UNIVERSITY OF CALIFORNIA, SAN DIEGO

A Connectionist Model of the Effect of Pro-drop on SVO Languages

A Dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy

in

Linguistics and Cognitive Science

by

Ezra Laurens Van Everbroeck

Committee in charge:

Professor Maria Polinsky, Chair
Professor Garrison Cottrell, Co-Chair
Professor Jeffrey Elman
Professor Andrew Kehler
Professor Ronald Langacker

2007

The Dissertation of Ezra Laurens Van Everbroeck is approved, and it is acceptable in quality and form for publication on microfilm:

Co-Chair

Chair

University of California, San Diego

2007

This dissertation is dedicated

to Gary and Masha

for simply refusing

to let me quit.

A man said to the universe:

"Sir I exist!"

"However," replied the universe,

"The fact has not created in me

A sense of obligation."

Stephen Crane

TABLE OF CONTENTS

List of Figures

LIST OF TABLES

ACKNOWLEDGMENTS

There is something to be said for taking eleven years to complete a Ph.D. program. It may not have been particularly efficient, but doing it the leisurely way has given me much more time to enjoy the company, teachings and advice of a large number of people in San Diego. I have no doubt that the memories of them will stay with me far longer than the intricacies of many theories.

First and foremost I want to thank my co-chairs, Gary Cottrell and Masha Polinsky, for having supported me as a student and as a person for more than a decade. Masha exposed me to a world of typological data in my first quarter at UCSD and the need to account for the diversity of the world's languages became a cornerstone for this dissertation. Gary let me join his unbelievable research unit in the Summer of 1998 – I am positive no linguist knows as much about the leech's central nervous system as I do – and proceeded to teach me more about machine learning than I was ever able to retain. In more recent years, they exhibited endless patience as they encouraged, coaxed, and occasionally threatened me into finishing this dissertation. I will never understand how they put up with my lack of progress for so long, but I do know that I am extremely lucky to have had them as advisers.

Next, I want to express my gratitude to the other members of the committee, and not only for being willing to read this dissertation critically. Ron Langacker's work on cognitive linguistics is the main reason I came to UCSD and to this day I believe it is the only linguistic framework that is fundamentally sound. Along similar lines, Jeff Elman's articles from the early 90s describing how simple recurrent networks can learn

linguistic structures are why I took up connectionist modeling. Finally, Andy Kehler's academic influence can be found in the section on probabilistic linguistics in the general discussion. I also owe him for many years of excellent advice and guidance in his capacity as chair of the department's Computing Advisory Committee.

Over the years, I have also benefited greatly from courses taught by many other faculty members at UCSD, both in Linguistics and in other departments. One person who stands out in this regard is Liz Bates. Her classes on language acquisition and aphasia were intensely demanding but absolutely worth it. In addition, the Competition Model of sentence comprehension she developed with Brian MacWhinney has been a major source of inspiration for my own work.

While wearing my other hat as a staff employee, I have had the good fortune of working with a group of wonderful colleagues in the Linguistics department. A special thanks is reserved for Rock Hunter as a gracious mentor in matters computational as well as personal. I am also greatly indebted to Julie Williams for her support and for being the kind of supervisor most employees can only dream of. Linda Murphy told me so many entertaining stories about the history of the department that I could have written a book or two; Vicki King has been such a constant source of useful information that I can't imagine working without her; Dennis Fink and Marc Silver deserve great credit for making the IT group work so well that I have been able to combine professional and academic careers.

There are also numerous people who have made the last decade enjoyable when studies and work threatened to overwhelm my sanity. Bill Morris was the big brother I had never had, from helping me with C programming to teaching me pool tricks. Wind Cowles and Gina Taranto suffered through the first few years of grad school boot camp with me and later demonstrated that there was hope for our cohort. When it was necessary to escape into fantasy altogether, the members of the Mesa D&D group were always happy to oblige and sit down for an evening of beer, pizza, and lots of fun.

Last but far from least, there are those who have filled my personal life. I would like to thank Pat Maxwell for hosting me when I first arrived in San Diego and checking up on me many times afterwards. Kim Hansen was the roommate who defied description but nevertheless meant well. And my parents have obviously helped me out as much as anyone could have, especially given how far I moved away. I also need to credit my father for proof-reading the dissertation and catching a large number of embarrassing mistakes.

At home, Lorenzo and Diego have simply been two marvelous kids, a constant source of parental pride (and some hair-raising frustration). I am looking forward to finally being able to spend more time with them. As for Alicia – my beloved spouse – no words can do justice to the sacrifices she has made for me. She has been the cornerstone of my life for the last ten years and this dissertation would never have been finished without her love.

VITA

1994 Master of Arts, Katholieke Universiteit Leuven

1996 Master of Science, Katholieke Universiteit Leuven and Edinburgh University

1997 Senior Coder, University of California, San Diego

1998 Programmer/Analyst I, University of California, San Diego

2000 Programmer/Analyst III, University of California, San Diego

2007 Doctor of Philosophy, University of California, San Diego

JOURNAL PUBLICATIONS

2003 Ezra Van Everbroeck, "Language type frequency and learnability from a connectionist perspective", *Linguistic Typology* 7.1: 1-50

2003 Maria Polinsky & Ezra Van Everbroeck, "Development of gender classifications: Modeling the historical change from Latin to French", *Language* 79: 356-390

OTHER PUBLICATIONS

1994 Ezra Van Everbroeck, *Is it all in the Mind? A Critical Survey of Cognitive Science into the 90s*, Unpublished Master of Arts thesis

1996 Ezra Van Everbroeck, *CLASPnet. A Modular Recurrent Neural Network for Spotting Clausal Properties*, Unpublished Master of Science thesis

1999 Ezra Van Everbroeck, "Could Sarah read the Wall Street Journal?", *UCSD Center for Research in Language Newsletter* 11.7

1999 Ezra Van Everbroeck, "Language type frequency and learnability: A connectionist approach", in *Proceedings of the 21st Annual Conference of the Cognitive Science Society*, Mahwah, NJ: Lawrence, 755-760

2004 Ezra Van Everbroeck, Maria Polinsky & Garrison Cottrell, "Does English need its pronouns? Simulating the effect of pro-drop on SVO languages", in *Language Learning: An Interdisciplinary Perspective. Papers from the 2004 AAAI Spring Symposium* (Eds. Paul Cohen, Andy Clark, Eduard Hovy, Tim Oates & Michael Witbrock), Menlo Park, CA: AAAI Press, 69-73

CONFERENCE PRESENTATIONS

1998 "New directions in grammar presentation and testing: Internet practice and intranet exams", CALICO 1998, San Diego State University (with Dr. Schane)

1999 "Language type frequency and learnability: A connectionist approach", Annual Cognitive Science Conference 21, Simon Fraser University

2004 "Does English need its pronouns? Simulating the effect of pro-drop on SVO languages", AAAI 2004 Spring Symposium, Stanford

AWARDS

1994 European Community, 1-year ERASMUS scholarship

1996 Belgian American Educational Foundation, 1-year fellowship

2003 University of California, San Diego, Division of Social Sciences
Certificate of recognition for outstanding service

ABSTRACT OF THE DISSERTATION

A Connectionist Model of the Effect of Pro-drop on SVO Languages

by

Ezra Laurens Van Everbroeck

Doctor of Philosophy in Linguistics and Cognitive Science

University of California, San Diego, 2007

Professor Maria Polinsky, Chair

Professor Garrison Cottrell, Co-Chair

We present four computational experiments that investigate the impact of null subjects (pro-drop) on the learnability of languages with a basic Subject-Verb-Object (SVO) word order and varying amounts of morphological marking on their nouns and verbs. The simulations show that the effect of pro-drop on language learnability is limited as long as some morphological marking is present. Contrary to expectation, rich agreement markers are no more useful in the simulations than nominal case markers or verbal Tense/Aspect/Modality markers. In the absence of morphological

marking, however, pro-drop leads to severe learnability problems in the simulations: overall performance on this language type is significantly worse (Experiment 1); additional exposure to language data is not as useful as with other types (Experiment 2); novel words are more problematic in this type (Experiment 3); and noun/verb homonyms also decrease performance for this type (Experiment 4). An analysis of the simulations shows that the main problem is accurately distinguishing nouns from verbs. These results suggests that the combination of pro-drop and no morphological marking should be unattested among natural languages.

To test this hypothesis we first survey various creole languages as they are SVO and typically lack morphological markers. However, cross-linguistic data shows that creole languages do not allow pro-drop unless they have also developed agreement markers. We then discuss Mandarin Chinese because it allows widespread pro-drop and features only minimal morphological marking. A closer look at the language reveals that Mandarin provides quite reliable cues for identifying nouns and verbs in the language. Crucially, these cues are acquired very early by children learning Mandarin. Similarly, children only very rarely use nouns as verbs (or vice versa) – unlike in English where pro-drop is not possible. Two other unusual properties of Mandarin Chinese that are also compatible with our experimental results are the relatively early acquisition of verbs and the presence of relatively frequent noun/verb homonymy. Mandarin is thus not a counter-example to the results of the simulations.

We end by situating our work in relation to various other approaches, such as the Competition Model, Optimality Theory, and probabilistic linguistics.

Chapter 1. Introduction

In pro-drop languages, such as Spanish and Mandarin Chinese, it is possible to omit just about any subject in a sentence (see section 2.3 for details). It is a phenomenon that has received a lot of attention in theoretical linguistics, largely because the presence of pro-drop in a language was believed to predict various other linguistic properties (Perlmutter 1971; Chomsky 1981; Gilligan 1987; Nicolis 2005; Falk 2006). The overall predictive power of the 'pro-drop parameter' has slowly diminished as data from more languages has become available (Newmeyer 2005), but there is still a commonly held belief that pro-drop can only occur in a language under either of two mutually exclusive conditions (cf. Huang 1984, 1989; Jaeggli and Safir 1989). Either the language must have rich verb agreement (as in Spanish, Italian or Swahili), so the extra morphology on the verb can replace the information from the unexpressed subject; or the language must have no significant morphological processes at all (as in Mandarin Chinese or Thai). In this view, a language that is in between these two extremes – i.e. it only has limited verbal morphology or perhaps rich nominal morphology – is predicted not to allow pro-drop at all.

We present four experiments that examine the validity of this account of pro-drop through the use of computational models. In our experiments, the models have to learn various types of artificial languages that only differ with respect to three relevant linguistic parameters. I.e., first, the presence/absence of morphological

marking on nouns (via case-marking); second, the presence/absence of morphological marking on verbs (via rich agreement and Tense/Aspect/Modality markers); and, third, the presence/absence of pro-drop. In their simplest form, these three parameters define a space of (2x3x2) 12 possible language types that include not only Spanish and Mandarin Chinese, but also a number of types that should not allow pro-drop according to the traditional linguistic account. We will see below how our models fared on these types. Contrary to expectation, we found that the presence of morphological markers (whether on nouns or on verbs) does not always benefit learning, and also that the presence of pro-drop (and the implied lack of information) only negatively affects learning in a single language type.

The task that the models have to learn is determining 'who did what to whom' for each input sentence of their language. We chose this task both because it is fundamental to the communicative function of language (Slobin and Bever 1982; Bates, MacWhinney et al. 1984; MacWhinney and Bates 1989). It is hard to imagine effective communication if it is not possible to correctly distinguish between the agent, the object of the action, and the action itself. In addition, this task can be tested using simple sentences that contain only a few words. Our implementation of the 'who did what to whom' task is relatively abstract, though, because the models do not have access to any conceptual, semantic or pragmatic information. The labels agent, patient and action thus only apply in a metaphorical sense. What the models really have to learn is to put each word in a sentence in one of three output bins that are labeled 'Subject', 'Verb' and 'Object'. All the information they have access to is formal in

nature – i.e. which word appears, which markers it carries, and what its position is in the input sentence.

The absence of semantic information in the models is obviously not intended to accurately reflect how children learn languages. There is a convincing body of research available demonstrating that even infants as young as a few months have rich cognitive capabilities (Baillargeon, Spelke and Wasserman 1985; Baillargeon 2004; Mandler 2004) and are interested in communicating with others (Givón 1979; Tomasello 2000). Still, we also know that children (as well as adults) can learn to distinguish 'grammatical' patterns in artificial and meaningless languages from ungrammatical ones (Saffran, Aslin and Newport 1996; Gomez and Gerken 1999, 2000; Hudson Kam and Newport 2005; Kaschak and Saffran 2006). We believe that pre-linguistic children are learning basic facts about their languages without the benefit of the incredibly rich semantic context which children beyond five years of age and adults take for granted. In the absence of strong evidence as to what words really mean for very young children, artificially limiting the models to purely formal information seems both a reasonable assumption and a good safeguard to ensure that the effects which we observe in our models can not (entirely) be attributed to conceptual factors that are not defined explicitly. Essentially, studying language learning in the limit is a reasonable method to explore the limits of language.

These experiments are part of a larger project examining the notion of the 'human language space' – i.e. what kinds of languages can be learned as a first language by children. Cross-linguistic studies estimate the number of human languages at about

6,000 – give or take a thousand depending on one's definition of what constitutes a language as opposed to a dialect (Comrie 1989). On the other hand, even if languages could only differ along twenty binary parameters (and we know both that there are more than twenty parameters and that most of them are not binary), there would be more than one million possible languages. The number of potential combinations of linguistic features must be orders of magnitude higher.

Obviously, we are not the first to investigate the space of possible languages via combinations of linguistic features. There is a rich literature of cross-linguistic work that describes implicational language universals (e.g. Greenberg 1963; Hawkins 1983, 1988; Dryer 1988, 1992). These universals define correlations between specific properties – e.g. languages with a basic word order of Subject, Object, Verb almost always have case-markers on the nouns. In the generative linguistic tradition, the Principles and Parameters approach inspired by Chomsky (1981) posits that there is a finite number of (syntactic) parameters that can be used to correctly describe all the world's languages. In this framework, children are believed to set these parameters on the basis of the linguistic input they receive (Gibson and Wexler 1994; Berwick and Niyogi 1996; Briscoe 2000; Yang 2006), and, once set, these parameters may affect other properties of a child's grammar (Newmeyer 2005). Along similar lines, most work in Optimality Theory is concerned with the (phonological) properties of human languages and how various constraints can be ranked with respect to each other to explain the kinds of languages that we find nowadays (Prince and Smolensky 1993; Samek-Lodovici 2001; Kuhn 2003; Tesar 2004).

However, there is a shared goal in the three approaches just mentioned which we believe to be unnecessarily narrow – i.e. they only want to account for the currently attested inventory of human languages. They thus share the assumption that all unattested language types should be ruled out somehow by the relevant theory. While this is a worthwhile goal from a descriptive point of view, it fails to do justice to how large the space of possible languages is. What we really want to know is why the unattested language types are unattested. Ideally, we should be able to distinguish possible unattested languages – i.e. languages that appear learnable by human beings – from unattested languages that appear impossible to learn for our cognitive systems. Simply put, languages cannot evolve in a direction which makes them unlearnable by the (relatively) limited cognitive abilities of very young children (Kroch 1989; Saffran 2002). Whereas adults may be able to produce and parse a 'hard' language due to their larger cognitive and conceptual capabilities, language learning children will re-analyze the input they hear in such a way as to have it make more sense. Inconsistent patterns are likely to be re-analyzed first, especially if they are not supported by high frequencies or an early age of acquisition (e.g. Marchman, Plunkett and Goodman 1995; Hudson Kam and Newport 2005).

Naturally, the fact that the space of possible languages is only sparsely populated may have a very simple explanation in the form of historical accident. It's conceivable that all current languages derive from a single proto-language and there hasn't been enough time for languages to change in such a way that they would occupy more of the space. It is also possible that radically different languages existed at one

point, but they disappeared as the result of non-linguistic factors such as natural catastrophes or genocide (Diamond 1999). While such events have likely played a role in the current inventory of human languages, they cannot be tested as easily in experiments as the learnability factor.

Testing the learnability of different language types is most easily done with computational models. At a practical level, they don't present the kinds of moral challenges we would face in trying to teach 'impossible' languages to human infants, and they provide an endless supply of reliable subjects for the experiments. In addition, they also allow us to systematically compare all the possible languages defined by linguistic parameters we are interested in. The main drawback of this approach is that there exist no models of human languages which do full justice to their complexities; even for English, there is no comprehensive grammar that accounts for all of its structures. Even if such grammars were available, though, they would have been of limited use due to the need to also test the learnability of languages for which there are no known counterparts. The only way one can truly compare all the languages that we are interested in is by systematically varying each parameter in turn. This provides us with minimal pairs of languages and relatively easy comparative analyses between them.

There are a few recent studies that have also used computational techniques to explore possible languages. Kirby (1997) investigated the relative clause accessibility hierarchy – e.g. any language that allows the head of the relative clause to function as a direct object will also allow it to function as a subject (Keenan and Comrie 1977). He

found that he could get this hierarchy to emerge in populations of modeled speakers if certain costs and benefits were associated with producing and understanding relative clauses. Christiansen and Devlin (1997) built neural networks to examine the effect of head position consistency on language learnability. They found that their models had fewer problems learning languages that were either fully head-final or fully head-initial than languages that were inconsistent in this regard. Interestingly, human languages show a similar preference for having the heads of each phrase appear in leftmost or rightmost position (Hawkins 1993, 1994). Monaghan, Gonitzke and Chater (2003) investigated the learnability problems associated with within-language word order inconsistencies (e.g. German is SVO in main clauses, but SOV in subordinate clauses). Using data from large corpora to generate the grammars for their neural networks, they observed that the negative impact of such inconsistencies could be quite small due to interactions between different sentence constructions.

This study is different from the ones just mentioned in several respects. Most importantly, we investigate a different part of the language space by using other parameters in the simulations – i.e. no one has used computational models to determine the effects of pro-drop or Tense/Aspect/Modality markers on language learnability. In addition, we are also more interested in the interactions between several of the parameters, rather than the main effect of a single parameter. Finally, we use considerably more data from actual languages to assess the validity of our computational results. E.g. we provide examples from many different languages (including several creoles) to give a detailed description of the pro-drop phenomena

we account for; and in several chapters we return to the linguistic properties of Mandarin Chinese to discuss its implications for the results of our models.

The work we present here is also quite different from the simulations in Van Everbroeck (1999, 2003). These studies looked at different linguistic parameters (including word order and the effect of accusative versus ergative alignment) and varied the complexity of the artificial languages by optionally including possessives, locatives, and relative clauses in the sentences presented to the models. Pro-drop was not included in his models. Moreover, we have added pronouns to the artificial languages to make them more realistic; natural languages invariably limit their use of full nouns, preferring to use pronouns and pro-drop instead (Sun and Givón 1985; Du Bois 1987; Chui 1992). The models of Van Everbroeck (2003) were also constructed differently, using phonological word representations at the input layer and just a current word analysis at the output layer. In the experiments described below, the input representation was much more compact and the output layer was used to gradually build up a representation of the entire sentence. Finally, the main goal of Van Everbroeck (2003) was to compare the simulation results to a implicational language universals described in the cross-linguistic literature. In this study, however, we are far more concerned with how the models compare to acquisition data from human languages. Much of the discussion will also be focused on a single language, Mandarin Chinese, as it might appear to contradict the results of our models.

The structure of the dissertation is as follows. In Chapter 2, we provide a thorough discussion of the main linguistic parameters which define the language space

explored in the experiments. Chapter 3 presents the implementation of the neural network models used in the simulations. Chapters 4 through 7 cover the four experiments. Experiment 1 explores the impact of pro-drop on language learnability. Experiment 2 investigates the benefits of additional training exposure and the acquisition of lexical versus grammatical knowledge. Experiment 3 looks at how the presence of novel nouns and/or verbs affects the performance of the models. Experiment 4 discusses the consequences of increasing levels of noun/verb homonymy on language learnability. Each experimental section also compares the relevant network results to what is known about similar natural languages. In Chapter 8, the General Discussion, we compare our findings to those of related work in four different lines of research (lexical learning; the Competition Model; probabilistic linguistics; Optimality Theory syntax). Finally, Chapter 9 concludes with brief summaries of the four experiments, an assessment of how our work reflects on the linguistic pro-drop parameter, as well as possible directions for future work.

