Does English Need its Pronouns?
Simulating the Effect of Pro-drop on SVO Languages

Ezra Van Everbroeck
UC San Diego
Linguistics 0108
9500 Gilman Drive
La Jolla, CA 92093
ezra@ucsd.edu

Maria Polinsky
UC San Diego
Linguistics 0108
9500 Gilman Drive
La Jolla, CA 92093
mpolinsky@ucsd.edu

Garrison W. Cottrell
UC San Diego
Computer Science & Engineering 0114
9500 Gilman Drive
La Jolla, CA 92093
gary@ucsd.edu

Abstract
We present the results from a large set of connectionist simulations exploring the effect of subject omission (pro-drop) on languages with an SVO word order. We show that pro-drop only affects the learnability of the languages if there are no cues available to tell nouns and verbs apart. We also argue that neither creoles nor Mandarin Chinese instantiate the language type which was unlearnable in the simulations.

Introduction
It is well documented that many of the world’s languages allow sentential subjects to remain unexpressed (i.e. pro-drop; Gilligan 1987): e.g. in Spanish, A comido la sopa ‘(S/He) has eaten the soup’ is a perfectly fine sentence even though it lacks an overt subject. Linguists have been studying this phenomenon for several decades, but it is still unclear to what extent pro-drop is correlated with other linguistic properties. While it was originally assumed that languages could only exhibit pro-drop if they also featured rich subject-verb agreement — like Italian and Spanish, where the agreement affixes on the verb essentially replace the information which is lost by omitting the subject (Jaeggli and Safir 1989) — it was soon pointed out that some languages which lack agreement systems altogether, like Mandarin Chinese and Thai, also exhibit frequent pro-drop (Huang 1984, 1989).

To gain insight into the interactions between pro-drop and other linguistic parameters, we have run 12,000 neural network simulations to systematically investigate how languages with different basic word orders and varying degrees of overt morphological marking are affected by the introduction of pro-drop. By using computational models and artificial languages, we can control each parameter separately and also look at combinations of properties (i.e. types of languages) which are unattested in the real world — neither of which would be possible if we limited ourselves to more traditional, grammar-based typological research. In order to keep the results and discussion sections manageable, we will restrict ourselves here to the network models which had to learn languages with a basic Subject-Verb-Object (SVO) word order. Cross-linguistically, the SOV word order is more frequent (Tomlin 1986; Dryer 1989), but SVO languages are more interesting for our purposes because they display more morphological variation than SOV languages (Siewierska 1996, 1998) and because a number of well-studied languages like English, Spanish and Mandarin Chinese are SVO.

Experiment
The use of neural networks for typological studies is a recent development in cognitive science, but there have already been several successful models (Christiansen and Devlin 1997; Lupyan and Christiansen 2002; Van Everbroeck 1999, 2003). For our simulations, we created simple context-free grammars capable of generating many different artificial languages, each of which represented a specific language type. The lexicon used to generate the training corpora consisted of 300 nouns (half animate, half inanimate), 100 verbs (half transitive, half intransitive) as well as several pronouns and morphological markers. The test lexicon contained new nouns and verbs but the same pronouns and markers. For each language type, the grammars were used to generate training and test corpora of 3,000 simple sentences. The possible sentences were SV and SVO, as well as V and VO for the language types in which subjects could remain unexpressed. In addition to word order and the presence/absence of pro-drop, the two other crucial parameters were the presence/absence of Subject/Object case-marking on the nouns, and the presence/absence of head-marking on the verbs. The latter could take the form of simple Tense/Aspect/Modality markers which essentially only help identify the verbs (Bybee 1985), valency markers indicating the number of arguments in the clause (e.g. McWhorter 1998), or rich subject-verb agreement.

The architecture of the model used in our simulations consisted of a simple recurrent network (Elman 1990) augmented with a recurrent layer for the output units (see Figure 1). The latter functioned as a short-term memory and led to much faster training. At the input layer, the networks
were shown one word of a sentence at a time. At the end of each sentence, a special ‘period’ pattern was presented to signal to the network that a new sentence was about to start. At the output layer, the networks were trained to build a representation of the entire sentence: there were separate banks of units for the subject, object and verb of the sentence and each bank had to be filled with the appropriate word from the sentence, or left empty if the sentence didn’t contain the element — e.g. the object bank was to remain empty in all intransitive sentences. The networks were trained using back-propagation for 10 epochs, and then tested on the sentences from the appropriate test corpus. The main error measure was the percentage of sentences for which the network got all three output banks correct; i.e. the pattern of activation over each bank of output units had to be closer to the correct word than to any other one.

