of the four languages to have regular and unambiguous D-marking on each of the nouns (at least when they are definite, as in this experiment). A correct understanding of Serbo-Croatian, on the other hand, requires paying attention to both some word order tendencies as well as the elaborate — and at times ambiguous — scheme of case markers. As one might expect, the English-speaking

children have to learn to use word order exclusively, while in Italian the location of the prosodic stress in the clause helps the children out when the word order by itself is ambiguous. These psychological results, then, are compatible with the connectionist networks which we have discussed: we have also seen in section 4.1.1 that a consistent D-marking system can be sufficient for an XXX/XD language; on the other hand, a consistent fixed word order language like SVO/XX was also learnable, be it at a somewhat slower rate; but in those cases where the fixed word order became inconsistent (recall the SOV language with prenominal relative clauses), the networks ran into problems.

Given the difference in complexity between current connectionist networks and children, it would obviously be premature to think that we can already model language acquisition correctly. Still, the large parallels between the results do provide empirical support for neural network researchers interested in modeling human language acquisition. In their article, Slobin & Bever (1982) also argue that all children first become sensitive to what they call the "canonical sentence schemas" of a language, i.e. the strategies which it uses to encode a simple, active, affirmative and declarative sentence. These schemas later form the basis for the children's understanding of more complex sentence types. Consequently, one of the limitations of the models presented here, namely the use of very simple sentences, may not be so important after all, especially in the light of the success of Elman's (1993) "starting small" simulations. Along similar lines, our focus on learning word order patterns and morphological markers is supported by cross-linguistic and language acquisition research which shows that children learn these cues first for transitive sentences in which an animate agent causes a significant change of state in the patient (Slobin 1981; Tsunoda 1981). A proper understanding of such sentences requires the

capacity to assign noun phrases to their correct grammatical relations — i.e. the same task which our networks had to learn.

5.3. *Problems and issues*

Given how little research has been done on using connectionist simulations for examining typological phenomena, it will come as no surprise that there are several properties of our network models which do not stand up to close linguistic scrutiny. However, it appears to us that they do not imply fundamental problems with the concept of computational modeling of typological data — rather, they highlight areas in which our connectionist networks are not yet sophisticated enough to deal with the full complexity of natural language.

Probably the single largest deficiency of our simulations is that they do not include any kind of semantic, pragmatic or discourse-oriented information. For example, the well-known fact that Subjects tend to be human, agentive and topical is not represented in the model at all (e.g. Talmy 1985; Du Bois 1987). It is exactly the absence of this kind of information which is likely responsible for three problematic network results. First, the Object-before-Subject languages perform too well. Just from looking at the MSE values, one would indeed fail to predict, for example, the almost complete absence of OSV languages. But if we take into account that the well-attested tendency for Subjects to front is not reflected in the model, this result is no longer as problematic. Second, the VSO networks perform too well. If we were to go by the results of the model, we would expect VSO languages to be much more common than they actually are; they should definitely be more frequent than the SOV type, which is plagued by more learnability problems. Our take on this discrepancy is that VSO languages may really be as easy

to learn as the models suggest, but it is again the very strong cross-linguistic tendency for Subjects — and topics, for that matter — to appear in clause-initial position which makes them less frequent in the real world. Indeed, many VSO languages have SVO as a frequent alternative order (Siewierska 1998). Third, the SOV networks don't perform well enough. Of the three Subject-before-Object fixed word order languages, the SOV models consistently underperformed, often by a large margin. On the other hand, the SOV/XD and SOV/HD models usually did a fairly decent job of keeping up with the VSO and SVO networks. This happens to be exactly what we could have expected following Greenberg's conclusion about SOV languages, namely that they have case systems. Notice that no such strong claim has ever been made about either VSO or SVO languages. To the extent that individual models like SOV/XD and SOV/HD correspond to what language typologists have observed, the simulations should be considered successful, not deficient.[24] In general, we feel that given the number of things which the models do simulate (approximately) correctly without having access to semantic and pragmatic representations, one can only wonder how much they would benefit from having access to such rich sources of information.

This last criticism cuts both ways, though, for one could also argue that the network results show that there may not be as much semantics and pragmatics involved in some parts of language comprehension as often believed: i.e. if computer simulations can exhibit certain properties of natural languages without semantics, then it may also play a minor role for human beings. Still, we do think that the limited "semantic" information which the network did have access to (i.e. the fact that words fall into different distributional classes) is of crucial importance: without it (i.e. the XXX/XX network) the models are at a complete loss.

Another deficiency with the models used in this paper is that the languages they learn are overly simplified. Even for the task of learning to distinguish Subjects and Objects, more linguistic properties must play a role than word order, case-marking, and H-marking. For example, it has been found that children acquiring Sesotho will at times make use of gestures (Demuth 1987); speakers of Italian pay particular attention to prosody (Bates *et al.* 1982), as do some English aphasics (Ansell 1980); and both for Lisu (Li & Thompson 1976) and Klamath (Barker 1964) semantic information like the degree of animacy of the different nouns involved, selectional restrictions of the verb, as well as the extra-linguistic context may be all a hearer can use to disambiguate a sentence. For example, sentence (11) in Klamath has two interpretations, as does its counterpart with the two noun phrases transposed (Barker 1964: 341).

(11) *sníewe'ts hiswaqs sleʔa*
woman man saw
"The woman saw the man."
"The man saw the woman."

Still, Klamath is learnable by humans as are the attested free word order languages without D-marking (Siewierska 1998). So, it actually beyond doubt that there are many more than three linguistic parameters which influence the relationship between the learnability of a language type and its frequency. However, we do not feel that the limitations of the current models should be held against them. Rather, they give us a baseline for further modeling of more complex artificial languages which could feature more linguistic properties: e.g. personal pronouns, person and number agreement, markers for aspect, tense and mood, more correlation pairs, various splits between Head-marking and Dependent-marking, splits of ergativity and accusativity, more possible word orders, bigger lexicons, modals, causatives, and different types of verbs (e.g.

psych-verbs, unaccusatives, unergatives, ditransitives). The interaction between some of these features (i.e. personal pronouns, agreement, pro-drop, and split case systems) has been the focus of a large set of simulations which we have recently completed. The results of these simulations are still being analyzed, but we have already noticed interesting interactions between specific word orders and the presence of null forms. Adding a rudimentary semantics to the artificial languages is something we hope to implement in the near future. In general, one has to realize that every parameter which is added to the simulations multiplies the number of models to be run by the number of possible values of this parameter. Consequently, there are many linguistic parameters which deserve to be included in the simulations but which have to be left out for the time being.