Chapter 2. Linguistic parameters

The space of possible languages which we explore in our computational models is defined by several linguistic parameters: i.e. word order, case-marking on nouns, morphological marking on verbs, and null subjects (or pro-drop). We briefly discuss each parameter below, covering both its linguistic background as well as the parameter values that we implemented in the experiments.

2.1 Word order

The surface word order parameter is used in cross-linguistic research to describe the most common and most natural sounding order with which the subject (S), verb (V) and object (O) appear in simple declarative sentences in a given language (Steele 1978; Hawkins 1983; Siewierska 1988; Dryer 1997). It is an important parameter in typology because the basic word order of a language is often correlated with various other linguistic features, such as the existence of prepositions and postpositions, or the relative order of relative clauses and their head nouns (Greenberg 1963; Vennemann 1975; Dryer 1988, 1992; Hawkins 1988; Nichols 1992). For the purposes of the current work, though, the most relevant property of word order is that it can suffice to communicate 'who did what to whom' in some languages. For example, the difference in meaning between the English sentences (1) and (2) is

entirely due to the order in which the noun phrases *the man* and *the kangaroo* appear in each sentence.

(1) The man kicked the kangaroo.

(2) The kangaroo kicked the man.

There are six possible orders of S, V and O, but only two of them (SOV, SVO) are really common, while two others occur hardly at all (OVS, OSV). The exact frequency of each word order among the languages of the world depends on one's methodology and sample size, so the numbers in Table 1 are based on a number of sources (Tomlin 1986; Dryer 1989; Siewierska 1996).

Table 1. Estimated frequencies of basic word orders in the languages of the world. (The numbers do not add up to 100% because some languages can not be classified as having a single basic word order (see Mithun 1987).

BASIC WORD ORDER					
SOV	SVO	VSO	VOS	OVS	OSV
51%	23%	11%	8%	0.75%	0.25%

We have built computational models of all six word orders in previous work (Van Everbroeck 2003), but we limited ourselves to the SVO word order for the experiments described below.[1] There are several reasons for this choice. First, SVO

[1] We have already run computational models like the ones described below for all six basic word orders. However, we still have to analyze the results for all the non-SVO languages.

accounts for almost a quarter of the world's languages, and these include a sizable number of languages which have been studied in detail. Second, SVO languages exhibit an interesting mixture of other linguistic features. For example, although there are more than 3,000 SOV languages, only a handful of them lack case-marking on their nouns. In fact, a representative sample of 300 of the world's languages (Siewierska 1996, 1998), shows that SVO languages are less likely to feature agreement marking (see section 2.2.2) on the verb or nominal case-marking (see section 2.2.1) than either verb-initial (VSO, VOS) or verb-final (SOV, OSV) languages. Finally, there are SVO languages which either do (e.g. Mandarin Chinese, Hebrew, Spanish) or do not (e.g. English, Vietnamese, Finnish) allow pro-drop (see section 2.3). The existence of many types of attested SVO languages has made it easier for us to compare the results from our language learning simulations to what is known about natural languages.

2.2 Morphological marking

While the two processes are theoretically independent because they apply to different lexical categories, we combine the description of morphological markers on nouns and verbs in a single section because the two share numerous properties (compare Nichols 1986; Song 2001). At the surface level, the phonological and morphological shapes of case markers and verbs markers are often quite similar in any given language. Functionally, both can identify the subject and object grammatical relations to indicate 'who did what to whom'. Moreover, when both case markers and

verbal markers occur in a single language, their behavior is often correlated. Still, given that the two styles of marking can occur independently, all the various combinations of case-marking and verb marking were examined in our artificial languages.

The similarities between morphological marking of nouns and verbs come as something of a surprise when we consider the many differences between the two lexical categories. (As a matter of fact, a recurring theme in our experiments will be how important it is to be able to tell the two apart in SVO languages.) Despite their many other disagreements, all linguistics theories of syntax and/or semantics take nouns and verbs to be two basic lexical categories (Langacker 1987; Baker 2003). Following seminal work by Genter (1981,1982, 2006; Gentner and Boroditsky 2001), the distinction is taken to be the result of the different properties of their conceptual prototypes. Whereas prototypical nouns refer to physical objects with many concrete perceptual qualities, prototypical verbs refer to processes in the world which are not as easy to delineate into separable wholes. For example, determining whether a given physical action (no matter how well it can be observed) should be described by the verb *to push* is harder than to determine whether a certain creature should be called *a rabbit*.

As a result of these conceptual differences, the meanings of verbs have lower referential densities than those of nouns – i.e. they have fewer reliable connections to other concepts. Dictionary entries show the average verb has more different word senses than the average noun. Verbs are also harder to remember, show more variation in translations than nouns, and are less likely to be borrowed in language

contact situations (Gentner 1981; Thomason and Kaufman 1988). More empirical support for the distinction comes mostly from the systematic differences in how they interact with other words (e.g. nouns are modified by adjectives; verbs by adverbs), and also from a growing body of psycholinguistic research which shows that nouns and verbs are processed differently in the brain (e.g. Brown, Marsh and Smith 1979; Caramazza and Hillis 1991; Federmeier, Segal et al. 2000; Tyler, Russell et al. 2001; Li, Jin et al. 2004). The distinction is also reflected in the current literature on language acquisition (Maratsos and Chalkley 1980; Gillette, Gleitman et al. 1999; Marshall 2003; Gleitman, Cassidy et al. 2005) and various types of language loss (Black and Chiat 2003; De Bleser and Kauschke 2003; Polinsky 2005).

In short, nouns and verbs are distinct lexical categories which are supported by at least partially different mechanisms in the brain. Nouns may attract case markers because of their conceptual primacy of nouns, but this does not entail that verbs are any less likely to receive morphological marking. On the contrary, verbs constitute the core of the sentences they appear in because their less specific semantics can accommodate the nouns that are being linked. Let us compare the two kinds of marking in more detail.

2.2.1 NOMINAL CASE MARKERS

Morphological case is a system to mark a dependent nouns for the type of relationship it has to its head (Blake 1994).[2] These heads can be prepositions or other nouns, but most often are verbs which need to have their argument slots filled. The types of relationships which case can potentially express are legion, so it is not surprising that no two languages use case-marking in the same way. Some languages such as Mandarin Chinese do not have any case-marking at all. Others, such as Daghestanian languages, use them in abundance, often displaying over 50 cases if all local forms are counted (see Comrie and Polinsky (1998) for discussion).

There are a few cases which appear in most of the world's languages, because they are used to identify the core participants in events and can thus be used to determine 'who did what to whom'. It is customary to refer to these participants as A (the agent or subject of a transitive verb), P (the patient or object of a transitive verb), and S (the patient/experiencer or subject of an intransitive verb). For example, in the transitive sentence *she kissed him*, the pronoun *she* is the A and the pronoun him is the P. In *she slept* the S is *she*. Note that we have used pronouns in this example because English no longer features case-marking on its nouns. These example also show that English has nominative/accusative alignment: i.e. it uses one case (nominative) to code subjects (A, S), and a different case (accusative) to mark objects (P).
Cross-linguistically, the nominative/accusative alignment system is most common,

---

[2] Note that we do not concern ourselves here with abstract Case assignment, an issue of much theoretical concern. Our interest is in overt morphological case.

with the absolutive/ergative system a distant second (Siewierska 1996).[3] In ergative languages there is one case (absolutive) which is used to mark P and S, and another one (ergative) that only marks A. The following sentence pair illustrates the use of these markers in Yidiny, an Australian aboriginal language (Dixon 1980).

(3) yiŋu waguuja gali-ŋ
This-ABS man-ABS go-PRES
"This man is going." (Yidiny)

(4) mujaam-bu waguuja wawa-l
Mother-ERG man-ABS look at-PRES
"Mother is looking at the man." (Yidiny)

In our simulations, we implemented the most common case-marking systems: i.e. accusative/nominative, absolutive/ergative, and neutral alignment (i.e. none of the participants are marked). Combined, they account for more than 90% of the large natural language sample analyzed in Siewierska (1996). As we will discuss below, we were able to further reduce this three-way distinction into a binary one – i.e. the presence of case-marking versus its absence – because the results of the simulations did not show significant differences between the nominative/accusative languages and their absolutive/ergative counterparts.

[3] There are other less frequently encountered case systems such as the tripartite one, in which different markers distinguish S from A as well as P (e.g. Pitta Pitta). We refer the reader to Blake (1994) and Falk (2006) for descriptions of these 'exotic' systems.

2.2.2 VERB MARKERS: AGREEMENT AND T/A/M

The first type of verb marking that we modeled in our simulations is agreement. Agreement is the matching of features between the verbal predicate and one or more nominal arguments in a sentence.[4] Typically, agreement markers on the verb share semantic or formal features with the subject (and objects) of the sentence (Nichols 1986; Corbett 2003). These features relate to the person, number and/or gender of the nominal arguments. In the case of person agreement, the marker on the verb indicates whether the nominal it agrees with is 1st, 2nd, or 3rd person. Number agreement usually distinguishes between singular and plural forms. Gender agreement, finally, implies that the form of the marker on the verb changes depending on other properties of the nominal argument (e.g. masculine vs feminine vs neuter as in many Indo-European languages, animate vs inanimate as in Algonquian languages). Agreement markers are often used to encode information about 'who did what to whom', as illustrated in the following sentences from Yimas (a language spoken in Papua New Guinea; Foley 1991). Note that neither the order nor the form of the nouns in the sentence is different between the two sentences. The agreement markers on the verb indicate which noun is the subject and which the object.

---

[4] This sharing of features distinguishes agreement from valency. Valency markers attach to a verb and indicate whether this verb is used transitively or intransitively (e.g. in the creole Tok Pisin, the suffix *–im* attaches to transitive verbs – McWhorter 1998). Although valency marking is uncommon cross-linguistically, we included such markers as an option in our early simulations. We discovered that it had essentially the same effect as the presence of the T/A/M marker described below.

(5) payum narmaŋ na-mpu-tay
man-PL woman-SG 3SG-3PL-see
"The men saw the woman." (Yimas)

(6) payum narmaŋ pu-n-tay
man-PL woman-SG 3SG-3PL-see
"The woman saw the men." (Yimas)

Cross-linguistically, agreement phenomena can be found in a large majority of the world's languages, though it is not as frequent in verb-medial languages (61%) as in verb-initial (88%) or verb-final (83%) ones – these numbers are from Siewierska (1996); see also Nichols (1986, 1992). Subject agreement is considerably more common than agreement with the direct object or indirect object, though the same alignment systems which we have described above also appear with agreement. In nominative/accusative languages, the same marker(s) on the verb will be present for the subjects of both transitive and intransitive clauses. Similarly, many absolutive/ergative case-marking languages use absolutive agreement, where the agreement marker cross-referencing the subject of intransitive clauses is also used for the object of a transitive clause. In our simulations, we again implemented the most common agreement systems: i.e. accusative/nominative, absolutive/ergative, and no agreement at all. As with case marking, we were able to further reduce this three-way distinction into a binary one – i.e. the presence of agreement versus its absence – because the results of the simulations did not show significant differences between the nominative/accusative languages and their absolutive/ergative counterparts.

Unlike agreement, the second type of verb marking specifies properties of the event described by the verb itself. These markers are commonly referred to as T/A/M markers because they express tense, aspect, and modality features for the language involved (Comrie 1976; Chung and Timberlake 1985; Bybee, Perkins and Pagliuca 1994). We cannot hope to do justice here to the complexity of these notions (and their interactions), but each of them characterizes the event described by the verb in a different way. Tense is easiest to grasp as it refers to the time axis and denotes the relation of the event to some reference point on that axis, usually coinciding with the moment of speech. The event can be prior to the moment of speech (past tense), coincide with the reference point (present tense), or precede it (future tense). The notion of aspect places the event in a larger time frame, and can (among many other things) distinguish whether the action is only just starting (inchoative), on-going (progressive), or has finished (perfective). The third notion, modality, can indicate whether the event has actually taken place (realis), is possible (potentialis) or impossible (irrealis), or is something that we would like to happen (optative), or that simply must take place (deontic modality).

Each event inherently has T/A/M features, but they can range from relatively straightforward (present, progressive, realis as in *he is swimming*) to quite complex (present, inchoative, potential in *he could start swimming*; past, perfective, deontic/irrealis as in *he should have finished swimming*). These examples from English illustrate that T/A/M values can be expressed by separate words such as modal verbs (*could, have*), but also by morphological markers such as *–ing*. Each natural language is different in

how it implements T/A/M marking, so we find languages such as Mandarin Chinese and Thai in which T/A/M properties are expressed via separate words, as well as more morphologically complex languages such as Turkish and Spanish in which various T/A/M markers can be stacked on a single verb and express an event as complex as *he should have started swimming* in a single verb form.

For our simulations, implementing a full-blown T/A/M feature system would have been impossible, if only because these conceptual notions are hard to describe algorithmically. However, we did want to include the cross-linguistic observation that some languages consistently add T/A/M markers to their verbs, while others do not. Natural languages also have relatively closed sets of such markers, so some of them are usually quite frequent in any given language. To model the effect on language learnability of an easily recognizable T/A/M marker, we used a binary T/A/M parameter in the simulations – i.e. there was a single such marker that was either present on all verbs, or absent from all of them.

2.3 Pro-drop

It is almost certain that all of the world's languages allow subjects to remain unexpressed in some types of sentences. For example, they are typically not expressed in imperatives, such as *drink your milk!* The pro-drop parameter which we are interested in has a much more narrow definition, and a much more limited cross-linguistic distribution (Gilligan 1987; Jaeggli and Safir 1989; Nicolis 2005). Moreover, the search for the necessary and sufficient conditions for pro-drop has been a topic of much

discussion over the years (see below). In order to clarify what we mean by pro-drop, we first present language data to illustrate the various kinds of unexpressed subjects. We will ignore constructions like the imperative because the definition of pro-drop has traditionally been limited to simple declarative clauses.[5] Also, we will not concern ourselves with unexpressed objects, a phenomenon that occurs far less frequently than subject pro-drop and that has received less attention (but see Cole 1987; Huang 1995; Speas 1996; Kim 2000).

A first kind of unexpressed subject can be found in single clause sentences where the subject has no real semantic reference. Examples are given in sentences (7) and (8). Notice that the English translations require the expletive element *it* to be grammatical. (The example from Urdu actually has an optional pro because the expletive *yi* 'it' is also possible (Huang 1995). In the Spanish example, no expletive element is allowed.)

(7) ∅ está lloviendo
pro is raining
"It is raining." (Spanish)

(8) ∅ vaazeh hai ki us ne jhoot bola thaa
pro obvious is that he-ERG lie spoken was
"It is obvious that he had lied." (Urdu)

[5] Unexpressed subjects are very common in complex sentences where the same referent is the subject of both clauses – e.g. *John went to the store and then ∅ bought an umbrella.* Such sentences are perfectly acceptable in English, but * *∅ bought an umbrella* is not.

The absence of expletives (or at least their optional presence) is a typical feature of pro-drop languages. However, similar phenomena can occasionally be observed in languages which are otherwise not tolerant of pro-drop. In Dutch, for example, it is possible to leave the normal expletive subject *er* unexpressed if another constituent has filled the structural position in the clause it would normally appear in. This is only possible when the referent of the subject is generic – see sentence (9) below (cf. Maling and Zaenen 1978). Similarly, even English marginally allows sentences such as (10) in which the usual expletive *it* has been left unexpressed.

(9) Gisteren is ∅/er hard gewerkt in de fabriek
yesterday is pro/it hard worked in the factory
"They worked hard in the factory yesterday." (Dutch)

(10) ∅ Seems like we have a problem.

The second type of unexpressed subject can be found when the subject refers to an underspecified entity. I.e. the sentence implies the existence of an actual referent but it is described as being arbitrary or generic in nature. For example, in the Spanish sentences below, the exact identity of the book editors in (11) and the bike thief in (12) is not known. As with the previous type, the literal English translations require the presence of an overt subject form. (More natural translations would use passives instead, e.g. *His bike was stolen*.)

(11) ∅ editan libros ahi
pro edit-3PL books there
"They edit books there." (Spanish)

(12) ∅ han robado su bicicleta
pro have-3PL stolen his bicycle
"Someone has stolen his bicycle." (Spanish)

The third kind of unexpressed subjects bring us close to the type of pro-drop we are interested in. The unexpressed element is now truly referential in nature, i.e. the speaker has in mind a specific referent in the world. For example, in the Spanish sentence (13), the 1st person pronoun *yo* 'I' is not overt. However, even English can still allow somewhat similar sentences with an unexpressed referential subject. In examples (14) and (15), the unexpressed subject is the speaker (cf. the 'diary drop' phenomenon – see e.g. Haegeman (1999; Haegeman and Ihsane 2001). It is important to note that English sentences like these are marginal. Referential pro-drop is only allowed in English in specific kinds of constructions – compare (14) to the far less acceptable **Drank a beer* – and it can only refer to the deictic 1st and 2nd person participants which are immediately salient in the conversation setting – a sentence such as (15) simply cannot mean 'He thought I saw something outside'.

(13) ∅ he bebido una cerveza
pro have.1SG drunk a beer
"I have drunk a beer." (Spanish)

(14) ∅ gone fishing

(15) ∅ thought I saw something outside

So, for a language to meet our pro-drop criterion, it must not only allow unexpressed subjects that are referential in nature, but these subjects must be able to refer to 3rd person entities. As the following two examples illustrate, sentences that meet these two criteria can easily be found in a language such as Spanish.

(16) ∅ ha visto un caballo
pro has seen a horse
"S/He has seen a horse." (Spanish)

(17) ∅ quiere comprar un coche nuevo
pro wants buy a car new
'S/He wants to buy a new car.' (Spanish)

Actually, leaving the subject unexpressed in such cases is generally the preferred linguistic strategy in pro-drop languages. More than half of the sentences in everyday language corpora of Spanish, Italian, Thai and Mandarin lack an overt subject (Bentivoglio 1992; Chui 1992; Tao 1996; Tardif 1996; Aroonmanakun 1999; Ueno and Polinsky 2002). When a pronominal subject does occur, it often has a contrastive meaning, which makes a literal translation with a neutral reading inappropriate. For

example, the Spanish sentence *Yo e bebido una cerveza* is closer in meaning to English 'It is I who have drunk beer' than to the neutral 'I have drunk a beer'.

An attentive reader may point out that our Spanish pro-drop examples all have one crucial feature in common, i.e. the verb has sufficient agreement marking to indicate the person and number of the unexpressed subject. In other words, one could argue that nothing is really missing at all. The same information that is expressed by pronouns in languages such as English and Dutch is carried by a morphological agreement marker in pro-drop languages. The observation of exactly this pattern led early analyses of pro-drop phenomena to postulate that the existence of rich verb agreement in a language was a necessary (and perhaps sufficient) condition for wide-spread use of unexpressed subjects (see Perlmutter 1971; Taraldsen 1980; Chomsky 1981; Jaeggli 1982; Rizzi 1982; Jaeggli and Safir 1989; Nicolis 2005). However, it was soon pointed out that the rich agreement condition does not hold for numerous other languages that also feature pro-drop, such as Mandarin Chinese and Japanese (Huang 1984, 1989; Gilligan 1987). In these languages, the absence of agreement features leads to considerable ambiguity with respect to the referent of the unexpressed subjects. In the following example from Mandarin, neither the person nor the number of the subject can be recovered from the sentence.

(18) ∅ lai-le
pro come-PERF
'[I/You/She/He/We/You/They] came.' (Mandarin)

We will return to the pro-drop phenomena in Mandarin in section 4.2.2 below when we discuss the modeling results, but here it suffices to say that many theoretical accounts of pro-drop nowadays distinguish between the 'rich agreement' variety exemplified by Spanish, and the 'discourse pro-drop' variety of Mandarin (Huang 1995; Huang 2000). The motivation for the latter name derives from the observation that the most important condition on pro-drop in languages such as Mandarin, Thai, Yiddish and Japanese is that the unexpressed subject must be a salient entity in the current discourse (Li and Thompson 1979; Prince 1998; Huang 2000; Shi 2000). Saliency can be a result of linguistic strategies – for example, having been mentioned explicitly in a prominent position in a previous sentence – but corpus studies have found that the unexpressed subject can also refer to a more general idea which has only been touched upon indirectly (but frequently) in the preceding context (see Li and Thompson (1979) and Li (1997) for examples). Because these two types of pro-drop appear to be quite distinct, it has been argued that they bear little relationship to each other (Bybee 1997; Li, Bates et al. 1992; Li 1998; Yang, Gordon and Hendrick 2006). As a result, speakers and hearers in discourse pro-drop languages might be producing and parsing sentences in fundamentally different ways from speakers of other languages. In these languages, the reasoning goes, grammatical cues are of far less importance than semantic ones, and speakers and hearers must engage in complex kinds of inferencing to disambiguate utterances.