Results

An ANOVA test reveals the importance of each linguistic parameter used for generating the artificial languages: pro-drop, case-marking on the nouns, and head-markers on the verbs are all statistically significant (see Table 1).

<table>
<thead>
<tr>
<th>Effect</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>3177.1</td>
<td>1</td>
<td>3177.1</td>
<td>34.31</td>
<td>.000*</td>
</tr>
<tr>
<td>Head</td>
<td>1.04E2</td>
<td>3</td>
<td>3.460</td>
<td>37.37</td>
<td>.000*</td>
</tr>
<tr>
<td>Pro</td>
<td>5403.1</td>
<td>1</td>
<td>5403.1</td>
<td>58.35</td>
<td>.000*</td>
</tr>
</tbody>
</table>

Table 1. ANOVA analysis results.

If we look at individual language types and how well the trained networks were able to generalize to the new sentences in the test corpora, we find that test performance was excellent (> 90%) for all but one combination of linguistic features (see Table 2; the percentages in each cell are averages over 20 networks with different initial weights). The problematic language type combines an SVO word order with pro-drop, no case-marking on the nouns and no marking on the verbs.

An analysis of the errors made by the various networks shows that telling the nouns apart from the verbs is the main problem caused by pro-drop. When all subjects are expressed, SVO languages are easy to parse because the first word is always the subject, the second word always the verb, and the third (if present) the object. (Note that these networks still make some mistakes because they are being forced to analyze sentences with nouns and verbs which they have never been trained on.) When there is pro-drop, however, this simple parsing strategy no longer works because the first word can be either the subject (in SV, SVO) or the verb (in V, VO); similarly, the second word can be either the verb or the object. Because the nouns and verbs are novel in the test corpora, some additional information is needed to identify which is which. The results in Table 2 demonstrate that marking the nouns (through case-marking) is as successful for the disambiguation task as marking the verbs (through any of its three types of marking). Interestingly, marking both simultaneously has very limited benefit. But if no marking is available to tell the nouns from the verbs, the networks fail to generalize to the novel sentences in the test corpus.

Discussion

A comparison of the network results in Table 2 with typological data on SVO languages suggests that some possible language types may be unattested (see Table 3). These ‘gaps’ in the space of possible languages could well be due to random historical events (Diamond 1997). On the other hand, the potential absence of natural languages which correspond to the type the networks found problematic suggests that some gaps could be motivated by cognitive factors such as learnability.

There are two groups of natural languages which may correspond to the supposedly unattested type in Table 3. The first group consists of creole languages, i.e. languages which developed over the last 400 years or so in contact situations where speakers from mutually unintelligible languages were forced to communicate (Bickerton 1981; Thomason and Kaufman 1988; McWhorter 1998).

<table>
<thead>
<tr>
<th>SVO</th>
<th>Head-marking</th>
</tr>
</thead>
<tbody>
<tr>
<td>No pro-drop</td>
<td>None</td>
</tr>
<tr>
<td>[-case]</td>
<td>96.2%</td>
</tr>
<tr>
<td>[+case]</td>
<td>97.0%</td>
</tr>
</tbody>
</table>

Table 2. Percentages of novel sentences of each language type which are parsed correctly.
All creoles share several linguistic properties, including an SVO word order and very little morphological marking on nouns and verbs. Many also derive partially from Spanish and Portuguese, both of which have frequent pro-drop. Nonetheless, if we look at the pro-drop phenomena observed in most creole languages, we find that they hardly ever go beyond the kinds of constructions which are also possible in English, a prominent non-pro-drop language. For example, it is possible to say the creole equivalent of ‘Seems like a good idea’ in Spanish-based Capeverdean Creole (Baptista 1995), Portuguese-based Papiamento (Kouwenberg 1990) and French-based Haitian Creole (DeGraff 1993). But the equivalent of Spanish *Está comiendo la sopa* ‘(S/He) is eating the soup’ is not acceptable, even in the creoles which derive from Spanish (Muysken and Law 1991). Moreover, in the few cases where creoles do allow pro-drop, they have also developed subject-verb agreement — e.g. Bislama (Meyerhoff 2000) and São Tomé Creole (Gilligan 1987). So, there is no evidence that creoles represent the language type which the networks have problems learning.