Another concern with the simulations presented in this paper is that they have not demonstrated how actual languages would avoid becoming unlearnable. While some networks obviously had problems acquiring their language types, what we really need to show is that the process of language evolution can prevent such unlearnable types from ever coming into existence. Luckily, linguistic case studies of language contact and language loss phenomena provide us with rich data sets which can be used to set up new computer simulations. The connectionist work by Hare & Elman (1995) has already demonstrated that neural networks can provide insight into patterns of language evolution, so we see no reason to assume that such an extension of our models would be impossible to implement. As the result of imperfect learning between generations of language speakers, children might learn to produce the same forms they observe in their input, but only after they have reinterpreted them to follow different regularities. For example, what were preposition for one generation of speakers of Niger-Congo became

something closer to case markers for the next (Givón 1975). Simulating this kind of imperfect learning is currently still beyond our reach as it would imply a more complex language model than anyone has created so far, but we hope that future work will show how attested patterns of language change allow problematic languages to develop into more child-friendly variants. There is no reason to believe that such changes would happen in a fast or efficient manner, but the use of computer models makes it possible to explore many alternative scenarios in parallel.

In summary, there is no question that the models presented above are linguistically and typologically too simple. However, connectionist simulations and typology have not been brought together in a consistent fashion, and this is an area where one has to "start small". Taking into account both the dearth of research done in this area and the number of positive results obtained in this study, we do not think that the simplicity of the model is fatal at this time, because it is possible to extend the network models to address the current deficiencies.

6. Conclusion

In this paper we have shown that applying connectionist modeling to cross-linguistic variation and language universals can be a fruitful enterprise. In many cases, we have seen that the results produced by the networks are compatible with what is known about the languages of the world. Analyzing the behavior of the models — especially the areas in which they experience problems learning a language type — can give insight into why we find the distributions of correlated linguistic properties which we do. We need to reassess the initial hypothesis that there may be a strong correlation between the frequency of a language type, on the one hand, and its frequency, on the other. It appears that we can only answer with a (cautious) "yes" to two of the three questions formulated earlier: namely, whether the neural networks are able to learn the attested language types, and whether they fail to learn the unattested types. However, we have also seen that there is no overwhelming evidence that the more frequent types are consistently learnt faster and/or better than their less common counterparts. This finding supports the hypothesis that historical events and geographical boundaries often account for the frequency patterns which we observe today (cf. Comrie 1989; Dryer 1989).

The models allow us to give experimental support to some of the functional hypotheses which have been proposed in typological literature. They also make it possible for us to investigate why certain language types are unattested — is their absence likely the result of historical accident or are there good reasons why human beings would be unable to learn such languages. The ability to systematically explore these gaps would be hard to achieve without some type of simulations, and connectionist networks appear to be well-suited for the task (cf. Christiansen & Devlin 1997). Finally, we have seen a number of times that the results produced

by the models can lead to the formulation of essentially linguistic hypotheses about which combinations of properties languages should possess. This proves that there is a potential for future interaction, especially as connectionist simulations become more sophisticated.

Acknowledgments
I am indebted to four anonymous reviewers, Chris Barker, Liz Bates, Gary Cottrell, Jeff Elman, Andy Kehler, Ron Langacker, Margaret Langdon, and especially Maria Polinsky for helpful feedback on earlier versions of this paper. All remaining errors are my sole responsibility.

References

Akhtar, Nameera (1999). Acquiring basic word order: evidence for data-driven learning of syntactic structures. Journal of Child Language 26: 339-356

Ansell, Barbara (1980). Aphasic Adults' Use of Morphemic and Word Order Cues for Sentence Processing. Doctoral dissertation, University of Washington

Barker, M.A.R. (1964). Klamath Grammar. Berkeley: UC Press

Bates, Elizabeth, Sandra McNew, Brian MacWhinney, Antonella Devescovi & Stan Smith (1982). Functional constraints on sentence processing: A cross-linguistic perspective. Cognition 11.3: 245-299

Bates, Elizabeth, Brian MacWhinney, Cristina Caselli, Antonella Devescovi, Francesco Natale & Valeria Venza (1984). A cross-linguistic study of the development of sentence interpretation strategies. Child Development 55: 341-354

Bates, Elizabeth & Judith Goodman (1999). On the emergence of grammar from the lexicon. In Brian MacWhinney (ed.), The Emergence of Language, 29-79. Hillsdale, NJ: Lawrence Erlbaum

Berko-Gleason, Jean (1982). Insights from Child Language Acquisition for Second Language Loss. In Richard D. Lambert & Barbara F. Freed (eds.), The Loss of Language Skills, 13-23. Rowley: Newbury House

Birner, Betty J. & Gregory Ward (1998). Information status and noncanonical word order in English. Amsterdam: John Benjamins

Burgess, Curt (1998). From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. Behavior Research Methods, Instruments, & Computers 30: 188-198

Bybee, Joan L. (1985). Morphology. A Study of the Relation between Meaning and Form. Amsterdam: John Benjamins

Bybee Hooper, Joan L. (1980). Child morphology and morphophonemic change. In Fisiak 1980, 157-187

Christiansen, Morten H. & Joseph Devlin (1997). Recursive Inconsistencies Are Hard to Learn: A Connectionist Perspective on Universal Word Order Correlations. In Proceedings of the 19th Annual Cognitive Science Society Conference, 113-118. Mahwah, NJ: Lawrence Erlbaum

Chung, Sandra (1998). The design of agreement: Evidence from Chamorro. Chicago: University of Chicago Press