The desire to investigate the validity of this distinction is one of the reasons we chose to implement pro-drop in our models. Our simulations make it possible to

Table 2. The space of possible language types defined by the three linguistic parameters in our simulations – i.e. case-marking on the nouns, verb marking, and pro-drop.

– PRO		N-marking	
		—	Case
V-marking	—		
	T/A/M		
	Agr		

+ PRO		N-marking	
		—	Case
V-marking	—		
	T/A/M		
	Agr		

compare how well two otherwise identical languages can be learned, if one has pro-drop with rich agreement (like Spanish), and the other one has pro-drop without such agreement (like Mandarin Chinese). With the same learning mechanism trying to acquire both, we can measure how much useful information is available in the input, despite the absence of rich semantics and the ability to make inferences. If it turns out that our simple neural networks can use the surface cues to determine 'who did what to whom' in such languages, then we consider it reasonable to assume that young children with more limited linguistic capabilities might do likewise.

2.4 Summary

In this chapter we have described the linguistic parameters that describe the space of possible languages our models have to learn. All these languages have SVO word order. Some have case-marking on their nouns, but others do not. With respect

to verb marking, some have rich agreement, some only a Tense/Aspect/Modality marker, and some no morphology on the verbs at all. Finally, some of the languages have pro-drop, but others do not. The space of possible languages is summarized in Table 2. We will use the format of this table to present the results of our experiments.

Chapter 3. Methods

We used artificial neural networks to set up our experiments. These networks are vaguely brain-like computational mechanisms consisting of interconnected units that can learn to associate specific patterns represented over a group of 'input' units (e.g. words in a sentence) with specific patterns over a group of 'output' units (e.g. their function in the sentence). We refer readers who are unfamiliar with neural network basics to Rumelhart, Hinton and Williams (1986) and Bechtel and Abrahamsen (1991) for relatively accessible introductions to how they function. A more mathematically oriented treatment can be found in Bishop (1995).

In the remainder of this section, we provide a more technical discussion of how we implemented our simulations. First, we describe the artificial languages that were generated to represent the various language types. Second, we present the architecture of the actual networks and how the units and their connections were organized. We end by going over the procedures for training and testing our networks on the corpora of each language type.

3.1 The artificial languages

As we wanted to test the various language types allowed by our three linguistic parameters, we first wrote a simple context-free grammar to generate the maximal type – i.e. the SVO language with case-marking, rich verb agreement, a T/A/M

(Tense/Aspect/Modality) marker, and no pro-drop. We then proceeded to selectively remove properties of this maximal type to create the various language types we wanted to compare. The context-free grammar consisted of formal rules to rewrite an initial S(entence) symbol into an N V N sequence (for transitive SVO sentences) or into an N V sequence (for intransitive SV sentences). The transitive and intransitive sentences were chosen with equal probability. The N symbols were then either replaced with nouns or pronouns selected from the lexicon, or deleted to create an unexpressed subject (see Table 3 for the probability of each scenario). Following cross-linguistic data collected by Ueno and Polinsky (2002), unexpressed subjects were twice as likely to appear in transitive sentences as in intransitive ones. Also following natural language data (Du Bois 1987), we configured the grammars to produce more nouns than pronouns in object position, but at least as many pronouns as nouns in subject position. The V symbols were simply replaced with an appropriate verb from the lexicon.

In order to implement case-marking, affixes were attached to all the nouns in the sentence. In the simulations of the accusative alignment system, there was a single (nominative) Subject affix which attached to the subjects of both transitive and intransitive sentences.[6] A different (accusative) Object affix was attached to all the

[6] In natural languages, case affixes often change shape depending on the gender and/or number of the noun they attach to. However, we feel there is no reason to believe a more complex case system including features like gender and/or number would have interacted in an interesting manner with the other parameters in our study.

Table 3. The actual realization of subject noun phrases in the artificial languages depended on whether they appeared with transitive (A) or intransitive (S) verbs. Unexpressed subjects were twice as likely with transitive verbs. Object noun phrases (P) were always realized overtly.

NP	Noun	Pronoun	∅
Subject (S)	25%	50%	25%
Subject (A)	25%	25%	50%
Object (P)	75%	25%	—

object nouns. In the ergative simulations, a single (absolutive) affix was used for intransitive subjects and objects. The second (ergative) affix only occurred with transitive subject nouns. When the language type to be generated was marked as [+case], these case affixes were all present. In [-case] types, they were all deleted. Pronouns in the simulations always occurred case-marked, as this is cross-linguistically the most common pattern (Siewierska 1996, 1998).

The verb marking parameter had three options. The implementation of T/A/M marking was quite simple: there was a single such marker, and it was either present on all verbs in a language, or absent from all of them. The implementation of agreement was much more complex. The artificial languages with agreement featured eight distinct verb markers depending on the person (1st, 2nd, 3rd), number (singular, plural) and gender (animate vs. inanimate) of both the subject and object in the clause. So, a verb could carry markers signaling its subject was animate 1st person plural (i.e. *we*) and its object inanimate 3rd person singular (e.g. *it* or a noun). If an artificial

language was marked for agreement, these markers would appear on all verbs in the language. Otherwise, there would be no agreement markers at all. Also, the same alignment system used for the nouns and pronouns would apply to the agreement marking, so a language could be consistently nominative/accusative or consistently absolutive/ergative, but not mixed.[7]

The lexicon used to generate the sentences in the training and test corpora contained 100 verbs (50 transitive; 50 intransitive), 8 pronouns (1st, 2nd, and 3rd person animate, as well as 3rd person inanimate; each of them in both agent and patient versions), and 300 nouns (150 'animate'; 150 'inanimate' – the latter were more likely to occur as objects (70% vs 30%) while the former were more frequently chosen as subjects (also 70% vs 30%). Note that the disparity between the type frequencies of the nouns and the verbs is based on a universal pattern observed in natural languages; nouns are typically far more numerous than verbs (Gentner 1981).

3.2 Network architecture

The architecture of the models used in our simulations is shown schematically in Figure 1. It consisted of a simple recurrent network (cf. Elman 1990) augmented with a separate recurrent layer for the output units (see below). The solid arrows in Figure 1 indicate full interconnectivity between the layers, so each unit in the source layer was connected to each unit in the destination layer. The dashed arrows leading to

[7] We have actually run thousands of simulations with various kinds of split-accusative and split-ergative language types, but we have not yet analyzed the results.

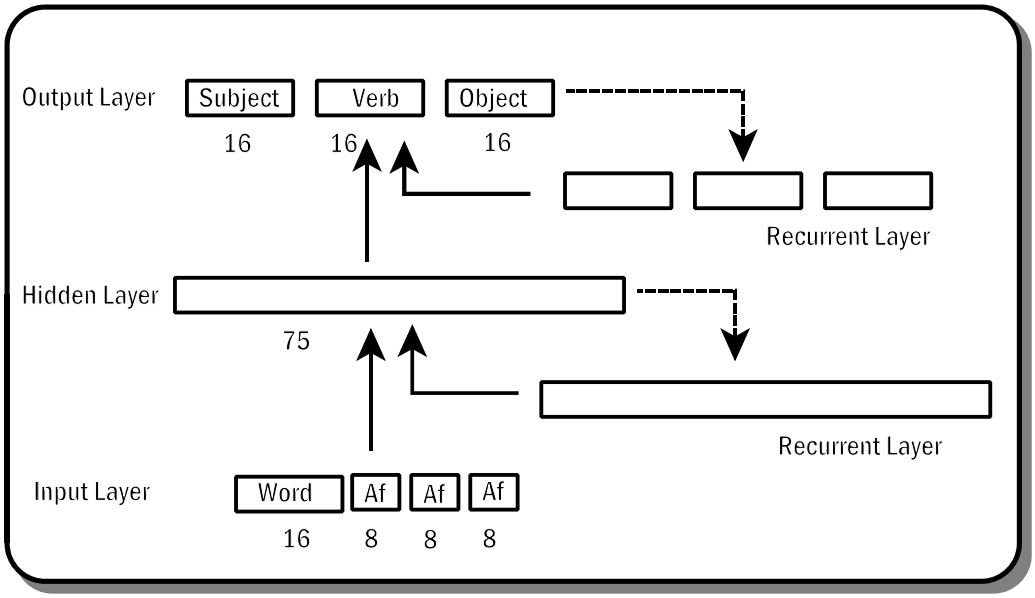

Figure 1. Architecture of the neural network. The input layer consists of four groups of units: one to represent a word, and three to represent affixes such as a case marker of agreement markers. The hidden layer receives input from both the input layer and a dedicated recurrent layer that stores the network's representation of the entire sentence seen so far. The output layer contains three groups of units to represent the words that have occurred as the subject, object and verb in the sentence. The output layer also has its dedicated recurrent layer to help the network 'remember' its interpretation of the sentence seen so far.

the recurrent layers indicate one-to-one value copying: i.e. at the end of each network update, the value of a unit in the source layer was copied exactly to the corresponding unit in the recurrent layer. The numbers in parentheses show how many units were in each layer. Within every layer except for the input layer all units also had connections to every other unit.

At the input layer, the individual words in each sentence were presented one at a time. The first 16 units were reserved for the representation of the word itself (i.e. noun, pronoun or verb). The remaining 24 units fell into 3 groups of 8 units, with each group potentially representing a morphological marker. Pronouns, which were inherently case-marked in the simulations, only used the main 16 units. A case-marked noun used the main 16 units for the noun and the first group of 8 units for the case marker. Similarly, verb agreement markers filled the first two groups of 8 units, while the T/A/M marker appeared in the last one. The actual representation for each word or marker was a unique string of 1's and 0's, with 6 to 8 of the 16 bits set to '1' for each word, and 2 or 3 of the 8 bits set to '1' for each marker.[8] There were also two special input patterns which followed each sentence. The first one signaled the end of a sentence to the network; the second one told it to reset the activation values of the output units in preparation for the next sentence.

The output layer consisted of 3 groups of 16 units each as this was sufficient for the networks to build up a representation of the entire sentence. The first group was intended to hold the pattern for the subject of each sentence (if present), the second group the pattern for the verb, and the third group the pattern for the object (if present). The desired output pattern for each word was identical to its input pattern.

---

[8] In Van Everbroeck (2003) we used more linguistically inspired input patterns, with words consisting of syllables which in turn were represented by phonemes and their phonological features. However, testing showed that switching to the shorter bit strings sped up learning without significantly affecting the network results.

Let us illustrate this with a simplified example. Imagine the SVO sentence '1001' (S), '0110' (V), '1010' (O). Given that the network only sees one word at a time, it would first see the subject pattern '1001'. The target output layer at this point would be '1001', '0000', '0000' – the target for the V and O groups is the blank '0000' pattern because no information about these words has been processed yet. After seeing the verb, the target output was '1001', '0110', '0000'. And after having seen the object, the correct output was '1001', '0110', '1010'. Although the network only processes one word at a time at the input layer, this output representation allows us to analyze its representation for the entire sentence. Note that the desired output for an unexpressed subject is a pattern of all zeros, as there is no discourse available from which the unexpressed subject can be reconstructed.

The purpose of the first context layer in the architecture is to provide the network with a 'memory' of the internal representational which it is has constructed of the sentence up to that point (Elman 1990). This enables it to integrate the incoming word with its representation of the sentence so far. In cases where the incoming word itself is ambiguous, the network may thus be able to disambiguate this word using the information it has stored about previous words in the sentence. The recurrence at the output layer simply allows the network to better "remember" its current interpretation of the sentence. We will see below that this recurrent layer at the output makes it difficult for the network to override its current interpretation of the sentence, but this turns out to model some interesting phenomena in child language processing.

3.3 Training and testing

To make sure our learnability results for each combination of language type were robust, we created 10 unique networks (i.e. experimental 'subjects') to learn each such combination of language parameters. As we used a different random seed for the initialization of the weights on the connections, the networks were also unique between different language types. Following standard experimental procedures, we discarded networks if they produced results which were more than two standard deviations removed from the mean of the relevant language type, and then trained and tested a new unique replacement network. This occurred infrequently. The results which we present below for each language type are thus averages over at least 10 unique networks.

To further ensure that the simulations produced reliable results, we also generated 10 different training corpora and 10 different test corpora for each language type. Each 'subject' was trained on a different subset of the relevant language. The corpora contained 3,000 sentences each, split evenly between transitive and intransitive verbs. Once a network had been trained on a given corpus, it could then be tested on a number of different test corpora. In all the cases we report here, the test corpora were generated from the same grammar which was used to create the training corpus. It would be easy to test how well a network trained on language A performs on language B, but it is not obvious to us how we would interpret such results.

The simulations were implemented using the Stuttgart Neural Network Simulator package. The learning algorithm was standard backpropagation, with small

learning rates (Rumelhart, Hinton and Williams 1986). We first trained all the networks for 10 epochs (i.e. 10 cycles through the full training corpus), but continued training for up to 100 epochs in some cases (see below). The error measure which we report is the percentage of sentences in a test corpus which was analyzed correctly. For a sentence to be counted as correct, we looked at the patterns of activation over the three groups of output units at the end of each sentence (i.e. after the end-of-sentence pattern was presented at the input layer). We then compared the activation pattern for each group to all patterns from the training and test lexica. Only if the actual activation pattern was closer to the target word (using Euclidean distance) than to all other words did a group count as correct. Only if all three groups (subject, verb and object) were correct did we count the entire sentence as correct.

3.4 Summary

In this chapter, we have described how we implemented our neural network models. We covered the properties of the grammar and lexicon used to generate the various artificial language types that constitute the language space we explore below. We also discussed the architecture of the neural networks and the types of representations used at the input and output layers. Finally, we described the procedures used to train and test the models of each language type.

Now that we have presented both the linguistic parameters (Chapter 2) and the details of the implementation, we can finally move on to the actual experiments.

Chapter 4. Experiment 1

In our baseline experiment we look at the interactions between the three major parameters in our simulations: pro-drop, case-marking, and verb marking. As we have mentioned in the linguistic background section, current linguistic analysis distinguishes 'agreement pro-drop' in languages such as Spanish from 'discourse pro-drop' in languages such as Mandarin. For the former type, the argument goes that the information carried by the agreement markers makes it possible to recover the unexpressed subject. For the latter, the missing information must be found outside of the sentence, in the linguistic (or even non-linguistic) context. The absence of semantic knowledge in our models entails that we can only directly model the agreement pro-drop languages. On the other hand, it allows us to determine how much of a discourse pro-drop language can still be learned on the basis of purely structural and formal information.

The models also make it possible to gauge the effect of pro-drop in the presence of not only agreement marking, but also case-marking, and a simple T/A/M marker. Cross-linguistic data suggests the languages without pro-drop should not present a problem for the models, and it is expected that the presence of rich agreement should be beneficial in pro-drop languages. The two questions we want to investigate in Experiment 1 are, first, under which linguistic conditions SVO languages

Table 4. Possible sentence structures in simple SVO languages with and without pro-drop.

SVO	1	2	3
Intransitive	S	V	
Intransitive, ∅	V		
Transitive	S	V	O
Transitive, ∅	V	O	

with pro-drop are learnable by neural networks, and, second, whether these conditions correspond to the ones which are attested in natural languages with similar features.

Before we present the network results, let us briefly consider why pro-drop can have a negative effect on the learnability of our artificial languages. The answer is quite straightforward if we compare the sentence structures in simple SVO languages with and without pro-drop (see Table 4). In fixed word order SVO languages without pro-drop, there are never any parsing inconsistencies between transitive and intransitive sentences: the first word of any sentence is guaranteed to be the subject and the second word is guaranteed to be the verb. If a third word occurs, it has to be the object. The presence of pro-drop in an SVO language causes this simple parsing method to break down. Instead of two possible linear orders of elements (SV, SVO), there are now four (V, SV, SVO, VO) and the function of neither the first nor the

second word can be taken for granted. Both the subject and the verb can occur in the first position, and both the object and the verb in the second position. Simply put, when an SVO language has unexpressed subjects, it can not be parsed as easily. Consequently, we expect that our SVO languages with pro-drop will be harder to learn than their non-pro-drop counterparts.

4.1 Network results

The simulation results for Experiment 1 are summarized in Table 5. Each number in the table is an average of the percentage of the 3,000 test sentences that were parsed correctly in the twenty simulations (10 accusative; 10 ergative) which modeled the SVO type as defined by the three linguistic parameters.[9] It is easy to see that *all* SVO types were learned well enough to allow for excellent generalization to new sentences (at least 98.7%), except for the type which features pro-drop but no head-marking or case-marking (73.4%). An ANOVA test shows that there is a main effect of the pro-drop parameter: the average of the test set results of all the SVO languages with pro-drop (94.7%) is significantly worse than the average of the

[9] An ANOVA test shows that there is no main effect for the alignment type: the average performance was 96.9% on all the accusative languages, and 97.2% on all the ergative languages, $F(1, 234) = 0.14$, $p > 0.71$. Cross-linguistic samples have shown that the nominative/accusative type is actually more common (63%; Siewierska 1996) in SVO languages than the absolutive/ergative one (33%), but both occur with sufficient frequency that we do not consider the models' ability to learn either one equally well a problem. In all the results presented below, the [+case] results will thus be the averages of the simulations with either kind of marking; the [-case] results are derived from the languages with the neutral alignment system.

Table 5. Results of Experiment 1. Average percentage of test sentences analyzed correctly by models learning each type of SVO language, as defined by three parameters: pro-drop, case-marking, and head-marking. Each number averages over 10 accusative simulations and 10 ergative ones.

– PRO		N-marking	
		—	Case
V-marking	—	99.5%	99.5%
	T/A/M	99.4%	99.5%
	Agr	99.4%	99.4%

+ PRO		N-marking	
		—	Case
V-marking	—	73.4%	98.8%
	T/A/M	99.2%	99.3%
	Agr	98.7%	99.0%

languages without it (99.5%), $F(1,234) = 40.3$, $p<0.0001$. This appears to confirm our prediction that pro-drop makes languages harder to learn. However, these percentages obviously average over many different types of SVO languages, including a large number that feature some at least some kind of marking which may mitigate the effect of pro-drop. It is more appropriate to look at the individual cells in Table 5, and just compare the twenty networks that learned the SVO language with pro-drop but without any kind of morphological marking (73.4%) to the ones that learned the equivalent language without pro-drop (99.5%). The significant effect of pro-drop is confirmed by a post-hoc Tukey test; MS = .82957, df = 228.00, $p<0.0001$.

The results in Table 5 also demonstrate that agreement marking is not the only antidote for the inconsistencies created by pro-drop. In fact, any kind of additional information is sufficient for the networks to generalize well to their test sets (> 98%). Post-hoc comparisons using Tukey tests show that the differences between the

pro-drop models with only case (98.8%) or only T/A/M marking (99.2%) or only agreement (98.7%) are not statistically significant . In these simulations, both case-marking and head-marking appear equally capable of compensating for the effect of pro-drop. Moreover, t-tests also show that there are no significant differences between any of the languages without pro-drop and those with pro-drop – with the exception of the one difficult language, of course. I.e. the no-marking pro-drop language type is significantly different from all the other types, but none of the other pairwise comparisons reach significance.

These results suggest there is no real learnability benefit for the languages that combine two or three sources of information. When there is no pro-drop, adding an extra feature does not affect how well the language can be learned; but when pro-drop is present in the language, any kind of marking improves learnability by at least 25%. At least for these simple artificial languages, the agreement markers do not appear to provide any information which cannot also be extracted from a consistent word order, case-marking or more limited head-marking. This finding is intriguing because rich agreement obviously contains more information than a simple T/A/M marker.