The second group of potential counter-examples can be found among the languages of South-East Asia, including Thai, Lao, Vietnamese, and the varieties of Chinese. Although not all of them are historically related, they share a large number of linguistic features, including an SVO word order and minimal noun and verb morphology, due to their pro-longed geographical proximity (Cooke 1968; Bisang 1996). Crucially, some of these languages also exhibit very frequent pro-drop. In Thai, for example, unexpressed subjects occurred in about every second sentence in a large corpus (Aroonmanakun 1999). Similar numbers are often mentioned for Mandarin Chinese (Huang 1984, 1989; Tao 1996; Tardif, Shatz, and Naigles 1997). We limit ourselves to a discussion of Mandarin here because it is the best documented of these languages. A closer analysis of Mandarin shows that there are a number of discourse and structural constraints on the usage of pro-drop. In general, pro-drop is only allowed when the unexpressed element can be readily recovered from the discourse or the situational context (Huang 1995). The structural constraints involve co-verbs and pivotal nouns in serial verb constructions (Li and Thompson 1981); in both cases, they prevent pro-drop from causing a verb to appear where a noun would normally be expected. In addition, Mandarin actually has some clues which help identify verbs (e.g. aspect markers, auxiliaries and co-verbs — Li, Bates, and MacWhinney 1993) and nouns (e.g. numerals, classifiers, prepositions, and the *ba* and *bei* particles — Chang 1992). Finally, there is considerable evidence from acquisition studies that children learning Mandarin Chinese prefer a fixed SVO word order, both in perception and production, and also don’t make mistakes in using nouns as verbs or vice versa (Erbaugh 1983, 1986; Miao and Zhu 1992). The combination of all these factors suggests to us that Mandarin, although it may come closer than most SVO languages, is not an instantiation either of the type which the models could not learn.

<table>
<thead>
<tr>
<th>SVO</th>
<th>Head-marking</th>
</tr>
</thead>
<tbody>
<tr>
<td>No pro-drop</td>
<td>None</td>
</tr>
<tr>
<td>[-case]</td>
<td>English</td>
</tr>
<tr>
<td>[+case]</td>
<td>Russian</td>
</tr>
<tr>
<td>pro-drop</td>
<td>(45.8%) Hebrew</td>
</tr>
<tr>
<td>[-case]</td>
<td>Polish</td>
</tr>
</tbody>
</table>

Table 3. Natural languages which correspond to the ones learned by the networks.

**Conclusion**

Our connectionist simulations of the effect of pro-drop on SVO languages show that the presence of unexpressed arguments need not create serious problems for a language learner, at least if some morphological marking is available to distinguish the nouns in the language from the verbs. If no such source of information is available, the networks fail to generalize to sentences containing novel words. This finding meshes well with what is known about human language processing because people, and especially young children, do poorly when faced with structural ambiguities (Trueswell et al. 1999; Hawkins 2002).

From a typological perspective, our results are relevant because they help determine which SVO language types are unattested for cognitive reasons as opposed to the result of historical accident. The fact that no real languages appears to instantiate the type which the networks found unlearnable, even when they could have been expected to do so as in the case of Spanish-based creoles, is a strong indication for us that computer simulations can make a valuable contribution to linguistic typology. On the other hand, we also want to stress that language simulations like ours currently have limited explanatory value because they do not include semantic or pragmatic information. Adding lexical semantics of the type described by Li, Burgess, and Lund (2000) is definitely possible, however, so this deficiency can presumably be addressed.

Other future work includes the analysis of the verb-final (mostly SOV) and verb-initial (VSO, VOS) language types, as well as closer look at Riau Indonesian and Singapore English, two SVO languages about which controversial claims have been made with respect to their linguistic features (Gil 1994, 2003).
References