Comrie, Bernard (1989). Language Universals and Linguistic Typology. Chicago: University of Chicago Press

— (1993). Language universals and linguistic typology: data-bases and explanations. Zeitschrift für Sprachtypologie und Universalienforschung 46.1: 3-14

Comrie, Bernard & Maria S. Polinsky (1999). Form and Function in Syntax: Relative Clauses in Tsez. In Michael Darnell, Edith Moravcsik, Frederick J. Newmeyer, Michael Noonan & Kathleen Wheatley (eds.), Functionalism and Formalism in Linguistics, 77-92. Amsterdam: John Benjamins

Croft, William (2000). Explaining Language Change. Harlow: Pearson

Demuth, Katherine (1987). Discourse functions of word order in Sesotho acquisition. In Tomlin 1987, 91-108

Derbyshire, Desmond C. (1985). Hixkaryana and Linguistic Typology. Arlington, TX: Summer Institute of Linguistics

Diamond, Jared (1999). Guns, Germs, and Steel: The Fates of Human Societies. New York, NY: W.W. Norton

Dixon, Robert M.W. (1980). The Languages of Australia. Cambridge: Cambridge University Press

Dowty, David R. (1991). Thematic proto-roles and argument selection. Language 67.3: 547-619

Dryer, Matthew S. (1988). Object-Verb Order and Adjective-Noun Order: Dispelling a Myth. Lingua 74: 185-217

— (1989). Large Linguistic Areas and Language Sampling. Studies in Language 13.2: 257-292

— (1992). The Greenbergian Word Order Correlations. Language 68.1: 81-138

— (1997). On the six-way word order typology. Studies in Language 21.1: 69-103

— (1998). Aspects of word order in the languages of Europe. In Siewierska 1998, 283-319

— (2000). Counting genera vs. counting languages. Linguistic Typology 4: 334-356

Du Bois, John W. (1987). The Discourse Basis of Ergativity. Language 63.4: 805-855

Elman, Jeffrey L. (1990). Finding structure in time. Cognitive Science 14: 179-211

— (1992). Grammatical Structure and Distributed Representations. In Steven Davis (ed.), Connectionism: Theory and Practice, 138-178. New York, NY: Oxford University Press

— (1993). Learning and development in neural networks: the importance of starting small. Cognition 48: 71-99

England, Nora C. (1991). Changes in Basic Word Order in Mayan Languages. International Journal of American Linguistics 57.4: 446-486
Fisiak, Jacek, ed. (1980). Historical Morphology. The Hague: Mouton de Gruyter
Fletcher, P.C., J.M. Anderson, Shanks D.R., R. Honey, T.A. Carpenter, T. Donovan, N. Papadakis & E.T. Bullmore (2001). Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. Nature Neuroscience 4: 1043-1048
Foley, William A. (1991). The Yimas Language of New Guinea. Stanford, CA: Stanford University Press
Gasser, Michael (1993). Learning Words in Time: Towards a Modular Connectionist Account of the Acquisition of Receptive Morphology. TR-384, Computer Science, Indiana University
Givón, Talmy (1975). Serial Verbs and Syntactic Change: Niger-Congo. In Charles N. Li (ed.), Word Order and Word Order Change, 47-112. Austin, TX: University of Texas Press
Greenberg, Joseph H. (1963). Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In Joseph H. Greenberg (ed.), Universals of Language, 73-113. Cambridge, MA: MIT Press
Hare, Mary & Jeffrey L. Elman (1995). Learning and Morphological Change. Cognition 56: 61-98
Harris, Alice C. & Lyle Campbell (1995). Historical Syntax in Cross-linguistic Perspective. Cambridge: John Benjamins
Hawkins, John A. (1983). Word Order Universals. New York, NY: Academic Press
— (1988). Explaining Language Universals. In John A. Hawkins (ed.), Explaining Language Universals, 3-28. Oxford: Basil Blackwell
— (1993). Heads, parsing and word-order universals. In Greville G. Corbett, Norman M. Fraser & Scott McGlashan (eds.), Heads in Grammatical Theory, 231-265. Cambridge: Cambridge University Press
— (1994). A Performance Theory of Order and Constituency. Cambridge: Cambridge University Press
— (1999). Processing complexity and filler-gap dependencies across grammars. Language 75.2: 244-285
Hewitt, B. G. (1979). The relative clause in Abkhaz (Abzui dialect). Lingua 47: 151-188
Hoffman, Beryl (1998). Word Order, Information Structure, and Centering in Turkish. In Marilyn A. Walker, Aravind K. Joshi & Ellen Prince (eds.), Centering Theory in Discourse, 253-271. Oxford: Clarendon Press
Holm, John (1988). Pidgins and Creoles. Cambridge: Cambridge University Press
Howell, Steve R. & Suzanna Becker (2001). Modelling language acquisition: Grammar from the lexicon? In Proceedings of the 23rd Annual Cognitive Science Society Conference, Mahwah, NJ: Lawrence Erlbaum
Jespersen, Otto (1922). Language: Its Nature, development and origin. London: Allen & Unwin
Katz, B.F. & M.H. Dorfman (1992). The Neural Dynamics of Conversational Coherence. In Andy Clark & Rudi Lutz (eds.), Connectionism in Context, 167-181. London: Springer-Verlag
Keenan, Edward L. & Bernard Comrie (1977). Noun Phrase Accessibility and Universal Grammar. Linguistic Inquiry 8.1: 63-99
Kirby, Simon (1997). Competing motivations and emergence: explaining implicational hierarchies. Linguistic Typology 1: 5-32
Kural, Murat (1997). Postverbal Constituents in Turkish and the Linear Correspondence Axiom. Linguistic Inquiry 28.3: 498-519
Lambrecht, Knud (1994). Information Structure And Sentence Form: Topic, Focus, and The Mental Representations of Discourse Referents. Cambridge: Cambridge University Press.
Langacker, Ronald W. (1993). Reference-point constructions. Cognitive Linguistics 4.1: 1-38
Langdon, Margaret (1966). A Grammar of Diegueño: The Mesa Grande Dialect. Doctoral dissertation, U.C. Berkeley
Lehmann, Christian (1984). Der Relativsatz: Typologie seiner Strukturen, Theorie seiner Funktionen, Kompendium seiner Grammatik. Tübingen: G. Narr
Li, Charles N. & Sandra A. Thompson (1976). Subject and topic: A new typology of language. In Charles N. Li (ed.), Subject and Topic, 458-489. New York, NY: Academic Press
MacWhinney, Brian & Elizabeth Bates, eds. (1989). The crosslinguistic study of sentence processing. Cambridge: Cambridge University Press
Maslova, Elena (2000). A dynamic approach to the verification of distributional universals. Linguistic Typology 4: 307-333