To determine why the different sources of information appear equally useful to the networks, we take advantage of one of the convenient features of computational modeling by opening up the black boxes and looking inside. We start with the networks learning the problematic language type, i.e. with pro-drop but without any marking, because they reveal the main source of the errors made by the networks. Figure 2 shows how well the networks performed on the subject, verb and object

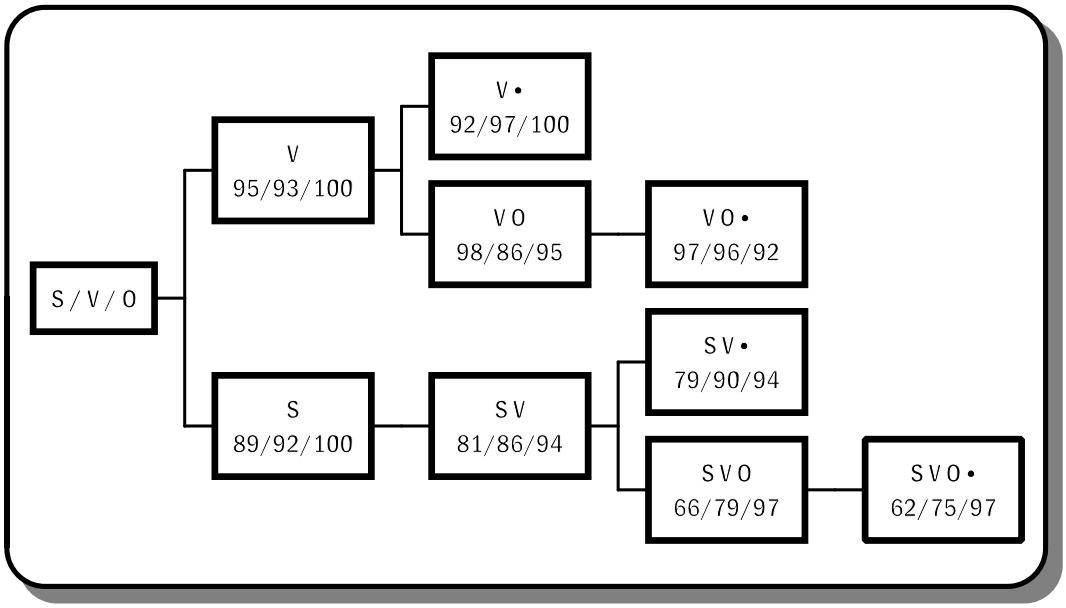

Figure 2. Word-by-word analysis of the performance on the four sentence types of the networks learning the difficult language type after 10 training cycles. The numbers in each box show the percentages of Subjects, Verbs, and Objects that were analyzed correctly after having seen the part of the sentence shown in the same box. E.g. after having seen just the Subject, 89% of the Subjects, 92% of the Verbs, and 100% of the Objects were correct at the output.

output groups after each word of the four possible sentence types (SV, V, SVO, VO). The first number in each cell is the percentage of subjects which were parsed correctly after the sequence shown; the second number is for the verbs, and the third for the objects. Each number is an average over the results from the 20 network which learned this language type. Figure 2 reveals that the networks learning the difficult language run into problems at the very beginning of every sentence: i.e. they only get 92% of the sentence-initial verbs and 89% of the sentence-initial subjects correct. In both cases, all errors are related to confusion between subjects and verbs, with the

networks apparently incorrectly categorizing some subjects as verbs and vice versa. (Note that performance on the object group is 100% in both cases, so the networks have learned that no sentence in their language can begin with an object.) Things change when the second word has been seen. When it is the object (in VO), the networks have little trouble parsing it correctly (95%). They also appear to deduce the subject group must be empty because its performance improves a bit (from 95% to 98%). However, the verb group does worse (86%), and this evolution continues after the presentation of the end-of-sentence symbol (75%). A different pattern can be observed when the second word is the verb (in SV): this time, the percentages of both subjects and verbs that are correctly analyzed drops steadily as the sentences continue, irrespective of whether the sentence is transitive or intransitive. The object group, however, still does not suffer from any major problems.

The errors in Figure 2 are the result of two different factors. The first one is memory degradation: as a sentence gets longer, the networks find it harder to maintain the correct strings of 1's and 0's for the early words. So, overall performance on the shortest sentence type (V, 91%) is noticeably better than performance on the two-word sentences (SV, 79%; VO, 73%), and the longest sentence type does worst (SVO, 58%).[10] It is not surprising hat neural networks should find it hard to store

[10] Note that the frequency of the four language types in the corpora is not a good predictor of their performance: the one-word sentence V actually occurs least frequently (13%), but still has the best performance; the most common type (SV, 37%) only performs about as well as VO (25%); and the sentence type SVO, which also occurs in 25% of the sentences, does worst. While it is beyond doubt that neural networks are sensitive to frequency – we will discuss some such examples below – its effect can be overridden by other factors.

precise (symbolic) strings, but the explanatory value of this factor is limited when we take into account that the networks learning all the other language types did not experience any problems keeping the correct words active in the output groups.

The second, and crucial, factor behind the errors is the subject/verb confusion which occurs when the networks see the first word in the sentence. In roughly 10% of the sentences, the networks seem uncertain what the correct output group is for the first word at the input layer, and end up copying it partially to both the subject and verb groups. They then never recover from this initial confusion, even when (at least theoretically) sufficient information becomes available later on in the sentence. For example, every three-word sequence must be an SVO sentence; the first word must thus be the subject. The inability to recover from an original parsing uncertainty can be interpreted as a weakness of the models. It is the result of the network architecture which uses the recurrent layer at the output to maintain a copy of the network's sentence analysis after the previous word. This strategy works well if the initial parse is correct, but it also causes initial errors to get fed back into the output layer and thereby perpetuates these errors. While such mistakes are sub-optimal from a machine learning perspective, they actually resemble some of the processing mistakes made by young children. For example, Trueswell, Sekerina et al. (1999) have reported that the 5-year-old subjects in their eye-tracking study often failed to recover from initial ambiguities in the garden-path sentences they were asked to parse. Self-paced reading experiments with older children – 6 and 7-year-olds in Felser, Marinis and Clahsen (2003) and 8 to 12-year-olds in Traxler (2002) – have also shown that children are far

more likely than adults to choose a simple syntactic analysis for a structurally ambiguous sequence even when conflicting semantic information is available. As a result, we consider the ability of the models to trip over problems in its input a valuable property for drawing attention to issues which need further exploration.

Finally, we need to consider exactly why the networks learning the pro-drop language without any marking fail to distinguish between some sentence-initial subjects and verbs. Given that the networks have been trained on all the words which appear in the test corpora, one would assume they should know the lexical category every word belongs to, and hence they should be able to accurately determine its function: if the first word is a noun, it must be the subject; if not, it must be the verb. The networks are capable of doing this correctly most of the time, but it turns out there are some nouns in the lexicon which they have not been trained on enough times to be able to categorize them correctly. Recall that nouns in the simulation are significantly less common in subject position than object position. When neural networks are trained, they initially pay most attention to frequent patterns in their input. The subject nouns appear with lower frequency than all the other categories (object nouns, subject and object pronouns, and verbs) so they are not trained on as much, and consequently are not learned as well. For some of them, this leads to confusion when they appear in sentence initial position in the test corpus. A smaller source of errors results from inanimate nouns that are used in subject position 30% of the time. In some cases, the trained network will never have seen this noun as a subject during training, though it will know it well as an object noun. One may wonder whether the relatively poor

performance on the 'difficult' language type is thus an artifact of the limited exposure the neural networks had to the subject nouns in this type. We will explore this hypothesis in Experiment 2, when we train the networks for more epochs. However, it is important to point out again that none of the other language types suffered from a similar problem. In all the other pro-drop types, the available morphological marking provided sufficient information for the 'who did what to whom' task to be learned within 10 epochs of training.

The results also point towards another generalization, namely that it is sufficient to either mark verbs or nouns in SVO languages with pro-drop (marking both simultaneously doesn't make a significant difference). This in turn makes sense if we realize that the ambiguities which pro-drop creates in SVO languages are always between nouns and verbs: i.e. between the verb and the subject in first position, and between the verb and the object in second position (see Table 4). If either the verbs or the nouns are easily identifiable by markers, then interpreting the sentence becomes almost trivial, even when the words in the test sentence are not familiar. In essence, the problem with SVO and pro-drop is not distinguishing between the subject and the object – the decision generally considered to be a problem in the 'who did what to whom task' – but between the action and the actors. This is why the presence of the T/A/M marker, which does not carry any information about the subject or the object, is still sufficient in our models to compensate for pro-drop.

4.2 Linguistic discussion

So far, we have looked at how well our neural network models can learn various kinds of artificial SVO languages with and without pro-drop. We will now compare these results to what is known about natural languages. Let us first consider SVO languages without pro-drop. We have seen that such languages are always easily learnable by our model, even if they don't have any kind of case-marking or head-marking. On the other hand, we have also seen that adding marking does not have a negative impact on the learnability of our SVO languages. The markers provide redundant information as to 'who did what to whom' and may well end up being ignored by the models. As far as the simulations are concerned, all such languages are learnable and could be expected to exist in the real world.

With respect to the SVO languages with pro-drop, we have seen that the models only have problems learning the type in which pro-drop occurred in the absence of both case-marking and head-marking. As soon as any type of morphological marking is present, the networks are able to generalize their knowledge to the test corpora with no difficulty. The expectation is that all of the pro-drop language types could exist, with the possible exception of the type without noun marking and verb marking. Table 6 summarizes the natural language data and puts real languages into the same cells which we used for the models. It almost goes without saying that most natural languages do not fit easily into a single cell. For example, English has case-marking for its pronouns, and Dutch verbs can agree with their

subjects in number (singular vs plural), but not person (1st vs 2nd vs 3rd). The classifications in Table 6 are based on descriptions of the most common patterns in these languages.

It is easy to see in Table 6 that some combinations of the linguistic features appear to be more common than others. We need to stress, though, that there exist hundreds of SVO languages for which insufficient data are available to categorize them into Table 6 – as a result, many more Indo-European languages are included in the table than would be part of a representative sample of the world's SVO languages. With respect to the three main parameters, there are several observations worth noting (compare Siewierska 1996, 1998): first, SVO languages without case outnumber their [+case] counterparts; second, agreement marking is a common feature among SVO languages; third, SVO languages with pro-drop are more frequent than those without. However, the interactions between the parameters are what really interest us. Quite striking in Table 6 is that case-marking appears to depend on the presence of agreement marking, whereas the opposite is definitely not the case. This generalization holds not only for pro-drop languages, where the frequent absence of nouns naturally gives an advantage to the verbs (see section 6.2.1), but also in languages without pro-drop. This finding meshes well with the cross-linguistic observation – involving all word orders, not just SVO – that verbs are more likely to have morphological markers than other lexical categories (Comrie 1976; Nichols 1992).

The apparent lack of attested SVO languages with case-marking but without agreement means there are three cells in Table 6 for which our simulations suggest that

Table 6. Natural language counterparts to the language types defined by the three linguistic parameters in our simulations.

– PRO		N-marking	
		—	Case
V-marking	—	Vietnamese	
	T/A/M	Norwegian English Dutch Swedish	
	Agr	French Bulgarian	Icelandic Finnish Latvian Russian German

+ PRO		N-marking	
		—	Case
V-marking	—	? Creoles ? Mandarin	
	T/A/M	Hebrew Pasamaqoddy Sundanese Indonesian	
	Agr	Spanish Italian Au Bemba Swahili Kinande	Polish Romanian Greek Czech Albanian Estonian

such languages could indeed exist. I.e. we would not be surprised if these cells become populated (however slowly) as data about more natural languages becomes available. The most interesting cell, however, is the one for the 'difficult' language type. The results from the models in Experiment 1 suggest languages with pro-drop but no more morphological marking should be significantly harder to learn than other types, so that the question becomes whether natural languages, which are also learned, may avoid this particular combination of linguistic features. However, there are two groups of natural languages which have been argued to instantiate this very type: i.e. creole languages, and some of the languages spoken in South-East Asia. We will now consider languages from both groups to see to what extent such languages are indeed counter-examples to the prediction made by the models.

4.2.1 Creoles

Creoles are languages that have developed in complex contact situations in which speakers of different native languages need to communicate with each other (Bickerton 1981; Thomason and Kaufman 1988; McWhorter 1998; McWhorter 2001; Muysken and Law 2001). For most of the current Pacific Ocean creoles, the contact situation was one of trade negotiations; for the Atlantic Ocean creoles the typical scenario involved African slaves (often from different regions) who had been taken to European plantations. In both cases, simple 'pidgin' languages developed first to allow

for basic communication.[11] These pidgins combined elements from the various native languages involved, with the European languages generally contributing the majority of the vocabulary, but not necessarily the sentence structures. The result often appeared to be a simplified form of the European language. Pidgins become creoles when they are learned as a first language by children growing up in the communities in which they are spoken. During this transition, the creoles develop more complex syntactic features and end up as regular languages which are used by their native speakers for a wide range of communicative purposes. As these creoles have come into existence relatively recently, and in a spontaneous fashion to boot – i.e. without a clear end product in mind or schools and grammars to guide it – they are a great place to look for the role that cognitive learnability issues can play in language evolution (see Bickerton 1981; Lightfoot 1991; Kihm 2000).

What makes creoles even more interesting for our purposes is the fact that they share certain linguistic features, including an SVO word order and very limited inflectional morphology – i.e. no case marking and no rich agreement systems (Bickerton 1981; McWhorter 1998, 2001; Roberts 1999; Holm 2000; Muysken and Law 2001). Why creoles spoken on different continents and based on different languages share so many features is a controversial issue: according to the monogenesis hypothesis, all creoles derive from a single source language (Thomason and Kaufman 1988); others believe that the similarities derive from the combination of

[11] In more unusual cases, creoles have arisen without a previous pidgin stage. For example, Hawaiian Creole is the result of imperfect learning of English (as the dominant language) by a native population (Bickerton 1981).

an impoverished language input (the pidgin) and either the innate language learning mechanism (the bioprogram; Bickerton 1981, 1984), or the lack of time for more complex features to have evolved (McWhorter 1997, 2001).

In addition, numerous creoles derive partially from Spanish and Portuguese, two uncontested pro-drop languages (Huang 1984; Jaeggli and Safir 1989; Grinstead 2000; Barbosa and Torres Morais 2001). The question thus becomes to what extent their pro-drop phenomena have survived into the creoles. The results from Experiment 1 suggest that SVO languages with pro-drop but without morphological marking should be rare or non-existent. If morphologically poor SVO creoles with pro-drop do indeed exist, then they would obviously falsify the predictions of the models. On the other hand, their absence would provide strong support in favor of the validity of the network results. The loss of pro-drop during creole genesis is exactly what the networks lead us to expect, given that creoles are SVO languages with very limited morphological marking.

A survey of creole languages shows that they fall into different categories with respect to the types of pro-drop they exhibit (Mufwene 1988; Nicolis 2007). There are some creoles which, quite like English, only allow pro-drop in non-initial conjoined sentences, where the available discourse information is directly available and it is hard to come up with an interpretation of the unexpressed element that is not the correct one. Such languages include Palenquero, a Spanish-based creole spoken in Colombia (Schwegler 1993), Philippine Creole Spanish (Lipski 2002) and Gullah, an

English-based creole spoken on the southeast coast of the United States (Mufwene 1988). Despite their Spanish heritage, Palenquero and Philippine Creole Spanish do not allow any kind of pro-drop in basic clauses. What we find instead are sentences such as (18) from Palenquero (Friedemann and Patiño 1983: 225) and (19) from Gullah (Mufwene 1988: 238) in which the unexpressed element is easily recoverable from the discourse.

(18) Tigre á teneba de to: ∅ á teneba yuka ...
Tiger TNS had of everything pro TNS had yucca ...
'Tiger had everything: [he] had yucca ...' (Palenquero)

(19) 'I can't drink the wine ... [it] gives me a headache.'
'I know Lady is dead ... [I] went to Lady's funeral.' (Gullah)

It is also easy to find creoles which feature non-referential, expletive pro-drop. This group allows a null subject where English normally requires the expletive element *it*. Relevant creoles include Spanish-based Capeverdean Creole (Baptista 1995), Philippine Creole Spanish (Lipski 2002), Portuguese-based Papiamento (Kouwenberg 1990; Muysken and Law 2001) and Saramaccan (Byrne 1987), Gullah (Mufwene 1988), and French-based Haitian Creole (DeGraff 1993; Déprez 1994). The following two examples illustrate this kind of non-referential pro-drop in Capeverdean Creole (Baptista 1995:9) and Papiamento (Muysken and Law 2001: 54).

(20) ∅ sta faze friu
pro is making cold
'[It] is cold.' (Capeverdean Creole)

(21) ∅ parse ku Maria ta kanta
pro seem that Maria ASP sing
'[It] seems that Maria sings.' (Papiamento)

The second kind of non-referential pro-drop allows indefinite and generic referents to remain unexpressed in subject position. This pattern is attested in French-based Mauritian Creole (Syea 1993; Lipski 2002) and Papiamento (Kouwenberg 1990; Muysken and Law 2001). The following two sentences illustrate this phenomenon. Example (22) is from Papiamento (Muysken and Law 2001: 54) and example (23) from Mauritian Creole (Syea 1993: 92).

(22) ∅ ta bende flor
pro ASP sell flower
'[They] sell flowers (here).' (Papiamento)

(23) ∅ fin koke Pyer so loto
pro ASP steal Peter his car
'[Someone] has stolen Peter's car.' (Mauritian Creole)

None of the creoles mentioned so far allow referential subjects to remain unexpressed. For example, in Papiamento – a creole that is relatively tolerant of

pro-drop – we so not find referential null subjects in sentences such as (24) from Muysken and Law (2001: 54). The Spanish equivalent in (25) is perfectly grammatical.

(24) *∅ ta kome
pro ASP eat
'S/He is eating.' (Papiamento)

(25) ∅ Está comiendo
pro be-3sg eating
'S/He is eating.' (Spanish)

One creole which has been described as having real referential pro-drop, just like Spanish and Portuguese, is Haitian Creole (DeGraff 1993). However, the proposal given by DeGraff depends crucially on reanalyzing some of the language's pronouns as clitics – a move which appears to be theoretically driven – and it has been refuted by Déprez (1994) and Roberts (1999). For example, in sentence (26) from Déprez (1994: 11) it is possible to put the adverb *toujou* 'always' between the pronominal *li* and the verb *ap travay*; real clitics don't allow such intervening material, so it makes more sense to analyze *li* as a regular pronoun and avoid a referential pro-drop analysis.

(26) Li toujou ap travay fò
he always ASP work hard
'He is always working hard.' (Haitian Creole)

The situation is different for Bislama, an English-based creole for which Meyerhoff (2000) has argued that it has real referential pro-drop like Spanish. However, Meyerhoff also describes in detail how Bislama has only started allowing pro-drop *after* it developed a (rudimentary) agreement system. In example (27) from Meyerhoff (2000: 207), *i* is the agreement marker and it can co-occur with a full noun phrase. Also note that at least in this particular example, there is ample discourse available to recover the unexpressed element.

(27) Denis hem i kam. ∅ i blok-em hem
Denis 3sg AGR come [he] AGR block-TRANS.3sg..OBJ 3sg
'Denis came. [he] stopped her.' (Bislama)

A similar phenomenon can be observed in Portuguese-based São Tomé Creole, which only allows pro-drop for first person singular subjects. Note that this is the only environment in this language in which agreement occurs and also that the referent is unambiguously recoverable from the discourse context (Ferraz 1987). The contrast between first and second person singular pronouns is illustrated in the following two examples taken from Gilligan (1987: 164).

(28) ∅/a'mi i-ka ba dumi'ni
pro/I 1sg-AOR go sleep
'[I] will go to sleep.'

(29) *∅/bo ka ba dumi'ni
*pro/(you) AOR go sleep
'You will go to sleep.' (São Tomé Creole)

The final language which requires mention here is Singapore English. It is interesting because it is still developing into a full creole (Zhiming 2001). While Mandarin (pro-drop) and English (no pro-drop) are the two main contributing languages, there is also some influence from Tamil (pro-drop) and Malay (pro-drop). The potential combination of Mandarin structures allowing rampant pro-drop of salient topics (see below) with English words showing hardly any inflectional morphology could make for exactly the kind of language our models had problems with. However, it remains to be seen what types of pro-drop will be allowed when Singapore English stabilizes. Going by the examples (30) through (32) from Zhiming (2001: 302), pro-drop may be limited to clauses containing verbs like *said*, i.e. from a single semantic class with some tense marking on the verb. Note that when the unexpressed subject is not co-referential with the subject of *said* (identified by the subscript letter 'i'), it is interpreted as referring to a previously expressed, but still salient topic (identified by the subscript letter 'k'). While the sentences are syntactically ambiguous, they normally have only a single preferred interpretation in any given context. In this regard, they are very much like the 'diary drop' form of pro-drop which we have described in section 2.3.