Massam, Diane (2000). VSO and VOS: Aspects of Niuean word order. In Andrew Carnie & Eithne Guilfoyle (eds.), The syntax of verb initial languages, 97-116. Oxford: Oxford University Press
Mithun, Marianne (1987). Is basic word order universal. In Tomlin 1987, 281-328
Morris, William C. (1998). Emergent Grammatical Relations: An Inductive Learning System. Doctoral dissertation, UC San Diego
Newmeyer, Frederick J. (2000). On the reconstruction of 'Proto-World' word order. In Chris Knight, Michael Studdert-Kennedy & James R. Hurford (eds.), The Evolutionary Emergence of Language: Social function and the origins of linguistic form, 372-388. Cambridge: Cambridge University Press
Nichols, Johanna (1986). Head-marking and dependent-marking grammar. Language 62.1: 56-119
Ozkaragoz, Inci Zuhra (1987). The Relational Structure of Turkish Syntax. Doctoral dissertation, UC San Diego
Plank, Frans (1980). Encoding grammatical relations: acceptable and unacceptable non-distinctness. In Fisiak 1980, 289-325
Polinskaja, Maria S. (1989). Object initiality: OSV. Linguistics 27: 257-303
Polinsky, Maria S. & Ezra Van Everbroeck (2000). Development of gender classifications: Modeling the historical change from Latin to French, manuscript
Pullum, Geoffrey K. (1981). Languages with object before subject: a comment and a catalogue. Linguistics 19: 147-155
Rijkhoff, Jan & Dik Bakker (1998). Language Sampling. Linguistic Typology 2: 263-314
Rumelhart, David E., Geoffrey E. Hinton & Ronald J. Williams (1986). Learning internal representations by error propagation. In David E. Rumelhart, James L. McClelland & the PDP Research Group (eds.), Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Volume 1: Foundations, 318-362. Cambridge, MA: MIT Press
Rumelhart, David E. & James L. McClelland (1986). On Learning the Past Tenses of English Verbs. In James L. McClelland, David E. Rumelhart & the PDP Research Group (eds.), Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models, 216-271. Cambridge, MA: MIT Press
Sapir, Edward (1921). Language. An Introduction to the Study of Speech. San Diego: Harcourt Brace
Sejnowski, Terry J. & Charles R. Rosenberg (1986). NETtalk: a parallel network that learns to read aloud. In James A. Anderson & Edward Rosenfeld (eds.), Neurocomputing. Volume 1, 663-672. Cambridge: MIT Press
Sgall, Petr (1995). Prague School Typology. In Masayoshi Shibatani & Theodora Bynon (eds.), Approaches to Language Typology, 85-144. Oxford: Clarendon Press
Siewierska, Anna (1988). Word Order Rules. London: Croom Helm
— (1996). Word order type and alignment type. Zeitschrift für Sprachtypologie und Universalienforschung 49.2: 149-176
— (1998). Variation in major constituent order: a global and a European perspective. In Siewierska 1998, 475-551
— ed., Constituent Order in the Languages of Europe. Berlin: Mouton de Gruyter
Slobin, Dan I. (1981). Introduction: Why Study Acquisition Crosslinguistically? In Dan I. Slobin (ed.), The Crosslinguistic Study of Language Acquisition. Volume 1: The Data, 3-24. Hillsdale, NJ: Lawrence Erlbaum
— ed. (1992). The Crosslinguistic Study of Language Acquisition. Volume 3. Hillsdale, NJ: Lawrence Erlbaum
Slobin, Dan I. & Thomas G. Bever (1982). Children use canonical sentence schemas: A crosslinguistic study of word order and inflections. Cognition 12: 229-265
Smith, Philip T. (1995). Are Morphemes Really Necessary? In L. Feldman (ed.), Morphological Aspects of Language Processing, 365-382. Hillsdale, NJ: Lawrence Erlbaum
Steele, Susan (1978). Word Order Variation: A Typological Study. In Joseph H. Greenberg (ed.), Universals of Human Language, 585-624. Stanford, CA: Stanford University Press
Talmy, Leonard (1988). Force Dynamics in Language and Cognition. Cognitive Science 12: 49-100
Thal, Donna & Melanie Flores (2001). Development of sentence interpretation strategies by typically developing and late-talking toddlers. Journal of Child Language 28: 173-193
Tomlin, Russell S. (1986). Basic Word Order. Functional Principles. London: Croom Helm
— ed. (1987). Coherence and Grounding in Discourse. Amsterdam: John Benjamins

Trueswell, John C., Irina Sekerina, Nicole M. Hill & Marian L. Logrip (1999). The kindergarten-path effect: studying on-line sentence processing in young children. Cognition 73: 89-134
Tsunoda, Tasaku (1981). Split Case-Marking Patterns in Verb-Types and Tense/Aspect/Mood. Linguistics 19: 389-438
Van Valin, Robert D. (1992). An Overview of Ergative Phenomena and Their Implications for Language Acquisition. In Dan I. Slobin (ed.), The Crosslinguistic Study of Language Acquisition. Volume 3, 15-37. Hillsdale, NJ: Lawrence Erlbaum
Vennemann, Theo (1975). An Explanation of Drift. In Charles N. Li (ed.), Word Order and Word Order Change, 269-305. Austin, TX: University of Texas Press

[1] These terms come with much linguistic baggage, but we will be using them in a fairly intuitive and atheoretical manner. Our computer simulations use artificial languages based on simple grammars, so the more abstract labels NP1 and NP2 could be considered more appropriate. The task of disambiguating between the two NPs remains the same, though, whichever names one gives them. Given that the artificial languages have no semantics whatsoever, the thematic roles of Agent and Patient fail to apply too — see Morris (1998) for a connectionist simulation which does look at the acquisition of thematic roles.