(30) $[...]_k$ then [my mum]$_i$ said [$\varnothing_{i/k}$ must call her sister]
(31) [Mei Mei]$_i$ said [$\varnothing_{i/k}$ finished lunch already]

(32) $[...]_k$ [Sar Che and Sar Ee]$_i$ said [$\varnothing i_{/k}$ go to the airport to fetch you and Li Sa]

In summary, our survey of null subject phenomena in creoles has not revealed any obvious counter-examples to the predictions made by the network simulations. While there are definitely creoles which allow pro-drop with non-referential phrases like weather constructions or arbitrary actors, there are no attested creoles which allow referential pro-drop without agreement, even if these creoles derive from pro-drop Romance languages (cf. Lightfoot 1991; Roberts 1999; Lipski 2002; Nicolis 2007). Thus, rather than being problematic for our simulations, the data from creole languages support them. We should point out that most creoles employ pre-verbal T/A/M markers (Bickerton 1981; Muysken 1981; Jara M. 1996; McWhorter 1997; Roberts 1999; Holm 2000; Muysken and Law 2001) – see for example sentences (18), (22), and (26). As we have seen, such markers can help identify verbs, and could act as a compensatory factor for pro-drop. The frequent availability of T/A/M markers makes the absence of pro-drop in creoles even more striking. Contact situations in which new languages evolve naturally with little or no prescriptive influences from a mixture of (typologically distinct) other languages present precisely the kind of scenario in which we would expect learnability issues to play an important role. So, it should not come as a total surprise that the results from our simulations and the data from natural languages point in the same direction.

4.2.2 MANDARIN CHINESE

Mandarin Chinese belongs to a group of languages from South-East Asia which share numerous typological features. The group also includes Thai, Lao, Cambodian, Hmong, Vietnamese, Cantonese, as well as various dialects of these languages. Although they are not all historically related, their prolonged geographical proximity has led to a large number of shared linguistic features, including isolating morphology (i.e. with minimal noun and verb morphology (Cooke 1968; Bisang 1996), SVO word order, tone, nominal classifiers, and question sentences that are formed with an overt question particle. The degree to which all these features are characteristic of an areal grouping remains open to debate – see Bisang (1996) for discussion of several hypotheses – but whatever the final analysis of these recurrent features is going to be, the isolating morphology and SVO word order make them relevant for the present study. 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 (Campbell 1969; Hatton 1975; Grima 1986; Aroonmanakun 1999, 2000). We will limit our investigation to Mandarin Chinese here because it is by far the best documented of these languages.

With respect to its word order, Mandarin Chinese presents a complex picture, because it exhibits some linguistic features which are associated with (S)VO languages (e.g. SVO sentences, prepositions, and auxiliaries which precede the verb) but also others which are typical of (S)OV languages (e.g. some SOV sentence patterns, possessive phrases where the head noun appears last, and most notably relative clauses

which precede the head noun — see Greenberg 1963; Li and Thompson 1981; Hawkins 1983; Lehmann 1984; Dryer 1991, 1992; Hawkins 1993). However, there are several lines of evidence supporting SVO as the basic word order: it is the most frequent order found in corpora of both written and spoken Mandarin (Sun and Givón 1985); it is the default interpretation assigned to ambiguous sentences by both adults (Li, Bates et al. 1992) and young children (Miao and Zhu 1992); and it is also the word order first acquired by children learning Mandarin (Erbaugh 1982; Chang 1992).

In terms of morphology, there is no dispute that Mandarin is an isolating language, and therefore does not present much morphological complexity. It does not have case-marking on its nouns (with the possible exception of the *ba*–construction – see Chang (1991), Li, Bates et al. (1992), Sun (1996), and Bender (2000)), or agreement markers on its verbs (Li and Thompson 1981; Huang 1984). It also does not mark tense or modality using morphological affixes, but it does feature a small set of aspectual markers, negation particles, and clause-typing markers (e.g. question particles). For example, the temporal interpretation of a sentence such as (33) from Li & Thompson (1981: 197) is unspecified, if taken out of the context.

(33) ta yao si
3SG want die
'S/He wants/wanted to die.' (Mandarin)

Let us now consider unexpressed subjects in Mandarin.[12] It is well documented that Mandarin allows extensive pro-drop in subject position, not only with weather verbs and arbitrary/generic referents but also with referential third person subjects (Li and Thompson 1981; Huang 1984, 1989; Chui 1992; Tao 1996; Tardif 1996; Li 1997; Tardif, Shatz and Naigles 1997). The following two examples from Huang (1984: 537) show that pro-drop can occur in main clauses (sentence (34)) as well as in subordinate clauses (sentence (35)). The examples also show that pro-drop applies across different persons.

(34) ∅ lai-le
pro come-LE
'[I/You/She/He/We/You/They] came.'

(35) Zhangsani shuo [$\varnothing_{i/k}$ bu renshi Lisi].
Zhangsan say pro not know Lisi
'Zhangsan said that [he] did not know Lisi.' (Mandarin)

While we are unaware of any large scale studies of exactly how common pro-drop is in Mandarin Chinese, Chui (1992) reported 52% pro-drop in a corpus of

[12] Whether the linguistic concept of 'subject' has a role at all in Mandarin Chinese has been an issue of considerable debate. Some deny its existence in the language and argue instead that the pragmatic topic-comment structure describes all phenomena (e.g. LaPolla 1990; Tao 1996). Others, however, conclude that cross-linguistic tests which identify subjects also apply in Mandarin (e.g. Chen 1989; Tan 1991) and that young children treat subjects and topics differently (Chien and Lust 1983). We refer interested readers to Chao (1968), Li and Thompson (1976, 1981), Hu (1991) and Shi (2000) for further discussion.

1,127 adult spoken clauses, and Tardif, Shatz and Naigles (1997) found that about 60% of subjects were left unexpressed in their corpus of child-directed speech.

Given its SVO word order, lack of morphology, and frequent pro-drop, Mandarin appears to instantiate the 'difficult' language type which the neural network models had a hard time learning. We will argue here that this conclusion is not entirely correct for a number of different reasons. First, it must be pointed out that pro-drop in Mandarin is not an unconstrained phenomenon. In their hefty grammar of Mandarin, Li and Thompson (1981) actually devote an entire chapter to a discussion of when pro-drop is allowed: they describe both pragmatic and structural constraints, with the former being most relevant (see also Huang (1995), and Prince's (1998) discussion of pro-drop in Yiddish, which shows similar strong discourse constraints despite having agreement). In general, pro-drop is only allowed in Mandarin when the unexpressed element can be readily recovered from the discourse or the situational context. For example, in response to the question "How did you feel about that movie last night?", sentence (36) from Li and Thompson (1981: 658) is perfectly acceptable (and easily understandable).

(36) ∅ yidi n dou bu xihuan ∅
pro a-little all not like pro
'[I] didn't like [it] a bit.' (Mandarin)

Similarly, in a paragraph which describes a house, it would be sufficient to mention the house explicitly only at the very beginning. The following sentences could

then leave further references to this topic NP unexpressed. On the other hand, when the referent is not readily available or if there is potential ambiguity, a nominal element has to be used (compare Chen 1989). For example, in an unusual question such as (37) the presence of the pronoun *ni* 'you' is expected and does not normally carry contrastive stress (Li and Thompson 1981: 667).

(37)	Ni	xihuan	bu	xihuan	Beiduofen	de	yinyue?	
	You	like	not	like	Beethoven	GEN	music	
	'Do you like the music of Beethoven?'							(Mandarin)

The structural constraints on pro-drop in Mandarin rule out unexpressed arguments in co-verb and serial verb constructions. In both cases, the noun phrase immediately following the verb must be present. For example, in sentence (38) the co-verb *gen* 'with' requires the pronoun *ta* 's/he', just like the serial verb *mingling* 'to order' in sentence (39) requires that the noun phrase which is shared between the two clauses be expressed (Li and Thompson 1981: 675).[13]

(38)	Wo	gen	*(ta)	xue	Yingwen
	I	with	*(3sg)	study	English
	'I study English with him/her.'				

[13] As pointed out to us by Roger Levy (p.c.), *mingling* can also be followed by verbs. However, this usage is quite rare from a statistical point of view, so children could be forgiven for ignoring it, if it occurs in child-directed speech at all. More generally, though, it is important to realize that none of the cues mentioned in this section need be completely reliable; to be useful, they only have to be associated with either nouns or verbs in most cases.

(39) Ta mingling *(ta) yong daozi
3sg order *(3sg) use knife
'He/she orders him/her to use the knife.' (Mandarin)

The fact we want to draw attention to is that these two structural constraints on pro-drop prevent verbs (*xue* in sentence (38), *yong* in sentence (36)) from appearing in a position where the listener expects a noun phrase. Recall from our analysis of the mistakes made by the models that distinguishing noun from verbs presents the main challenge for networks learning a morphologically poor SVO language with pro-drop. More specifically, the networks occasionally confused subjects and verbs in sentence-initial position and then failed to recover from this confusion. We also saw that any kind of marking on just one category, either noun or verb, was sufficient to make such a language easily learnable in the simulations. While co-verbs and serial verbs are not typically considered as markers, they can still function as such in Mandarin. The point we want to make here is that in addition to their regular semantic meaning, co-verbs and serial verbs also signal to the parser that the next constituent must be a noun phrase. Because at least the co-verbs form a limited (closed class) set which can be learned quite early, they provide valuable information about the lexical category of the followin word, especially when this word is still unfamiliar to the language learner.

As long as there are other ways to tell nouns from verbs, it is simply not necessary for a language like Mandarin Chinese to have case marking or rich agreement to express 'who did what to whom'. Once we start looking for cues to

identify nouns or verbs in the language, it is not very hard to find several candidates. As mentioned earlier, Mandarin has a small set of aspect markers which suffix to the verb (i.e. -*le* for perfective aspect; -*zai* and -*zhe* for durative aspect; and -*guo* for experiential aspect). The following sentence exemplifies the use of –*le*, the most frequent of these aspect markers (Li and Thompson 1981: 190).

(40)	Tamen	fa	le	wu	shi	ge	qingtie
	they	issue	PFV	five	ten	CL	invitation
	'They sent out fifty invitations.'						(Mandarin)

For someone learning Mandarin, the presence of a suffix like -*le* is a very good indicator that the word it is attached to is a verb and not a noun. This piece of knowledge makes it easier to interpret other occurrences of the same word, even when –*le* is absent. In addition to the aspect markers, there are other classes of words that can serve a similar function because they normally only precede verbs and not nouns. These include auxiliaries, negation words, and adverbs. Obviously, the smaller the class of words, and the more frequent its words occur, the more likely it is that a language learner will be able to use them early on to determine the lexical category of other words.

We find a similar situation when it comes to identifying nouns in Mandarin. Next to the co-verbs we have already mentioned, the language also features prepositions which have to be followed by their dependent noun phrases, as well as limited derivational morphology specific to nouns. Perhaps more importantly,

Mandarin has a rich system of classifiers. These are words that must accompany demonstratives, numbers, and some quantifiers and that often express a measure or another property of the noun they precede. Mandarin has a large number of specific classifiers for reference to certain classes of referents, not unlike English *a flock of geese* or *a pride of lions*, but also a few more generic ones such as gè which can be used with various kinds of nouns. The following two sentences illustrate the use of classifiers – example (41) is from Li and Thompson (1981: 110) and example (42) from Li and Thompson (1981: 104). Once one has learned that e.g. *gè* is almost always followed by a noun (or larger noun phrase), one can use it as a reliable cue for determining the lexical category of the following word. We will investigate in section 5.2.2 below whether children learning Mandarin indeed make use of this strategy.

(41) Zuotian you yi chang dianying
yesterday exist one CL movie
'Yesterday there was a movie.'

(42) san ge ren
three CL person
'three people' (Mandarin)

Another processing indicator for finding a noun phrase is the *ba* particle which has been mentioned already. The majority of SOV sentences in Mandarin are in fact S *ba* O V, so *ba* signals both the marked word order as well as the presence of a following noun phrase. When *ba* is present, this noun phrase cannot be omitted, as in

sentence (43). Syntactically, *ba* has been described (among other things) as an object marker (Chang 1991; Chang 1992; Sun 1996) as well as a verb (Bender 2000), but from our perspective, it crucially functions for the language learner as a noun identifier, especially in a position in the sentence where a verb is normally expected. Example (43) is taken from Li and Thompson (1981: 466).

(43)	Wo	*(ba)	cha	bei	nong	po	le	
	I	BA	tea	cup	make	broken	PFV	
	'I broke the teacup.'							(Mandarin)

In summary, we have seen that Mandarin may well be an SVO language with pro-drop and without case marking or rich agreement, but this still does not instantiate the language type which the models failed to learn well. Unlike with the artificial language type, various cues are available to signal whether a word is a noun or a verb, and there is no reason to assume children learning the language would not be able to make use of them to learn about the lexical category of new words (Saffran 2002). However, we do not have frequency data to determine how often at least one of these cues is available to a language learner. Moreover, Mandarin Chinese presents other learning difficulties, such as its frequent noun/verb homonymy – a phenomenon we will investigate in Experiment 4. We will return to these issues below. At this point, we can conclude that although Mandarin gets far closer than most languages, it is not a pure instantiation of the 'difficult' type – the available cues for distinguishing nouns

and verbs are likely part of the solution to the learnability problems which it apparently presents.

4.3 Summary

In Experiment 1, we tested neural networks to see how well they could artificial SVO language types that explored all possible combinations of three linguistic parameters: pro-drop, case-marking, and verb markers. We found that the presence of pro-drop in the absence of morphological markers on the nouns and verbs presented a serious learnability challenge for the models. They experienced problems determining whether the first word in a sentence was the subject or the verb and were then unable to recover when more information became available.

A comparison between the results of the simulations and the language types attested among natural languages revealed that there are several artificial types with case marking but without rich agreement which are unattested but which appear to be possible human languages. However, the simulations also suggested that pro-drop in the absence of compensatory morphological marking leads to a language which is significantly harder to learn. Evidence for this prediction was found in creoles. They are among the most morphologically impoverished languages, but they do not have pro-drop even if some of their contributing languages do. Mandarin Chinese, with its very poor morphology and extensive pro-drop, is probably the most challenging natural language for the modeling results of Experiment 1. However, an examination of Mandarin structure showed that it has considerably more cues distinguishing nouns

and verb than the 'ideal' artificial type without marking that the networks were faced with. Consequently, Mandarin is not a serious counter-example to the results of our simulations.

Chapter 5. Experiment 2

We have so far used our models to explore cross-linguistic phenomena. It is also worthwhile to bring a developmental perspective to the simulations. Instead of looking for natural languages which are exemplars of the type the networks had problems learning, we now want to consider how the 'difficult' type can shed light on the observation that some sources of linguistic information appear easier to learn than others (e.g. Bates, McNew et al. 1982; Slobin and Bever 1982; Bates, MacWhinney et al. 1984). Recall from the first experiment that neural networks which had to learn a language with a consistent word order (i.e. without pro-drop) never experienced any problems with their task. Similarly, the networks which had access to reliable morphological information (whether on the verbs or the nouns) also learned to parse sentences of their language within 10 training cycles. When neither word order nor morphology was available to the networks, the same amount of exposure to the language was clearly insufficient for good generalization to novel sentences. This finding is especially intriguing because the sentences in the test corpora contained only known words, and all these words were nouns, pronouns or verbs. So, it should have been possible for the models to have learned whether the first word was a noun (and therefore the subject) or a verb. Instead, our analysis of the errors made by the networks showed that they were sometimes confused about the lexical category of the first word and gave partial activation to both the subject and verb output banks. This

suggests that learning a language becomes harder when the only cue for parsing is the lexical identity of the words themselves, rather than the structural position the words appear in or the morphological markers they carry.

This doesn't mean that it is impossible to learn the lexical category a word belongs to in the absence of extra cues. The mere sum of the contexts a word occurs in probably carries sufficient information if the available language corpus is large enough. For English, it has been shown that the lexical categories of nouns and verbs can be acquired pretty easily by statistical learning mechanisms that pay attention to the distributional context in which each word occurs (Redington, Chater and Finch 1998; Mintz, Newport and Bever 2002; Levy and Manning 2003; Monaghan, Chater and Christiansen 2005; Christiansen and Monaghan 2006). This result was valid for corpora of adult English and child-directed speech. Moreover, the categories of noun and verbs were learned noticeably better than any other lexical class. Similar results were reported by Li, Burgess and Lund (2000; Li 2006) in their study of large English and Mandarin corpora. Using a combination of word co-occurrence data as well as unsupervised neural networks (self organizing maps), they found that their model would cluster nouns and verbs in different areas of the map, with the distance between individual words determined by their similarity in usage. It seems reasonable to assume that children acquiring their languages will also make use of distributional information (Aslin, Saffran and Newport 1999; Saffran 2002; Christiansen and Monaghan 2006), but they have the additional advantage of living in a meaningful world, where words are labels that can be attached to observable entities and relations in a physical world.

Unsurprisingly, words with concrete referents are thus also learned earlier (Gentner 1981, 2006). Still, it is an open question whether the networks can learn the problematic language or not. Given the lack of co-occurrence patterns and access to meanings, is this language learnable at all or was there simply insufficient exposure in the simulation described in Experiment 1?

Given our focus on determining the role of structural patterns, it is non-trivial to either train the networks on large natural corpora or to provide them with rich semantic information. Instead, we increase the amount of exposure the models have to their languages by training them for more epochs. All other model parameters are kept the same.

5.1 Network results

We started by training all the language types for 20 additional cycles, as this proved to be sufficient for all types but one to reach near-perfect performance. The results are summarized in Table 7. It is easy to see that additional training is indeed helpful for the neural networks. Performance on all but the problematic type is essentially perfect, with more than 99.9% of the test sentences analyzed correctly. More importantly, the SVO language with pro-drop but without morphological marking sees its performance increase from 73.4% (after 10 epochs) to 91.4% (after 30). This type still performs much worse than any of the other language types (post-hoc Tukey comparisons show it to be significantly different from every other type; MS = .60389, df = 228.00, $p<0.0001$), but it also demonstrates that the required

Table 7. Results of Experiment 2. Average percentage of test sentences analyzed correctly by models learning each type of SVO language, as defined by three parameters: pro-drop, case-marking, and head-marking. Each number averages over 10 accusative networks and 10 ergative ones. Each network was trained for 30 epochs.

– PRO		N-marking	
		—	Case
V-marking	—	99.9%	99.9%
	T/A/M	99.9%	99.9%
	Agr	99.9%	99.9%

+ PRO		N-marking	
		—	Case
V-marking	—	91.4%	99.9%
	T/A/M	99.9%	99.9%
	Agr	99.9%	99.9%

information to solve the 'who did what to whom' task is available in the training corpus – the networks just have a harder time extracting it than when they can use word order or morphological marking.

Given the improved performance after 30 epochs, we decided to see how well the 'difficult' type could be learned if time was not an issue. (We did not continue training on the other types because they were already performing at ceiling.) The numbers in Table 8 show that the models continue to score better on the sentences in the test set as training continues, getting close to but never quite reaching the 99.9% scores achieved by the other networks after only 30 epochs.

With the performance of these networks approaching perfection after 200 epochs, it becomes easy to identify the remaining source of the errors. The word-by-word analysis of the mistakes still made by the networks learning the 'difficult' language type is shown in Figure 3. It is easy to see that sentences with

Table 8. Average percentage of test sentences analyzed correctly for an increasing number of training epochs by the models learning the 'difficult' language type: i.e. SVO, with pro-drop, but without case-marking or verb marking. Each percentage averages over the results from 20 networks.

THE 'DIFFICULT' TYPE	
Training epochs	Performance
10	73.4%
30	91.4%
50	95.6%
75	96.5%
100	97.0%
125	97.3%
150	97.5%
175	97.7%
200	97.8%

pro-drop (V, VO) no longer present a problem at any point during parsing. But sentences which have an expressed subject (SV, SVO) show lower performance. This result is largely due to the sentence length effect discussed earlier – i.e. the longer SVO

sentences lead to more mistakes, because more information has to be stored in the recurrent layers. Moreover, subjects appear in first position, have to be remembered longer and are thus vulnerable to being forgotten more easily.