[2] However, many languages do not force explicit coding of the Subject and Object in all environments (Plank 1980). The resulting ambiguity is typically only theoretical because contextual or semantic information will assist the hearer in choosing between, for example, "The man ate a burrito" and "A burrito ate the man". In addition, clauses in which both the Subject and the Object are expressed by lexical noun phrases are rare in many languages (Du Bois 1987), in effect avoiding many ambiguities altogether.

[3] As pointed out by an anonymous reviewer, "[m]any of the initial attempts to model linguistic issues, such as the much-touted Rumelhart and McClelland [1986] simulation of English past-tense learning, were so seriously flawed from a linguistic and psycholinguistic point of view as to be practically useless." We feel that our work does not suffer from the same flaws, both because we have taken care to avoid the mistakes made in those simulations (see Fodor & Pylyshyn 1988 and Pinker & Prince 1988 for discussions of the main problems found in early connectionist research), but more importantly because we have let the cross-linguistic data drive the modeling: for example, we have looked at both accusative and ergative case-marking systems because both are attested in the languages of the world.

[4] We would like to point out that infants who have only been acquiring language for a limited time are unlikely to have the same level of articulated semantic information available to them which adults take for granted. Similarly, personal pronouns do not always feature prominently in the initial language output of young children. In both these areas, then, our connectionist simulations may actually be somewhat better models of unsophisticated language learners than appears to be the case at first sight.

[5] We do not necessarily expect our models to fail to learn *all* the unattested language types, because some of them might be unattested for historical reasons (i.e. an accidental gap) as opposed to more fundamental problems, e.g. because they are unlearnable.

[6] The numbers do not add up to 100% because for some languages there are not enough data available, and others have continued to defy characterization into a single type (e.g. Dutch, which has SVO in main clauses, but SOV in subordinate clauses, as well as all "free" word order languages). Also, the existence of languages with a basic order of OSV is still a topic of debate.

See Pullum (1981) for a number of candidates and Polinskaja (1989) for a critical look at these languages. Finally, Dryer (1997) has suggested that the six-way word order typology should be replaced with a binary one: OV vs VO.

[7] Hawkins (1993, 1994) turns the requirement of uniqueness into an axiom, suggesting that we would otherwise not be able to determine which phrasal category to construct. As will become clear in the discussion of how our neural networks parse language, there is actually no need for any single category to predict a unique phrasal category *as long as the context in which this category occurs disambiguates between the various phrasal categories associated with it* (cf. Elman 1992).

[8] More detailed information about how the simulations were implemented — the grammar files, the phonological and syllabic constraints, and the results for each output unit — can be found on-line at http://ling.ucsd.edu/~ezra/lintyp/ .

[9] In more recent simulations, we have replaced the phonological input representation with a much shorter variant in which words are represented by random, but unique strings of 0's and 1's. The results presented below are not affected much by this change, but the models train faster because there are fewer units and therefore fewer connections. The use of a linguistically plausible phonology is obviously still desirable in principle because it increases the chances of obtaining results which are linguistically motivated.

[10] It is important to mention that some linguistic knowledge was built into the models. For example, they did not have to discover for themselves how to represent the phonology of the artificial language at the input layer, or which word classes to distinguish at the output layer. Especially for the latter, it would be desirable to have a self-organizing network which would avoid any preconceived set of valid linguistic notions. On the other hand, it is not clear how to construct such a network, especially in the absence of any lexical semantic information which would provide the network with clues as to which words typically fall into which classes (e.g. nouns which refer to places are often part of locative phrases). The work by Burgess (1998) is a step in the right direction, but since it is based solely on English corpora, including it in a typological model would raise numerous other problems.

[11] We selected 30 training cycles because training for more epochs did not lead to noticeably better results on the training or test corpora. Also, the use of larger corpora was not pursued because tests showed no substantially different results for corpora with 5,000 or 10,000 sentences. It is also worth mentioning that the networks were only trained on a small number of the possible sentences of their languages because the grammars generated more than one million possible sentences for each language type: the simplest lexicon contained 150 animate nouns, 150 inanimate nouns, 50 transitive verbs and 50 intransitive verbs. Hence, even ignoring the 30% chance that an inanimate noun was chosen for the Subject or an animate noun for the Object (cf. Du Bois 1987), there were already 1,125,000 (150x50x150) possible transitive sentences, and 7,500 (150x50) possible intransitive sentences. The ratio of transitive to intransitive sentences in

the corpora was not as lopsided as these numbers suggest. Each sentence generated had an equal probability of being transitive or intransitive.

[12] The desired output values which the network had to produce were not 0 and 1, but rather 0.1 and 0.9. There are two reasons for this slight modification: first, absolute value like 0 and 1 are impossible to attain for a system which uses the standard logistic activation function (Rumelhart, Hinton & Williams 1986), so a lot of learning time can be wasted while the network tries to move its output on a unit from, for example, 0.000003 to absolute 0. Second, both 0.1 and 0.9 allow the network to err on the safe side a bit by signaling for example 0.05 and 0.95. This again makes learning easier, without negatively affecting the network's overall performance.

[13] The MSE is derived from the SSE (i.e. Sum of Squared Errors) by dividing the SSE by the number of patterns (i.e. the words and sentence-final periods) in the corpus. The SSE can in turn be calculated by finding the differences for each pattern between the desired activation values of each of the output units and the actual value produced by the network for these units; then squaring all these differences (to avoid that positive and negative errors would cancel each other out); and finally summing all the squared values. Though the exact MSE numbers in the tables are not too important — they depend to some extent on the random values which were initially assigned to each of the connections — differences of at least one order of magnitude typically are. So, in Table 8 the difference between 0.0001 (SVO/XX) and 0.0004 (XXX/HD) is not significant, but the difference between 0.0003 (SOV/XD) and 0.0031 (SOV/XX) is.

[14] Greenberg (1963) refers to a "dominant" rather than "basic" word order, but we will assume that the two are interchangeable for our purposes. See also section 2.1 above.

[15] It is evident that in more complex sentences, the absolute position of a constituent may not be as stable as it is here: e.g. even in SVO sentences, a fronted topicalized element may "push" the Subject into the second linear slot. However, there is some evidence that infants may be mostly ignoring such non-canonical sentences when they are learning about the basic word order (if any) of their language (see the discussion of Slobin & Bever 1982 below). Elman (1993) has presented connectionist results which are compatible with this hypothesis: he found that recurrent neural nets with a limited capacity were confused by complex sentences and based their generalizations on simple clauses instead. The fact that parts of long sentences turned into noise for the limited network was actually a positive factor, because "the increased variability may retard learning and keep the network in a state of flux until it has enough data to make the reasonable approximations at the true generalizations" (95).

[16] Though we haven't run any simulations to verify its importance for our models, there is probably a third principle involved as well: i.e. the strong tendency for languages to have the Subjects of intransitive clauses and Subjects of transitive clauses appear on the same side of the verb (both to the left or both to the right), at least in neutral contexts. A violation of this principle would be a language which has SVO in transitive clauses and VS in intransitive ones. Such a language would obey the first two principles, but the fact that there appear to be few natural

languages which have this kind of word order split suggests to us that they may be harder to acquire than languages in which the Subjects always appear on the same side of the verb. To our knowledge, such languages are not attested.

[17] We have preliminary results of a new set of simulations using a different network architecture which show that ergative VOS languages are definitely learnable by connectionist networks. We hope to report on these results in the near future.

[18] Because of the optional character of the new additions, the new SOV grammar could still generate basic SOV sentences, in addition to more complex ones with possessives or locative phrases. For both Subject and Object NPs, the probability that it would now contain a possessive NP as well was 50%. Similarly, there was a 50% chance that a locative phrase would be generated in any sentence. These probabilities are higher than what one will typically find in natural language but they also make it easier for us to determine how the presence of these more complex elements affects the learnability of the various language types.

[19] These are not absolute regularities. Papago, for example, can express possession via juxtaposition of two noun phrases, though it also allows the use of a marker to signal the relationship between them (Langacker 1993; see also Linguist list message 12.1577 by Joost Kremers). It appears, though, that there are very few languages — even creoles with their impoverished morphology (Holm 1988) — which are able to express possession only via mere juxtaposition. Next, it has been argued that Diegueño, like most other Yuman languages, has an overt marker for the Subject, but not for the Object (Langdon 1966).

[20] Obviously, the network can be said to be doing "the right thing" by hedging its bets until it has received more information. It is a limitation of the output representation used that it is impossible for the network to signal that it has figured out the correct interpretation of a word which occurred earlier in the sentence. We have been addressing this issue in a new set of simulations in which the output representation gradually builds up a representation of the entire sentence, so that reinterpretations of previously seen words and phrases can be investigated.

[21] Note that the networks still had to learn their languages from scratch and that a "starting small" approach in which the networks are gradually exposed to more complex input sentences might lead to different results (Elman 1993). Such an investigation would merit a paper to itself, though.

[22] As one might have expected the numbers in Table 11 go down if the grammars with relative clauses also include resumptive pronouns in the gap position. Keenan & Comrie (1977) already suggested that resumptive pronouns typically appear in relative clauses to make processing easier. Hence, the neural networks appear to follow the human data here too. The reason is that these pronouns make the word orders of main clauses and relative clauses become more similar and this in turn makes it easier to develop a single strategy for analyzing clauses in such languages.

[23] Obviously, young children have access to much more semantic and conceptual information than the syntax-only connectionist models presented in this paper. And it is exactly this knowledge that can let children (Bates & Goodman 1999) or neural networks (Howell & Becker 2001) bootstrap into syntax and morphology. The point we would like to make here, though, is that infants who have only just begun acquiring their native language(s) can not yet be assumed to have a good understanding of, for example, the selectional restrictions of verbs or the connotational differences between "synonyms". It is important not to lose track of the qualitative and quantitative differences in linguistic knowledge between the early stages of language acquisition and adult competence.

[24] The existence of a language like Abkhaz (Hewitt 1979), which appears to be SOV/HX, remains problematic for the model. On the other hand, it has recently been argued that SOV may not be the preferred word order which its frequency would lead one to believe. Newmeyer (2000) claims that there are no attested cases of languages changing their word order to SOV (except where language contact was involved), whereas there have been many instances of SOV languages which have evolved to a different word order.