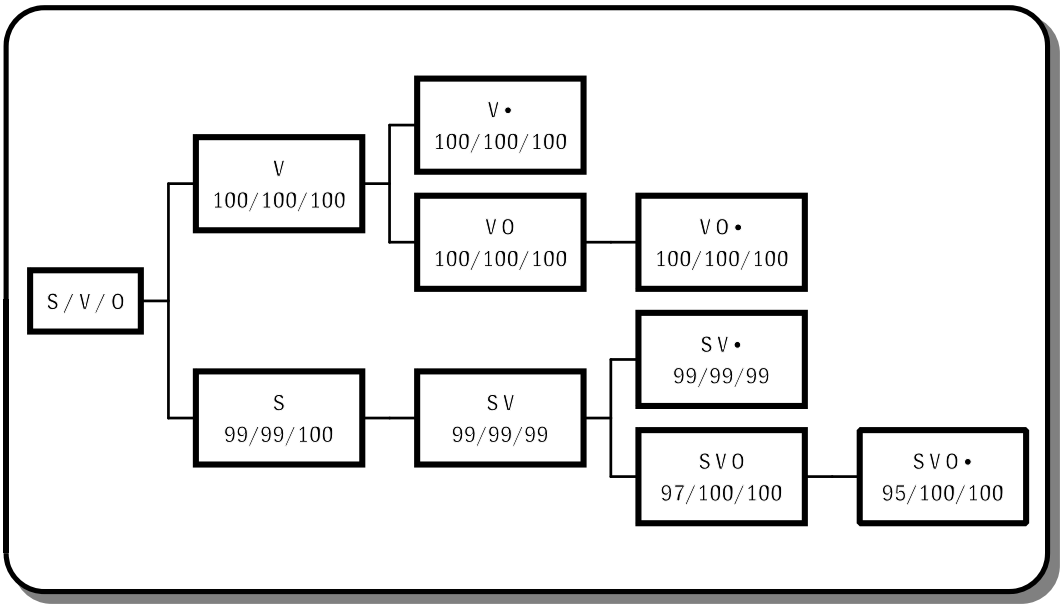

Figure 3. Word-by-word analysis of the performance on the four sentence types of the networks learning the difficult language type after 200 training cycles. The numbers in each box show the percentages of Subjects, Verbs, and Objects that were analyzed correctly after having seen the part of the sentence shown in the same box. E.g. after having seen just the Subject, 99% of the Subjects, 99% of the Verbs, and 100% of the Objects were correct at the output.

However, it would be a mistake to think of the subjects as a uniform category. For example, the 95% performance on SVO subjects is misleading because it averages over pronouns (100% correct) as well as nouns (90% correct). The lower performance on these nouns results from their having lower type and token frequencies than the other words which could appear in the same position. In an average training corpus of 3,000 sentences, about 750 nouns appeared as the subject of a sentence (half in transitives; half in intransitives); the other sentences contained either an overt pronoun

(1,125) or a verb (due to an unexpressed subject 1,125) in sentence-initial position. As a class, nouns were thus less likely to be found as the first word. In addition, individual nouns also had low token frequencies. The 750 subject nouns tokens were distributed over the 300 nouns in the lexicon, so each noun (and especially the inanimate ones, which were less likely to occur as subjects) appeared in subject position only a few times. In comparison, each of the 4 subject pronouns had a token frequency of more than 250, and each of the 100 verbs was seen more than 10 times in sentence-initial position. During training, the learning algorithm made the neural networks pay more attention to the more common types and tokens, so the subject nouns were learned relatively late.

This finding may come as a surprise, given the advantage nouns generally have in the acquisition of natural languages (see above and section 6.2.1). However, the noun bias in children is taken to derive from the conceptual primacy of early noun referents (i.e. entities) over those of verbs (i.e. relationships). In the simulations, the subject nouns received no such conceptual support to help overcome their lower frequencies. In addition, there are obviously some nouns in child-directed speech which have much higher token frequencies than the majority of the verbs, so these nouns would be even easier to learn for children. Implementing frequency differences within lexical categories for our simulations is one of the refinements we are planning to pursue. We are confident that we will be able to capture much of the desired phenomena because the networks did not have any major problems with the object nouns. The latter were learned well despite the fact that the nouns themselves were the

very same words which could also appear in subject position. Even after only 10 epochs, 94% of the nouns were already processed correctly in object position, whereas performance on these same words as subjects was much lower (39% in intransitive sentences; 26% in transitive ones). The explanation for this difference is threefold: first, objects occurred closer to the end of the sentence so they were easier to remember; second, objects always followed verbs which the networks had learned to be transitive, so there was helpful syntactic context which prevented any ambiguity about the role of the object noun; finally, following cross-linguistic tendencies, the relative type and token frequencies of object nouns were much higher when compared to the object pronouns, so the networks were also better prepared for parsing nouns in this position.

5.2 Linguistic discussion

The goal of the second experiment was to determine whether our networks could learn to solve the 'who did what to whom' task on the basis of lexical identity alone, i.e. in the absence of a fixed word order and/or morphological marking. We have seen that the answer to this question is essentially affirmative, although there are two caveats. First, the amount of training needed is about an order of magnitude higher. Second, overall performance on the sentences in the test corpora remained below performance on the other language types, even after the additional training. These findings thus corroborate our initial hypothesis that learning a language becomes harder when each and every word has to be learned separately. The issue we

take up in this section is whether the network results are compatible with what is known about the acquisition of natural languages by children. We first provide a general sketch of early learning of syntax and morphology, and then look in more detail at the acquisition of Mandarin, as it is one of the few human languages that is similar to the language type the models took so much longer to learn.

5.2.1 EARLY GRAMMAR ACQUISITION

As there is a wealth of data available nowadays about how children learn the linguistic structures in their languages, it is relevant to compare the findings of this literature here to the acquisition strategies exhibited by our models. This comparison shows that there are some differences between the two – often related to the implementation of the models – but the similarities in the kinds of language patterns that both children and networks pay most attention to are noticeably more striking.

We feel that the majority of empirical language development research in the last three decades has provided ample evidence that human language is first and foremost a learned ability, with children developing increasingly complex linguistic representations on the basis of the language data available to them as well as their own increasing cognitive skills.[14] One important line of evidence in favor of this position is

[14] Our theoretical linguistic position is thus quite compatible with usage based approaches such as Cognitive Grammar (Langacker 1987) and other work in cognitive linguistics (e.g. Barlow and Kemmer 2000; Goldberg 2002). This position is by no means uncontroversial. For recent criticism, see Newmeyer (2003, 2005).

that young language learners are remarkably conservative in how they initially use their languages. Let us illustrate this phenomenon with several examples.

The conservative nature of early language can be observed at many levels of analysis, but its effect is always the same – children limit their use of linguistic elements to the exact contexts in which they have observed them. So, newly acquired morphological markers, whether prefixes or suffixes, first only appear with a limited set of words (Slobin 1973; MacWhinney 1978; Tomasello 1992). It has also been found that the first produced morphemes are not analyzed by the language learner as separate and separable segments, as opposed to being taken to be root words themselves (e.g. Armon-Lotem and Berman 2003). Similarly, the argument roles which help define the meaning of relational words are initially quite limited, with e.g. the verb *to hit* implying the existence of a *hitter* and *hittee* more than the general roles of agent and patient (Tomasello 1992). Specific verbs are often only used in specific constructions (e.g. the verb *to give* for the ditransitive – see Goldberg, Casenhiser and Sethuraman 2004; Theakston, Lieven et al. 2004). Entire constructions may in turn only be used in a limited manner (e.g. the 'I think' in *I think Daddy's sleeping* does not immediately appear next to 'You think' or 'He thinks' – Diessel and Tomasello 2001), and a construction like the passive may at first appear only with specific verbs (Maratsos 1998). Also, most words are at first only used in specific discourse contexts, which only gradually expand to fill the full adult range – this even applies to more abstract terms like expressions for causal and temporal relations (Levy and Nelson 1994). Finally, young language learners tend to produce a large proportion of

utterances which are either completely invariant (i.e. rote), or have a single open slot in a fixed position – the combination of these two classes consistently combined for at least 88% of the multi-word utterances by eleven children between the ages of 1;0 and 3;0 in a recent corpus study by Lieven, Pine and Baldwin (1997). What ties all these phenomena together is that, over time and often one element at a time, children expand the use of their morphemes, words and constructions beyond the initial boundaries. The mechanisms underlying the expanded usage are not fully understood, but it has been argued that the child needs to acquire a 'critical mass' of the relevant elements before he or she can develop the relevant generalization to support more general production (Marchman and Bates 1994; Tomasello 2000; Wilson 2003; Marchman, Martínez-Sussman and Dale 2004).

The fact that children start as conservative learners appears to be at odds with the results we observed in the simulations: if grammatical knowledge is initially restricted to individual words rather than morphemes, then we would expect the lexical identities of the words (as well as their categories) to be learned before more generic parsing strategies based on morphological markers or word order. The models, on the other hand, have no problems acquiring these more abstract mechanisms for determining 'who did what to whom'. The reason the children and the models behave differently in this regard lies in a crucial property of the input data: i.e. whereas the children hear words with wildly different frequencies, the languages the models were exposed to each word of a given class (e.g. transitive verbs, animate nouns) with equal probability. For the children, learning to comprehend and produce the high frequency

content words can make an immediate impact on their communication skills, so it is not surprising that these are also acquired first. Because none of the words in the artificial languages stand out, our networks essentially skip this first step and start by looking for patterns which are frequent in their input, i.e. morphological markers in the languages which have them. As we have seen, they also quickly notice the value of word order as a processing cue for the languages without pro-drop.[15]

While our networks may fail to model the very first stages of language acquisition, their behavior is entirely compatible with what has been observed in children who have moved beyond the fixed forms only phase and whose language has started to exhibit more 'grammar' – i.e. structural patterns which generalize over the shared properties of several learned words. Interestingly, some of the very first parts of grammar to be acquired productively are the basic strategies for expressing 'who did what to whom' in the target language, both when these strategies involve word order and morphology (Slobin 1973; Clark 2001). However, when both position and markers play a role, the elements which are actually observable and have a phonological form associated with them, i.e. the morphological markers, are generally learned first (Akhtar

---

[15] It would obviously be desirable to make the artificial languages more realistic by introducing variable frequencies in their vocabularies, and we plan to do so in follow-up simulations. There is ample data available from other connectionist language research showing that neural networks are sensitive to frequency patterns in their input languages – e.g. just about all the papers investigating the acquisition of the English past tense point in this direction (e.g. Rumelhart and McClelland 1986; Plunkett and Marchman 1991, 1996; Plaut, McClelland et al. 1996) – so we see no reason to believe similar phenomena would not be observable in our simulations. Also note that the exact relationship between frequency and acquisition is still a topic of some debate – see e.g. Carroll and White 1973; Ellis and Morrison 1998; Anderson and Cottrell 2001; Weekes, Chan et al. 2004.

and Tomasello 1997). With respect to these markers, it has often been observed that they are learned better (easier and faster) when they have invariant phonological forms, appear at the end of words, express salient meanings, and occur with above average frequency in the input (e.g. Lieven 1997; Peters 1997; Gil 2006). Morphemes which meet these criteria can turn into 'inflectional imperialists' – i.e. they are overgeneralized in contexts where the adult language requires other, less consistently present or phonologically more variable morphemes (Slobin 1973, 1985). The same general cognitive mechanisms which underlie the child's ability to notice the order in which morphemes appear in words are presumably also involved in the attention which is paid to order in which words appear in sentences. The resulting generalizations lead to a significant preference for a specific fixed word order, even when the adult language allows much more variation – in fact, when utterances exhibiting these less frequent word orders are heard, they are often interpreted as if they were following the standard order, even when the result is nonsensical (Slobin 1973; but see Tomasello (2000) for different results with novel verbs). In summary, young language learners are without doubt capable of extracting consistent morphological markers and word order patterns which are used to express 'who did what to whom'. In this regard, they are quite similar to how our networks learn to solve their task.

Evidence for how the presence of consistent cues plays an important role in the acquisition of an appropriate parsing strategy for a given language can be found in an experimental cross-linguistic study by Slobin and Bever (1982) – see also Bates, MacWhinney et al. 1984; Thal and Flores 2001. Their experimental subjects were

children learning four different languages in four age groups between 2;0 and 4;4. The languages involved were English, Italian, Turkish, and Serbo-Croatian – all spoken in the same region of the world but featuring quite different strategies for marking subjects and objects. Slobin and Bever compared how well children from each of the age groups had acquired the regular parsing strategies of their language to determine who did what to whom by asking them to act out (using toy animals) sentences containing two animate nouns and a transitive verb which was compatible with both nouns – e.g. *the squirrel scratches the dog*. For the children learning English or Italian, the experimental parameters were word order (Verb Noun Noun; NVN; NNV) and prosody (stress on the first noun, or the second noun). The children learning Serbo-Croatian or Turkish were presented with sentences which also had varying word orders, but the second parameter this time was the possible presence of case markers. The main results of the experiment are shown in Table 9 (Slobin and Bever 1982: 241). The numbers in Table 9 indicate the percentages of sentences which were grammatical in each language and which were also processed correctly by the children – i.e. the children used the correct toys to be the agent and patient in the sentence they heard.

Probably the two most striking observations to draw from Table 9 concern the acquisition of Turkish and Serbo-Croatian. The former is being learned much earlier than the other three languages, whereas the latter appears to be much more problematic for the children. As pointed out by (Slobin and Bever 1982), the explanation for the different acquisition profiles lies in how each of the languages codes subjects and objects. Turkish is an SOV language but it also allows a

Table 9. Percentages of reversible transitive sentences processed correctly by children learning four different languages. The numbers are taken from Slobin & Bever (1982: 241).

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

considerable amount of word order flexibility in main clauses (Ozkaragoz 1987; Kural 1997). However, it is the only one of the four languages tested to exhibit consistent and unambiguous case marking on each of its nouns (at least when they are definite as in the sentences used in the experiment). The presence of reliable cues in Turkish is in stark contrast with Serbo-Croatian, because parsing strategies for the latter require paying attention to both some word order tendencies as well as the elaborate (and sometimes phonologically ambiguous) system of case markers. The relatively simple and consistent encoding scheme in Turkish is thus learned early by children, with performance in the first Turkish age group equaling that of the oldest age group of Serbo-Croatian children. The English children, unsurprisingly, have to learn to use word order, and this proves to be a feasible task, albeit initially harder than interpreting the more salient case markers. Finally, the children learning Italian also have to learn to

combine two different categories of cues, i.e. agreement markers on the verbs and prosodic stress. But the agreement markers are quite reliable so the stress information is only needed occasionally to disambiguate sentences. The availability of a usable default strategy proves sufficient in Italian.

The parallels between the types of grammatical cues which are useful for both young language learners and our networks are interesting, and we feel that they further support the notion that neural network research can be relevant for the study of human language acquisition. Slobin and Bever (1982) also argue that all children initially become sensitive to what they call the 'canonical sentence schemas' of their respective languages, i.e. the mechanisms which the languages use to encode who did what to whom in simple, active, declarative sentences. These sentence schemas then become the basis for how children learn to comprehend (and later produce) more complex utterances. For example, a longitudinal study of the production data of seven children learning English showed that (mental) perception verbs such as *know, think* and *see* initially appear only in formulaic clauses such as *I know X* before these verbs are used in other constructions (Diessel and Tomasello 2001). As a consequence, the fact that our simulations are limited to using very simple sentences in the artificial language corpora may not be that far removed from the type of sentence children initially pay attention to. As demonstrated in Elman's (1993) well known 'starting small' simulations, neural networks can also benefit from learning environments in which they are initially limited to just being able to process the simple (core) sentences of a language. Once the ability to parse such sentences has become entrenched in the

connection weights, they are then more capable of processing more complex sentences than if they had tried to learn sentences of all levels of complexity from the very beginning. Along similar lines, our focus on learning morphological markers and word order cues is supported by cross-linguistic acquisition data showing that young learners generally acquire these cues first for transitive sentences in which an animate agent causes a significant change of state in a patient (Slobin 1973, 1981; Tsunoda 1981; Herr-Israel and McCune 2006). The ability to understand such sentences obviously entails the capability to distinguish verbs from nouns, as well as the ability to assign which noun phrase is the subject and which is the object – i.e. the very same task which our networks have to learn as well.

In this brief summary of early grammar acquisition in children, we have outlined how young language learners start using morphological markers and word order. We have seen that children become sensitive to these processing cues only after they have already learned a number of words as separate, unanalyzed elements. In our simulations, on the other hand, the absence of salient, high frequency words in the training corpora causes this first stage in child acquisition to be skipped, and the networks immediately start learning the morphological markers and word order cues. For both children, and networks, however, there is an observable preference for features that occur frequently and consistently, and that are easily recognizable.

5.2.2 Acquiring Mandarin Chinese

In our earlier discussion of Mandarin Chinese (see section 4.2.2), we looked at the properties of the adult language and concluded that it is almost an instantiation of the language type our networks have problems with – while Mandarin has rampant pro-drop, it also has a small number of morphological markers and word order cues which can be used to identify the nouns and verbs in the language. The simulations show that knowing the correct lexical category each word belongs to is crucial for solving the 'who did what to whom' task in an SVO language with pro-drop. These findings suggest that a closer look at the acquisition of Mandarin is in order, because it allows us both to verify whether children pay attention to morphology and word order when neither is a very reliable cue in the adult language, and to gauge how well the noun-verb distinction can be made by young language learners. One can imagine a scenario in which children learning Mandarin pay little attention to the morphology of the language, because so little of it is present in the data they are exposed to. However, the acquisition data for Mandarin we review below reveal almost the opposite picture: children learning the language make immediate use of the formal cues which are available to them, and the categorical distinction between nouns and verbs is acquired very early.[16]

---

[16] We will not discuss the pro-drop parameter here. It has been argued that children learning Mandarin display relatively limited amounts of pro-drop in production (Erbaugh 1982; Miao and Zhu 1992), but it is hard to determine in the early stages of language acquisition whether unexpressed elements are due to knowledge of the grammar – i.e. topics can remain unexpressed – or to performance constraints on the number of arguments which can be made overt (see Freudenthal, Pine and Gobet 2002; Aronoff 2003). It does not seem

Let us first consider the word order data. As mentioned earlier, adult Mandarin features a mix of SVO and SOV sentences, with the former being more common (Sun and Givón 1985). This trend is even stronger in child-directed Mandarin: in the sample analyzed by Erbaugh (1982), about 10% of adult-to-child utterances were not SVO, whereas the number in adult-to-adult speech was 20%.[17] It does not come as a surprise then that children learning Mandarin who have reached the two-word stage produce are much more likely to produce SV or VO sentences, rather than OV (Erbaugh 1982, 1992; Chang 1992). The Noun-Noun groups SO or OS are essentially unattested (see below). When the children have reached two years of age, full SVO sentences make up the majority of sentence production, and the numbers of SV and VO sentences start to drop (Miao and Zhu 1992). It is only when children are three years old that they start to reliably produce grammatical SOV sentences, and it takes them several more years before they have mastered the full range of topic-fronted constructions. The same

---

unreasonable to us that both factors may play a role. As the children grow older, both their exposure to the language and their cognitive capabilities increase. Consequently, they have ample opportunities to hone their ability to keep track of multiple discourse referents, and consequently determine which one should be omitted (in production), or interpret the likely referent of an omitted subject (in comprehension).

[17] It is worth mentioning that child-directed speech in Mandarin shares many features with child-directed speech in other languages. For example, the sentences are kept shorter, pronounced more slowly, and often feature exaggerated prosodical contours (Grieser and Kuhl 1988). One possible adaptation to the properties of the languages proposed by Erbaugh (1982) is that adults "drill" children on specific question-answer structures. Because question words in Mandarin appear in situ (i.e. they are not fronted to the beginning of the sentence), these drills may reinforce syntactic cues about the linguistic contexts in which the words that are being questioned occur (compare Kaschak and Saffran 2006). While interesting, this proposal requires further study to determine whether it is a phenomenon that is truly special about Mandarin.

initial preference for simple SVO sentences has also been observed in experiments which investigated comprehension. Miao and Zhu (1992) report that young learners of Mandarin (63% at 3;6 and 75% at 4;6) are much more likely than adults (44%) to impose an SVO analysis on a Noun-Verb-Noun sentence when the first noun refers to an inanimate entity and the second one to an animate entity. When both nouns had animate referents, the first one was taken to be the subject 88% of the time by children aged 3;6, and 90% for children 4 to 5 years of age. These numbers are remarkably similar to the ones reported for children learning English (see Table 9 above), as is the preference for a syntactic SVO parse in the face of contradictory semantic data. The shift towards a semantic parsing strategy happens gradually. Even seven-year-olds still use a syntax-based interpretation in some constructions (Chien and Lust 1983; Miao and Zhu 1992). In summary, children learning Mandarin initially construct linguistic representations which are much more dependent on a rigid SVO word order than a study of the adult language would lead one to expect; they first adopt this simple sentence schema for both production and comprehension and only expand on it as their general cognitive skills develop.

The initial reliance on strict word order in the acquisition of Mandarin is often linked to the lack of morphology in the language (Erbaugh 1982; Chang 1992; Miao and Zhu 1992). While every sentence containing at least two words is bound to carry word order information, there is no such guarantee in Mandarin that any morphological markers will be present. Nonetheless, children learning Mandarin Chinese appear to pay as much attention to morphology as children learning other

languages. Comprehension of markers can be hard to determine in young children, but the production data are much easier to analyze. Probably the two most frequently uttered grammatical morphemes by children learning Mandarin are *–le* and *gè*.

The *–le* suffix has two functions in the adult language: first, when attached to the verb it marks perfective aspect; second, in sentence-final position, it indicates that an utterance is highly relevant for the current situation (Li and Thompson 1981; Li and Bowerman 1998). A single sentence can display a *–le* suffix attached to the verb as well as one in sentence-final position (e.g. S V-*le* O -*le*), but only a single one is used when this verb is also the last lexical word in the sentence (S V-*le*; not S V-*le* –*le*). In acquisition, -*le* is already used productively by two years of age, usually following a verb in sentence-final position and often with both perfective and current relevance meanings, as in examples (44) and (45) from Erbaugh (1992: 423; compare Huang 2006).

(44) dǎ-pó-le
hit-break-*le*
'[I] have broken [it].'

(45) kū-le
cry-*le*
'[I] have cried.'

In the spontaneous speech corpus analyzed by Erbaugh, the children averaged one *–le* every two minutes, and produced it more than twenty times as frequently as

the next most common aspect marker, i.e. progressive marker *zài*. Another indicator of how salient this marker is during early acquisition is the common error in child Mandarin of following a sentence-final verb with *–le le*, effectively double-marking the perfective and current relevance meanings.

Whereas *–le* is a suffix that attaches almost exclusively to verbs, the classifier *gè* is a reliable indicator that a noun will follow shortly. Classifiers in Mandarin Chinese appear obligatorily between, on one side, numerals, demonstratives and certain quantifiers, and, on the other side, nouns. Mandarin has more than a hundred different classifiers, with each of them being used to refer to nouns with specific properties (Li and Thompson 1981): e.g. the classifier *tíao* indicates that the following noun is an extended, long thing, while *kē* is used for small, round, hard referents. Prescriptive grammars of Mandarin state that the appropriate classifier should be used in each context, but analysis of actual usage in adult speech shows that a small set accounts for the large majority of the cases, with the very general classifier *gè* 'one' being most frequent (Erbaugh 1986, 2006; Hu 1993; Tai 1994; Chien, Lust and Chiang 2003). It is also *gè* that is first used as a classifier, both with animate and inanimate nouns. Spontaneous usage starts around 2;0 and it quickly ranks as the fourth most frequent word in children's speech (Chang 1992). Unsurprisingly, the more specific and less common classifiers are only learned slowly and it takes many years before they are all acquired correctly.

We have focused here on *–le* and *gè* for two reasons. First, they are both acquired early and used frequently. As such, they demonstrate that children learning Mandarin are quite like children learning other languages in that they quickly adopt grammatical elements that occur with considerable frequency in their input, that exhibit invariant phonological forms, and that have an immediate functional use for them. As Erbaugh (1992) puts it, rather than ignoring what little morphology can be observed, the children she studied just "reveled in the surface morphology that Mandarin does have" (442). In the absence of innate knowledge about the properties of every human language in existence, young language learners must extract from their input whatever bits of grammar are accessible to them at each developmental stage. In the case of Mandarin, this means that children have no way of knowing how relatively impoverished the morphology of the language really is. It remains to be studied whether the paucity of markers may actually make it easier for them to acquire that morphological markers that are available.

Our second reason for describing *–le* and *gè* is that acquiring these two grammatical elements has an additional benefit for the language learner beyond being able to express perfective aspect or obligatory classifiers. Once a child learning Mandarin has figured out that the suffix *–le* is reliably added to verbs, while the marker *gè* is a good indicator that the next word is a noun, he or she has also picked up a very useful strategy for learning the lexical categories of other words (compare Höhle, Weissenborn et al. 2004). As we have seen, knowing which words are nouns and which are verbs can be sufficient to determine 'who did what to whom' in SVO

languages with pro-drop. Crucially, there is reason to believe that children learning Mandarin Chinese are indeed very aware of whether a word they know is a noun or a verb. The evidence for this can be found in the phenomenon of zero-derivation, i.e. using a word belonging to one lexical category as if it belonged to another lexical category, but without changing its morphological form. Obviously, zero-derivation is very rare in languages which have rich marking systems, whether on nouns or on verbs. On the other hand, it is a very common phenomenon in English, a morphologically poor language – e.g. *bottle* is usually a noun, but becomes a verb in *The first brewery to bottle beer made a fortune*. Moreover, it has been well documented that many children learning English produce novel zero-derivations starting around 2;0 (Clark 1982, 1993, 2001): e.g. *to key* meaning 'to insert a key' or *to water* meaning 'to paddle in water' (Clark 2001: 386). Given the lack of morphological markers, one would expect to find similar data in child Mandarin. But, as Erbaugh (1982, 1992) has observed, one simply doesn't. Children learning Mandarin do not 'experiment' with lexical category usage, even in play settings.[18] Another line of evidence supporting this finding comes from the absence of Noun-Noun utterances in the two word stage – the equivalent of English *mommy sock* is not found, because children always combine a verb with a noun. Exactly how the children learn to determine the lexical categories of words remains an open question, but we feel it is safe to posit that distributional information (Li, Burgess

[18] With respect to this curious lack of zero-derivation, Erbaugh (1982: 508) cites the following couplet penned by James Matisoff, a linguist who specializes in South-East Asian languages: "In child Chinese there's nothing worser // Than using nouns as verbs, or vice versa."

and Lund 2000; Christiansen and Monaghan 2006; Shi 2006), prosodic patterns (Grieser and Kuhl 1988), as well as morphological cues like *–le* and *gè* play an important role (Höhle, Weissenborn et al. 2004; Tardif 2006).

In summary, there is considerable evidence from acquisition studies that children learning Mandarin Chinese initially construct a simpler language with rigid SVO word order and some salient morphology. Their internal representations of Mandarin are thus arguably less complex than what the adult language allows. The more complex constructions and word order variations of Mandarin are only used productively when the children are older and have had much more experience with the language. In addition, we find it very telling that children learning Mandarin don't produce the kinds of noun-to-verb zero-derivations which are so common in other languages with limited morphology. The fact that they treat nouns as a distinct word class from verbs is a strong indication that their developing language system is sensitive to the importance of lexical category knowledge when neither morphology nor sequential word order provide reliable processing cues.

5.3 Summary

In Experiment 2, we investigated whether additional exposure to the difficult language type, i.e. the SVO language with pro-drop but without morphological marking, led to better performance by our networks. We found that this is indeed the case, with average performance increasing from 73.4% after 10 epochs, to 91.4% after

30 epochs, and 97.0% after 100 training cycles. However, the value of this increased performance is called into doubt by the fact that all the other language types had reached ceiling (99.9%) by 30 epochs. What the models showed, then, is that difficult type can be learned, but at the same time also that it is very hard to learn compared to other language types.

From a cross-linguistic perspective, this finding helps explain the absence of human languages that fully correspond to the difficult type. When neither word order nor morphological markers are available as consistent and reliable cues, a language learner instead has to fall back on learning the lexical category of each individual word by keeping track of the sentence contexts in which it appears. This strategy works quite well for frequently occurring words (such as the ones children learn first), but it is easy to see that it becomes cumbersome for words which are not observed that often. For those – the large majority of the vocabulary of a language – it is much more convenient if morphological markers and/or word order information can be used as a shortcut to determine the lexical category of the word. We have also seen that young language learners are very capable of extracting such information from their input. Even when marking is scarce, as it is in Mandarin Chinese, children pick up on the cues which are available to them. The basic strategies for learning a language remain constant, resulting in acquisition profiles which can be quite similar for typologically distinct languages – e.g. the reliance on word order in both English and Mandarin. On the other hand, the fact that children learning Mandarin are much more aware than their counterparts learning English of whether a word is a noun or a verb

demonstrates that the acquisition process adapts itself to the language to be learned. While the absence of pro-drop in English allows nouns to be used as verbs creatively, children learning Mandarin can not rely on word order as a processing cue. They use their nouns as nouns and their verbs as verbs because sentences otherwise become incomprehensible.

Chapter 6. Experiment 3

The main goal of Experiment 3 is to determine how well our networks can generalize their parsing strategies to sentences containing novel words. Being able to demonstrate successful generalization is desirable because it is a requirement for modeling human cognitive processes. Especially with children, there is ample evidence that they apply linguistic knowledge gleaned from one set of language data to other forms. In production, for example, many two-year-olds learning English have no problem generating the plural form *wugs* for a nonce noun such as *wug* that they have just learned in an experimental setting (Berko-Gleason 1958; Tomasello, Akhtar et al. 1997). Similarly, the well documented phenomenon of overregularization in the acquisition of the English past tense also indicates an ability to generalize linguistic patterns from one set of words to other ones (Marchman, Plunkett and Goodman 1995; Marchman 1997). In comprehension, the need to generalize is even more striking, because every sentence that contains a novel word or a new way of combining familiar words depends crucially on our (usually effortless) ability to generalize from previously seen words and constructions to novel ones. Young language learners in particular are faced with a multitude of novel words that they have to categorize and interpret correctly. To make matters worse, children tackle this task with relatively limited cognitive abilities and linguistic knowledge.

To test the generalization capabilities of the model, we investigated how well our trained networks from Experiment 2 could solve the 'who did what to whom' task for sentences containing novel words. Not unlike children who hear new words, the networks had to use their knowledge of previously seen markers, words, and sentence types to make sense of the novel forms. We were especially interested in the ability of the networks to determine the lexical category of the new words, because the previous experiments have shown that being able to tell nouns from verbs plays an especially important role in SVO languages with pro-drop. Recall that even the simplest intransitive sentence in these languages can begin with either a noun (SV) or a verb (V, following pro-drop), so immediate disambiguation is required to avoid confusion. The results discussed below demonstrate that the generalization task is very feasible for all of the language types, except for the 'difficult' one. We will use this finding to argue that basic sentence parsing strategies may well be learnable with both less syntactic and less semantic processing than is often assumed.

There is another question that we want to address in Experiment 3 as well, namely does it matter whether these novel words include nouns, verbs, or both? As we will discuss below, our results relate to the nouns-first versus verbs-first debate, which has been a topic of much discussion in language acquisition (e.g. Gentner 1982; Choi and Gopnik 1995). It has been argued that verbs play a significantly larger role in pro-drop languages because they are never left unexpressed and can therefore be the only word in a simple sentence (e.g. Tardif 1996; Tardif, Shatz and Naigles 1997). We will show below that the networks provide support for this claim.

6.1 Network results

The implementation of the generalization experiment was as follows: first, we created 400 new words (300 nouns, 100 verbs) for the artificial languages by generating strings of 16 bits (with 6 bits set to '1') that were different from any existing word in at least two bit positions (i.e. minimum hamming distance of 2). We then created three new lexicon files by replacing either just the nouns, just the verbs, or both the nouns and verbs from the existing lexicon file. Note that the pronouns and morphological markers (if present in the language) were not replaced with novel counterparts under any condition, as the goal of the test was exactly to determine the usefulness of these more grammatical elements. Next, we combined the original grammars for each language type with each of the three new lexicon files to generate three new test corpora with 3,000 sentences. The final step was to take each network from Experiment 2 that had been trained for 30 epochs and have it process these three new test corpora. Obviously, a network which had originally been trained on sentences from a specific language type was also tested on new corpora from the same type. The performance measure remained the same: i.e. the percentage of the 3,000 test sentences in each corpus for which all three output slots (Subject, Verb, Object) were correct at the very end of the sentence. For these simulations, however, the actual pattern in each of the output slots was compared to all the words from both the training and the test lexica to ensure it was closest in Euclidean distance to the correct one.

The results for the test corpora containing novel words are summarized in Tables 10 through 12. Table 10 shows the results for novel nouns, Table 11 for novel verbs, and Table 12 for novel nouns and verbs. Each percentage is again an average over the twenty different networks which learned each language type. A cursory reading of the three tables reveals that the results again separate themselves quite naturally into two sets: on one side, there is poor performance on the difficult type (i.e. the type which has pro-drop but no marking on either nouns or verbs); on the other side, there is successful generalization for all the other language types. Rather than discuss similar results for each of the three tables separately, we first go over the interesting findings for all the language types but the difficult one. We then turn to the latter.

6.1.1 THE 'EASY' LANGUAGE TYPES

Ignoring the difficult type, the first observation to be made about the results in Tables 10, 11 and 12 is that the networks have very few problems with the generalization task. The presence of novel words in the test corpora barely affects the networks: when we compare the numbers with those for corpora containing all familiar words (see Table 7 above), the largest difference which can be found is less than 3%. Put differently, all the language types in all these conditions get at least 97% of the test sentences correct, even when they contain both novel nouns as well as novel verbs (Table 12). These results demonstrate that our connectionist models can

Table 10. Results from Experiment 3: novel nouns (30 epochs).

– PRO		N-marking	
		—	Case
V-marking	—	99.9%	99.9%
	T/A/M	99.9%	99.9%
	Agr	99.9%	99.9%

+ PRO		N-marking	
		—	Case
V-marking	—	80.2%	98.6%
	T/A/M	99.8%	99.8%
	Agr	99.7%	99.8%

Table 11. Results from Experiment 3: novel verbs (30 epochs).

– PRO		N-marking	
		—	Case
V-marking	—	98.9%	99.3%
	T/A/M	99.4%	99.5%
	Agr	99.2%	99.1%

+ PRO		N-marking	
		—	Case
V-marking	—	52.8%	97.5%
	T/A/M	99.0%	99.2%
	Agr	98.5%	98.6%

Table 12. Results from Experiment 3: novel nouns + verbs (30 epochs).

– PRO		N-marking	
		—	Case
V-marking	—	98.7%	99.0%
	T/A/M	99.1%	99.3%
	Agr	99.0%	99.0%

+ PRO		N-marking	
		—	Case
V-marking	—	44.2%	97.3%
	T/A/M	98.8%	99.0%
	Agr	98.4%	98.4%

generalize successfully to sentences containing completely novel words – we will return to the implications of this finding in section 8.3.

The results of Experiment 3 also let us assess the value of different sources of linguistic information. We can only compare the utility of word order and morphological marking, as none of these types depends on lexical category knowledge alone. It is easy to see that both word order and markers can provide the networks with sufficient cues for the generalization task. In Table 12, the networks for which word order is the only source of information still get 98.7% of the test sentences correct. The numbers are similar for the networks whose only cue was case-marking (97.3%) or agreement markers (98.4%). Across the three tables, word order always outperforms morphology, and verb marking is always more useful than case marking. However, given the small absolute differences as well as the absence of cross-linguistic data supporting this ranking, the relatively poor performance of case marking is likely the result of implementation details rather than any cognitive factors. The appropriate conclusion is that both word order and morphological markers can carry enough information for solving our 'who did what to whom' task. Another recurring result is that there is limited benefit from having access to multiple sources of information at the same time: neither combining word order with morphological marking, nor noun markers with verb markers leads to consistently better results than a single type of information.

The other question we are interested in concerns the effects of novel nouns versus novel verbs – i.e. Table 10 versus Table 11. However, all the pairwise comparisons of the cells in these two tables show only small differences, with the biggest one being just 2.1% for the language types with case-marking and pro-drop, but no verb marking (i.e. 99.6% in Table 10 vs 97.5% in Table 11). Here as well, we see no reason to believe these small differences are motivated by cognitive factors as opposed to details of the implementation. Moreover, the fact that the combination of novel nouns and verbs (Table 12) hardly affects the models' performance is another indication that lexical identity (and lexical category) knowledge plays a very small role in how the models parse these languages. Instead, the networks learning these language types relied on the grammatical cues which were present in the test corpora. It did not matter whether these cues consisted of word order patterns or morphological markers. The networks effortlessly applied the same parsing strategies to the novel sentences.

6.1.2 THE 'DIFFICULT' LANGUAGE TYPE

The only language type which fared poorly on the generalization task is the one which featured pro-drop but no case-marking or verb-marking – i.e. the same type which also benefited from additional training cycles in Experiment 2 to achieve decent performance on test corpora with familiar words. As the performance of these networks varied considerably between the experimental conditions, we will discuss them separately.

Let us begin with the results for novel nouns (Table 10). When the trained networks were shown sentences containing novel nouns, performance dropped to 80.2% (vs 91.4% on familiar words in Experiment 2). Post-hoc Tukey comparisons show this result to be significantly different from all of the other language types; MS = .21752, df = 228.00, $p<0.0001$. The errors were almost exclusively due to problems with the novel nouns: only 69.5% of the nouns were processed correctly as opposed to 99.5% of the (familiar) pronouns. However, a closer look reveals that it is only the nouns in subject position that caused the overall decrease in performance: 36.0% of intransitive subject nouns and 25.4% of transitive subject nouns were analyzed correctly, whereas performance on the object nouns was 95.1%. The different treatment of subject and object nouns implies that the networks still recognized the verbs from the training corpora. When they saw a transitive verb, they knew that the following word had to be the object of the sentence, and they were able to process this word accordingly, even if it was completely novel. The novel subject nouns, on the other hand, occurred in sentence-initial position, where there was no sentential context available to help interpret them. Due to the lack of morphological markers, these novel forms could be either nouns (S in SV or SVO) or verbs (V in pro-drop V or VO). As we have seen before, this kind of structural ambiguity in sentence-initial position leads to considerable confusion in the networks and partial copying of the novel word to both the Subject and Verb output banks. Even though the next word in the sentence (i.e. the familiar verb) could have been used to disambiguate the first word, the poor quality of the representation of the first word deteriorated even more as the other

words of the sentence were processed. The result was a pattern that usually failed to meet our strict performance measure of being closer to the correct word than any other one.

While novel nouns already caused definite problems for the difficult type, the presence of novel verbs had an almost catastrophic effect (Table 11): only 52.8% of the sentences in these test corpora were parsed correctly. Post-hoc Tukey comparisons again show this number to be significantly different from all the other language types; MS = 1.0930, df = 228.00, $p<0.0001$. We expected these errors to be due almost exclusively to the novel verbs, but a step-by-step analysis of the four possible sentence structures revealed a much more complex picture (see Figure 4).