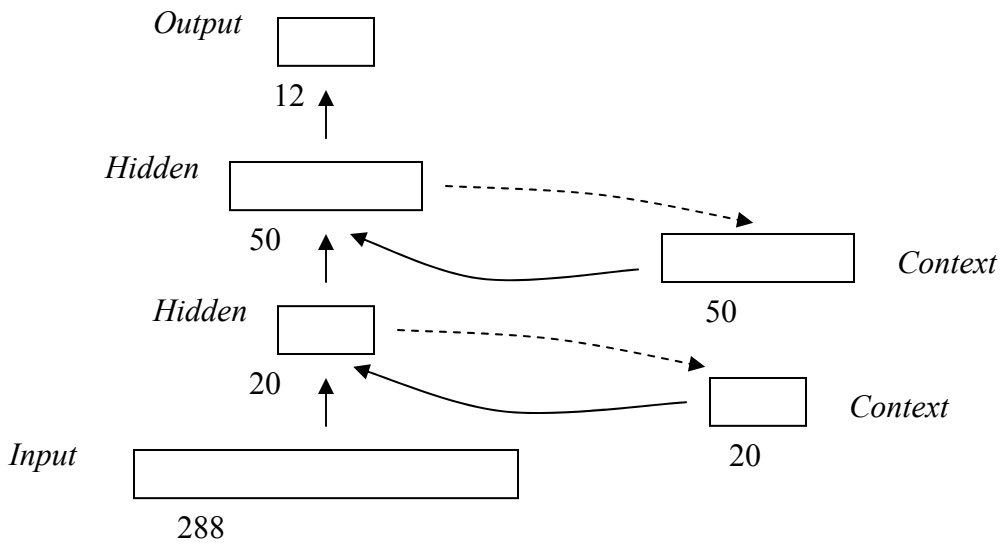

Figure 1. *The architecture of the network: 288 input units, which contain a phonological representation of a single word, feed into a small hidden layer with 20 units which in turn sends activitation to a larger hidden layer. Both hidden layers have context units to act as short-term memory. The second hidden layer then feeds into the output layer. It contains seven units to signal the functional class the input word belongs to, four units to indicate the word class, and one unit to signal whether the input word belongs to the main clause or a subordinate one. Full lines between layers represent full interconnectivity; dashed lines indicate that there is a perfect one-to-one copy from units in the source layer to their counterparts in the context layer.*

Table 1. *Estimated frequencies of language types with basic word orders involving Subject, Object and Verb.*

SOV	**SVO**	**VSO**	**VOS**	**OVS**	**OSV**
51%	23%	11%	8%	0.75%	0.25%

Table 2. *Number of language families which combine a word order like VSO with another property like prepositions.*

Correlation	**VSO**	**SVO**	**SOV**
Prepositions	31	42	5
Postpositions	4	6	97
Noun — Genitive	13	9	7
Genitive — Noun	3	12	24
Noun — Relative Clause	8	12	15
Relative Clause — Noun	0	1	10
Noun — Adjective	9	13	21
Adjective — Noun	9	8	22

Table 3. *The mnemonic system for referring to network models.*

Language	**Fixed word order**	**Head-marking**	**Dependent-marking**
SOV/HD	✓	✓	✓
SOV/HX	✓	✓	—
SOV/XD	✓	—	✓
SOV/XX	✓	—	—
XXX/HD	—	✓	✓
XXX/HX	—	✓	—
XXX/XD	—	—	✓
XXX/XX	—	—	—

Table 4. *A sample sentence in each of the four possible SOV languages.*

Language	**Subject**	**Object**	**Verb**
SOV/HD	pedi-uχ	splov-Emt	moχE-on-ak
SOV/HX	pedi	splov	moχE-on-ak
SOV/XD	pedi-uχ	splov-Emt	moχE
SOV/XX	pedi	splov	moχE

Table 5. *The twelve output units of the network can be divided into three distinct groups. The first group indicates the functional class the input word belongs to (e.g. Subject vs Object). The second group signals the category of the input word (e.g. Noun vs Verb). The last unit indicates that the input word belongs to a subordinate clause.*

Functional class	**Word category**	**Level**
1. Subject	1. Noun	1. Subordinate clause
2. Object	2. Verb	
3. Possessive	3. Adposition	
4. Locative	4. Pronoun	
5. Main Verb		
6. Subordinate Verb		
7. Relative Pronoun		

Table 6. *The possible configurations for the first eleven output units. Each configuration represents a possible desired pattern of activation for a word of an input sentence.*

Pattern	**Linguistic counterpart**
1000000 1000	Subject noun
0100000 1000	Object noun
0010000 1000	Possessive noun
0001000 1000	Location noun
0001000 0010	Locative adposition
0000100 0100	Main verb (always in main clause)
0000010 0100	Subordinate verb (always in subordinate clause)
0000001 0001	Relative pronoun (always in subordinate clause)
1000000 0001	Resumptive Subject pronoun (always in subordinate clause)
0100000 0001	Resumptive Object pronoun (always in subordinate clause)
0000000 0000	Sentence-final period (always in main clause)

Table 7. *The desired (regular) and actually produced (italics) values for each of the output units at the end of training for a single sentence in the training corpus with the structure [[[Possessive Subject] [Location Postposition] Verb] [Possessive Subject] Object Verb]. Shaded cells indicate that the output unit had an activation value larger than 0.2.*

Word	Subj	Obj	Poss	Loc	Main V	Sub V	Rel Pron	N	V	Adp	Pron	Lev
lægGutil	0.1 *0.06*	0.1 *0.07*	0.9 *0.92*	0.1 *0.10*	0.1 *0.08*	0.1 *0.09*	0.1 *0.11*	0.9 *0.90*	0.1 *0.09*	0.1 *0.07*	0.1 *0.09*	0.1 *0.71*
GjæXæuX	0.9 *0.95*	0.1 *0.11*	0.1 *0.06*	0.1 *0.08*	0.1 *0.06*	0.1 *0.09*	0.1 *0.08*	0.9 *0.91*	0.1 *0.08*	0.1 *0.08*	0.1 *0.10*	0.1 *0.77*
tjElew	0.1 *0.11*	0.1 *0.10*	0.1 *0.09*	0.9 *0.93*	0.1 *0.05*	0.1 *0.04*	0.1 *0.06*	0.9 *0.93*	0.1 *0.09*	0.1 *0.08*	0.1 *0.06*	0.1 *0.59*
poX	0.1 *0.08*	0.1 *0.12*	0.1 *0.05*	0.9 *0.94*	0.1 *0.04*	0.1 *0.07*	0.1 *0.08*	0.1 *0.11*	0.1 *0.09*	0.9 *0.91*	0.1 *0.09*	0.1 *0.48*
sjEts	0.1 *0.11*	0.1 *0.06*	0.1 *0.09*	0.1 *0.08*	0.1 *0.18*	0.9 *0.82*	0.1 *0.08*	0.1 *0.10*	0.9 *0.89*	0.1 *0.10*	0.1 *0.06*	0.1 *0.17*
slOfsil	0.1 *0.07*	0.1 *0.11*	0.9 *0.92*	0.1 *0.09*	0.1 *0.06*	0.1 *0.08*	0.1 *0.08*	0.9 *0.94*	0.1 *0.06*	0.1 *0.06*	0.1 *0.12*	0.9 *0.93*
spisiskEguX	0.9 *0.88*	0.1 *0.12*	0.1 *0.10*	0.1 *0.11*	0.1 *0.19*	0.1 *0.06*	0.1 *0.12*	0.9 *0.88*	0.1 *0.13*	0.1 *0.07*	0.1 *0.10*	0.9 *0.90*
jigEEmt	0.1 *0.09*	0.9 *0.93*	0.1 *0.07*	0.1 *0.06*	0.1 *0.05*	0.1 *0.09*	0.1 *0.13*	0.9 *0.93*	0.1 *0.09*	0.1 *0.08*	0.1 *0.08*	0.9 *0.81*
bælo	0.1 *0.08*	0.1 *0.08*	0.1 *0.09*	0.1 *0.11*	0.9 *0.85*	0.1 *0.15*	0.1 *0.11*	0.1 *0.08*	0.9 *0.92*	0.1 *0.08*	0.1 *0.11*	0.9 *0.92*

Table 8. *Results for simple grammar, accusative system, Subject-before-Object languages.*

		VSO		**SVO**		**SOV**		**XXX**	
		Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
MSE	**XX**	0.0002	0.0004	0.0001	0.0004	**0.0031**	**0.3982**	**0.0959**	**0.8208**
	XD	0.0002	0.0007	0.0002	0.0007	0.0003	0.0093	0.0007	**0.0422**
	HX	0.0002	0.0006	0.0001	0.0004	0.0004	**0.0210**	**0.0816**	**0.1663**
	HD	0.0002	0.0005	0.0002	0.0005	0.0002	0.0093	0.0004	0.0028

Table 9. *Results for simple grammar, accusative system, Object-before-Subject languages.*

		VOS		OVS		OSV	
		Train	**Test**	**Train**	**Test**	**Train**	**Test**
MSE	**XX**	**0.0607**	**0.1399**	**0.0031**	**0.6706**	**0.0799**	**0.4637**
	XD	0.0002	0.0025	0.0003	**0.0198**	0.0003	0.0013
	HX	**0.0417**	**0.0715**	0.0004	**0.0169**	**0.0703**	**0.1291**
	HD	0.0002	0.0007	0.0002	0.0012	0.0002	0.0009

Table 10. *In our SVO and VSO languages, the transitive and intransitive sentences always start with the same two elements in the same linear order. Only the first element is shared in SOV and VOS languages, while there is no overlap at all in OSV and OVS.*

	SVO			VSO			SOV			VOS			OSV			OVS		
Intransitive	S	V		V	S		S	V		V	S		S	V		V	S	
Transitive	S	V	O	V	S	O	S	O	V	V	O	S	O	S	V	O	V	S
	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
Identical?	✓	✓		✓	✓		✓	✗		✓	✗		✗	✗		✗	✗	

Table 11. *Results for simple grammar, ergative system, Subject-before-Object languages.*

		VSO		**SVO**		**SOV**		**XXX**	
		Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
MSE	**XX**	0.0002	0.0004	0.0001	0.0004	**0.0032**	**0.3764**	**0.0961**	**0.8045**
	XD	0.0002	0.0007	0.0002	0.0005	0.0003	0.0043	**0.0501**	**0.1399**
	HX	0.0002	0.0009	0.0002	0.0004	0.0004	0.0113	**0.0852**	**0.1638**
	HD	0.0002	0.0005	0.0001	0.0005	0.0002	0.0014	**0.0350**	**0.0625**

Table 12. *Results for simple grammar, ergative system, Object-before-Subject languages.*

		VOS		**OVS**		**OSV**	
		Train	**Test**	**Train**	**Test**	**Train**	**Test**
MSE	**XX**	**0.0671**	**0.1267**	0.0037	**0.6695**	**0.0784**	**0.4997**
	XD	**0.0656**	**0.1363**	0.0004	**0.0158**	**0.0739**	**0.1454**
	HX	**0.0477**	**0.0819**	0.0002	**0.0141**	**0.0699**	**0.1342**
	HD	**0.0121**	**0.0304**	0.0002	0.0025	**0.0755**	**0.1149**

Table 13. *Results for grammar with adpositions, Subject-before-Object languages.*

		VSO		**SVO**		**SOV**		**XXX**	
		Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
MSE	**XX**	**0.0527**	**0.0773**	**0.1177**	**0.5228**	**0.1852**	**0.5365**	**0.2350**	**0.7202**
	XD	0.0003	0.0011	0.0009	0.0192	0.0010	0.0190	0.0011	0.0432
	HX	**0.0510**	**0.0825**	**0.1137**	**0.1341**	**0.1744**	**0.2153**	**0.2242**	**0.2820**
	HD	0.0003	0.0010	0.0004	0.0012	0.0004	0.0018	0.0005	0.0033

Table 14. *Results for grammar with adpositions, Subject-before-Object languages, revised case marking.*

		VSO		SVO		SOV		XXX	
		Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
MSE	**XX**	0.0006	0.0029	0.0030	0.0235	0.0310	**0.3021**	**0.1333**	**0.6111**
	XD	0.0007	0.0018	0.0037	0.0431	0.0015	0.0167	0.0037	**0.4281**
	HX	0.0006	0.0039	0.0009	0.0130	0.0233	0.0587	**0.1253**	**0.1887**
	HD	0.0006	0.0015	0.0010	0.0111	0.0009	0.0051	0.0013	0.0169

Table 15. *Results for grammar with relative clauses, Subject-before-Object languages.*

		VSO		SVO		SOV	
		Train	**Test**	**Train**	**Test**	**Train**	**Test**
MSE	**XX**	0.0215	0.0293	0.0103	**0.3826**	**0.1766**	**0.6607**
	XD	0.0065	0.0090	0.0022	0.0289	0.0726	**0.1503**
	HX	0.0203	0.0306	0.0025	0.0203	**0.1275**	**0.1730**
	HD	0.0054	0.0067	0.0014	0.0033	0.0397	0.0446

Table 16. *Percentages of reversible transitive sentences processed correctly.*

Age (months)	English	Italian	Serbo-Croatian	Turkish
24-28	58%	66%	61%	79%
32-36	75%	78%	58%	80%
40-44	88%	85%	69%	82%
48-52	92%	90%	79%	87%