First, in the sentences without pro-drop (SV and SVO), the networks recognized the familiar sentence-initial noun (or pronoun) without any problems. They also did a reasonable job (87%) on the next, novel word by correctly parsing it as the verb. If the sentence ended there (SV), this verb interpretation for the novel word became even stronger (92%). If the sentences continued with a (familiar) object noun (SVO), this word clearly came unexpected and interfered with the stored verb representation: at the end of SVO sentences, only 55% of the verbs were still correct. The difference between the transitive and intransitive sentences here is partly due to the increased length of the former (and the resulting increased memory load), but also to the difference in frequency between SV and SVO sentences in the training corpus: overt subjects were more likely to occur with intransitive verbs, so the network apparently assumed that the novel verbs following an overt subject would be

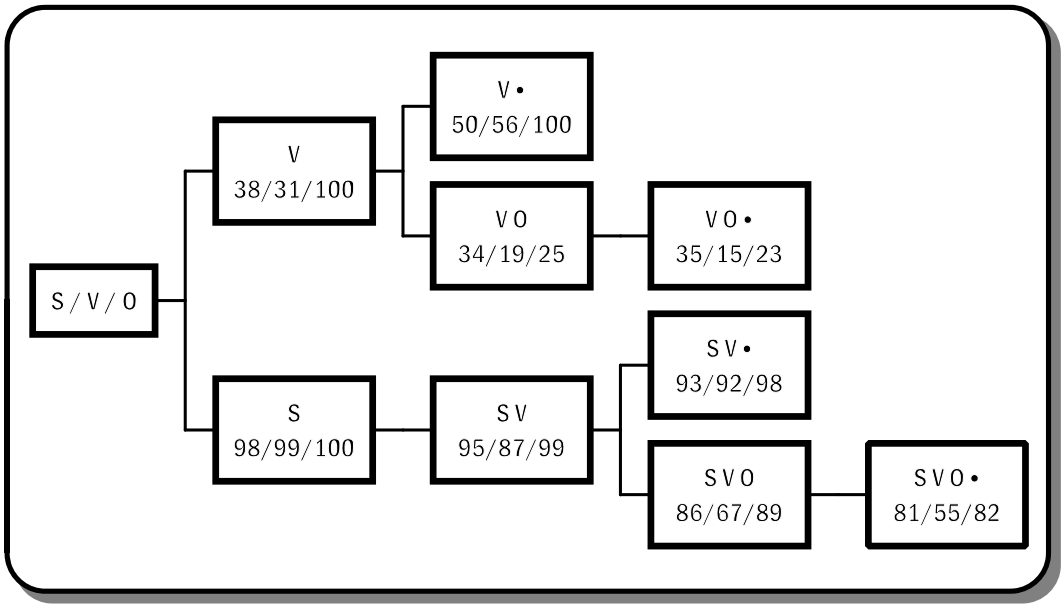

Figure 4. Word-by-word analysis of how the networks that were trained on the difficult language type for 30 epochs parsed test sentences with novel verbs. The numbers in each box show the percentages of Subjects, Verbs, and Objects that were analyzed correctly after having seen the part of the sentence shown in the same box. E.g. after having seen just the Verb, 38% of the Subjects, 31% of the Verbs, and 100% of the Objects were correct at the output.

intransitive as well. When this assumption turned out to be false, the resulting confusion negatively affected all three output banks.

Second, in sentences with pro-drop (V and VO), the presence of a novel word in sentence-initial position also led to immediate and severe problems: only 38% of the subjects and 31% of the verbs were correct after parsing the first word. (Note however that the networks still remembered that objects (100%) could never occur as the first word in a sentence.) If the sentence ended there (V), the networks managed some

limited recovery as both the subject and object output banks improved to more than 50%. But when a noun or pronoun followed as the object (VO), even the fact that it was a known word was insufficient to remove the original uncertainty: performance on all three output banks dropped considerably, with only 15% of the novel verbs parsed correctly at the end of the sentence.

Finally, let's examine the results when both nouns and verbs were novel (Table 12). Given the negative effect of both separately, it comes as no surprise that even fewer test sentences were processed correctly when both were novel. The overall performance on these test sentences was only 44.2%, a number which post-hoc Tukey comparisons again show to be significantly worse than any of the other language types; MS = .82842, df = 228.00, $p<0.0001$. For the sentences with pro-drop (V and VO), the errors made by the networks here are almost the same as the ones shown in Figure 4 for novel verbs. Most of the additional effect of the novel nouns can be found in the sentences with an overt subject (SV and SVO), as there are no longer any familiar sentence-initial nouns to help guide the interpretation of the sentences. Consequently, the presence of novel nouns leads to the kind of immediate confusion described above for Table P. It is only exacerbated by the following novel verb form. Performance at the end of the sentence on both SV (73.3%) and SVO (33.3%) sentences was thus noticeably worse than when only the verbs were novel (87.6% for SV, 45.9% for SVO). The fact that these networks still got 73.3% of SV sentences correct may be surprising, but it is a reflection of the fact that the first word in the training set was most likely to be the subject, and the second word the verb. As a result, the networks

developed a default sentence interpretation which reflected this tendency. The test corpora had the same distribution of the four possible sentence structures as the training corpora so this default interpretation worked well for novel SV sentences.

In short, the networks learning the difficult language type experienced major problems with the generalization task, both when the nouns were novel but especially when the verbs were unfamiliar. The reason why only this one language type was problematic leads us again to the usefulness of the different sources of linguistic information for generalization tasks. What sets the difficult language type apart from all the other ones is that the only reliable source of information for parsing sentences was the lexical identity of each word and the lexical category it belonged to. Recall from Experiment 2 that the networks learning the difficult language type were the ones which benefited most from 20 additional training epochs. The extra exposure to the lexical items in the lexicon allowed them to learn the noun/verb status of every word, and this distinction in turn enabled them to become much better at deciding 'who did what to whom'. In this experiment, the networks' increased knowledge of the words in the training corpora was of limited use when they were asked to process completely novel words. Unlike grammatical word order patterns or morphological markers, knowledge of the properties of a single word simply cannot be used when this word is not present in the sentence. Consequently, lexical identity knowledge is

not as useful a cue for the models as the other sources of information we have looked at.[19] We will return to the implications of this finding in section 8.1.

The second question we wanted to investigate with this experiment concerns the effects of novel nouns as opposed to novel verbs. In the other language types, we only found very small differences between the two. But the results from the difficult language type suggest that novel verbs can cause significantly more problems than novel nouns, at least in pro-drop languages. As we have seen, the presence of pro-drop is the crucial factor here as it causes verbs to appear in sentence-initial position. When these verbs are novel, the uncertainty about how to parse the sentence starts with the first word, and the networks only seldom recover from the resulting structural ambiguities. This suggests that the problems experienced by the networks in this condition are due to the frequency of pro-drop rather than individual properties of novel verbs. In other words, performance in the novel verbs condition is not worse because the novel words are verbs but because these novel words occur more

[19] Additional support for this conclusion can be found by comparing the networks' performance after 10 epochs of training versus 30 epochs (see Table 13 below). The data shows that overall performance actually became somewhat worse with more training (45.8% after 10 epochs; 44.2% after 30 epochs) when both the nouns and verbs were novel. In essence, additional exposure to the lexical items in the training set created a greater dependence on them for parsing, and this optimization for the training set worked against the networks when they were faced with all novel words.

Table 13. The performance of the networks learning the difficult language type on four different test corpora after 10 and 30 epochs of training.

Training cycles	Test corpus, 'difficult' language type			
	Familiar words	Novel nouns	Novel verbs	Novel nouns + verbs
10	73.4%	69.3%	49.1%	45.8%
30	91.4%	80.2%	52.8%	44.2%

frequently in initial position than novel words do in the novel nouns condition (see section 5.1).

It is worth pointing out that the same pattern could already be observed after only 10 epochs of training (Table 13). At this time, the networks faced with novel verbs only got 49.1% of the test sentences correct versus 69.3% of the sentences with novel nouns. The percentages in Table 13 also show that the networks learning the difficult language type initially paid more attention to the verbs in the training corpora. When these were learned after 10 epochs, there was a relatively small difference between the networks faced with novel nouns (69.3%) and those seeing all familiar words (73.4%). The novel verbs condition, on the other hand, did much worse (49.1%), partly due to the position of the novel word effect just described, and partly due to the sudden uselessness of the lexical knowledge already learned by the networks. However, it would be wrong to think that the networks first learned all the verbs, and only then the nouns. The numbers in Table 13 demonstrate that nouns and

verbs both benefited from the additional training: the jump from 69.3% to 80.2% in the novel nouns condition was the result of better knowledge of the verbs in the test corpus, not the unfamiliar nouns; similarly, the limited improvement from 49.1% to 52.8% in the novel verbs condition shows that most of the useful information which could be gleaned from nouns had already been extracted after 10 epochs.

In summary, the ability to identify nouns and verbs was crucial for the networks learning the difficult type and it was more important for verbs than nouns. In the absence of other grammatical cues, novel words could be impossible to categorize correctly, especially when they occurred in the sentence-initial position. Additional training did not help significantly because it only increased the dependence on knowledge of familiar words.

6.2 Linguistic discussion

In this section, we evaluate how the networks' performance reflect on the two main questions that inspired Experiment 3: first, do the models support the typological finding that pro-drop languages favor early verb learning over noun acquisition; second, do they display sufficient generalization capability.

6.2.1 Nouns first or verbs first in pro-drop languages

The hypothesis that nouns are learned before verbs in all languages has recently received a lot of attention in studies of language acquisition, so we briefly

discuss here how our simulations reflect on this issue. Prototypical nouns refer to concrete, animate entities which can be observed easily in the world. Prototypical verbs, on the other hand, have meanings that are harder to observe and thus require more linguistic context to establish. These conceptual differences suggest that nouns should be easier to learn. Studies of children learning English have consistently supported the primacy of nouns hypothesis (e.g. Gentner 1982; Goldfield 2000).[20]

However, analyses of other languages has revealed a more complex picture, with Korean (Choi and Gopnik 1995; Gopnik, Choi and Baumberger 1996; Choi 2000; Kim, McGregor and Thompson 2000), Mandarin Chinese (Tardif 1996, 2006; Tardif, Shatz and Naigles 1997; Tardif, Gelman and Xu 1999), and Tzeltal (a Mayan language; Brown 1998) all being described as 'verb-friendly' languages in which the first verbs are learned at least as quickly as the first nouns. These three languages are typologically quite distinct in their word order and morphology (Korean: an SOV language with case-marking and tense/polarity verbal morphology; Mandarin: an isolating SVO language with no case-marking and limited verbal morphology; Tzeltal: a VOS language with some case-marking and rich verb agreement), but a common feature is that they all allow referential pro-drop. The putative connection between early verb

[20] There is a related but separate issue which we do not go into here, and that is whether the lexical category of nouns develops before that of verbs. The existence of the abstract category can be tested by seeing how freely children apply morphological markers associated with the category: e.g. English *–ing* for verbs, or plural *–s* for nouns. For most languages, the noun category develops before the verb category (Tomasello 1992; Akhtar and Tomasello 1997; Tomasello, Akhtar et al. 1997; Childers and Tomasello 2002; Naigles 2002), perhaps because more nouns are learned initially and the required critical mass to establish the category is thus reached earlier.

acquisition and pro-drop is simple: when nouns are left unexpressed, verbs not only become relatively more frequent, but they are also more likely to appear in prominent positions (i.e. as first or last word in a sentence), or even be the only word left in the sentence. As a result of their increased salience, children would presumably pick up verbs more easily. The developing consensus is that nouns do indeed have a conceptual acquisition advantage, but some languages simply have linguistic properties which can make the verbs sufficiently salient to offset this advantage (Kim, McGregor and Thompson 2000; Sandhofer, Smith and Luo 2000; Gentner and Boroditsky 2001; Bornstein, Cote et al. 2004).

Comparing our simulations to the acquisition data, we find that the most straightforward method of determining whether the networks learn their nouns or verbs preferentially is also the least interesting one. We can easily measure that the verbs in the lexicon are learned first (i.e. after fewer training cycles), but this finding is hardly surprising because it just reflects that the verbs in the simulations (as in natural languages) have higher token frequencies than the nouns. What is missing from the simulations is the conceptual and representational advantages which animate nouns have over all other kinds of words. Until we can implement this type of cognitive advantage in a computational model, a direct comparison of the acquisition rates of nouns and verbs is of limited use.

However, there is another way in which we can test whether verbs or nouns may be at a disadvantage. Instead of comparing how quickly the words from both lexical categories are acquired, we can sidestep the conceptual representation issue by

Table 14. The impact of novel verbs on all the language types. The numbers in the cells are the differences between the performance of each language type on test corpora with familiar words vs test corpora with novel verbs.

– PRO		N-marking	
		—	Case
V-marking	—	1.0%	0.7%
	T/A/M	0.5%	0.4%
	Agr	0.7%	0.8%

+ PRO		N-marking	
		—	Case
V-marking	—	38.6%	2.4%
	T/A/M	0.9%	0.7%
	Agr	1.4%	1.4%

Table 15. The impact of novel nouns on all the language types. The numbers in the cells are the differences between the performance of each language type on test corpora with familiar words vs test corpora with novel nouns.

– PRO		N-marking	
		—	Case
V-marking	—	0%	0%
	T/A/M	0%	0%
	Agr	0%	0%

+ PRO		N-marking	
		—	Case
V-marking	—	11.2%	0.3%
	T/A/M	0.1%	0.1%
	Agr	0.2%	0.2%

looking at the generalization task. Our reasoning is as follows: if the presence of pro-drop in a language leads to greater dependence on verbs during processing, then the presence of novel verbs should have a bigger impact on the languages with pro-drop. This prediction is confirmed across the board. Tables 14 and 15 summarize the difficulty caused by the presence of novel verbs or novel nouns for each language

type. Each of the numbers is the difference between the performance on the corpora with familiar words (91.4% for the difficult type; 99.9% for all other cells) and those with novel verbs or novel nouns. Obviously, the difficult language type is affected most by the presence of novel verbs, but an ANOVA test shows that the presence of novel verbs also leads to significantly worse results in the five other types with pro-drop when compared to their five counterparts without, $F(1,198) = 66.299$, $p<0.0001$. This result is compatible with the cross-linguistic generalization that pro-drop leads to increased salience for verbs (and thus earlier acquisition). Similarly, our prediction that novel verbs should have a bigger effect than novel nouns in pro-drop languages is also confirmed. Even if we exclude the difficult language type, novel verbs lead to significantly worse performance in pro-drop languages than novel nouns, $F(1, 198) = 191.07$, $p<0.0001$.

In short, these modeling results bolster the arguments made for languages such as Korean, Mandarin and Tzeltal by linking the existence of pro-drop to an increased reliance on verbs. However, they also indicate that languages may differ with respect to the noun vs verb bias in acquisition. Even pro-drop does not always entail a big boost for early verbs. For example, Italian is typologically similar to Mandarin in combining pro-drop, an SVO word order and a lack of case marking. However, it also features rich verb agreement, and this difference may be (part of) the reason why children learning Italian initially acquire fewer verbs than children Mandarin Chinese – though still more than children learning English (Caselli, Bates et al. 1995; Tardif, Shatz and

Naigles 1997; Camaioni and Longobardi 2001). Note that in the simulations, the presence of rich verb agreement also greatly reduces (38.6% vs 1.4%) the impact of novel verbs. One of the reasons rich agreement may slow down verb learning is that it reduces the morphological transparency of verbs – i.e. each verb can take many different forms and it will take a child a while to determine that they are all based on the same verb. Similarly, in language production, the child not only has to know how to say the root of each verb, but also any required agreement markers. It seems reasonable to assume this is a more complex task than producing a verb by itself in an isolating language like Mandarin.

6.2.2 Generalization in language development

The main goal of Experiment 3 was to determine whether our neural networks can generalize their processing strategies for determining 'who did what to whom' to sentences containing novel words. We needed to demonstrate this ability in our simulations because linguistic generalization is such a common phenomenon that it can reasonably be thought of as a requirement for any model of language learning. One of the areas which has been studied extensively in this regard is how early and how quickly children acquiring English become productive with novel nouns and verbs. In the seminal work of Berko-Gleason (1958), the focus was on productive morphology; she found that the 4 to 7-year-olds in her study had few problems learning a label for a novel object, e.g. *wug*, and also immediately treated it like a regular count noun in production, e.g. by referring to multiple instances of the object as *wugs*.

Interestingly, the labels for novel actions were not learned as quickly, and the children were also more reluctant to treat them as regular verbs by adding verbal morphology such as *–ing* or *–ed*. Similar 'wug test' experiments have since been carried out with increasingly younger children (e.g. 17 to 21-month-olds in Tomasello, Akhtar et al. 1997), but the basic findings have remained the same: children can reliably learn novel nouns at an earlier age than novel verbs, and whereas children become productive with nouns around two years of age, verbs are only used as freely by the same children when they are three to four years old (Olguin and Tomasello 1993; Behrend, Harris and Cartwright 1995; Forbes and Farrar 1995; Tomasello, Akhtar et al. 1997; Gillette, Gleitman et al. 1999; Abbot-Smith, Lieven and Tomasello 2001, 2004; Childers and Tomasello 2002; Lieven, Behrens et al. 2003; Gleitman, Cassidy et al. 2005; May Vihman and Vija 2006; Uziel-Karl 2006).

It has also been found that within the larger category of verbs, the transitive-intransitive distinction is important in early language development. Between 2 and 3 years of age, children use verbs conservatively and only gradually become willing to use a novel verb with a different number of arguments than they have heard in the experimental setting. So, if a verb has only been presented in transitive contexts (e.g. *Bert is dacking Ernie*) then young children are unlikely to produce *Elmo is dacking*, or even to agree that a scene showing the same action without a second participant can be described as *dacking* (Tomasello 1992; Olguin and Tomasello 1993; Akhtar and Tomasello 1997; Abbot-Smith, Lieven and Tomasello 2001; Fisher 2002; Naigles 2003; Gleitman, Cassidy et al. 2005; Naigles, Bavin and Smith 2005).

Why children should be sensitive to the transitive/intransitive distinction in verbs is currently a topic of much debate between proponents of syntactic bootstrapping and those who advocate a more cognitive 'constructivist' approach. Before we discuss both viewpoints in more detail, we should point out that the two sides are not that far apart, at least when compared to the traditional distinction between the 'formal' linguists favoring a fully innate syntactic Universal Grammar, and the 'functional' linguists advocating that language is essentially about semantic and pragmatic phenomena (but see Tomasello and Abbot-Smith (2002), Lidz, Waxman and Freedman (2003), and Tomasello (2004) for a revival of this old discussion). Both proponents of syntactic bootstrapping and constructivists believe that children initially build very detailed lexical representations that reflect the linguistic contexts in which the children have observed the forms (e.g. Gillette, Gleitman et al. 1999; Cameron-Faulkner, Lieven and Tomasello 2003; Savage, Lieven et al. 2003; Gleitman, Cassidy et al. 2005). Both sides also believe that commonalities between these representations support the gradual development of more abstract generalizations and grammatical patterns. For both sides, then, children's parsers are performing a distributional analysis of the words they hear (Cartwright and Brent 1997; Mintz 2003).

What the two sides do not agree on is whether there is a role reserved for innate syntactic biases. Proponents of syntactic bootstrapping argue there is experimental proof of very early linguistic knowledge in infants, and that much of this knowledge is the result of innate (probably language-specific) processing mechanisms that pay attention to the formal characteristics of the language input observed by the

child (e.g. Naigles and Kako 1993; Saffran, Aslin and Newport 1996; Aslin, Saffran and Newport 1998; Gillette, Gleitman et al. 1999; Fisher 2002; Naigles 2002, 2003; Gleitman, Cassidy et al. 2005). Evidence cited in support of this position includes the finding that 2-month-olds can already distinguish the order of syllables and words in a sentence, as long as it is presented in a single coherent prosodic unit (Mandel, Kemler Nelson and Jusczyk 1996). Similarly, an analysis of prosodic properties of child-directed speech from English and Japanese corpora revealed that the end of utterances, and even individual phrases, can be detected using acoustic information alone (Fisher and Tokura 1996). At the word level, it has been demonstrated that 9-month-olds can learn the stress patterns of polysyllabic words and generalize them to novel words; this is an ability that could be used to detect similarities between words and thus form a basis for word classes (Gerken 2004). Finally, even quite complex linguistic patterns such as non-adjacent dependencies have been observed to be learnable by 18-month-olds. Gomez (2002) found that the children in her experiment would distinguish between different artificial languages on the basis of non-adjacent relationships if no simple adjacent cues were available. What all these experiments have in common is that the children have been found to be sensitive to formal patterns which lack an obvious semantic counterpart. Likewise, the transitive/intransitive distinction mentioned above is taken to reflect innate argument principles for argument mapping (such as the theta criterion), combined with language-specific data about the number, position and type of arguments generalized on the basis of external input (Mintz 2003; Gleitman, Cassidy et al. 2005).

It is exactly this need for innate structural principles that is called into question in 'emergentist constructivism'. Constructivists posit that the striking abilities displayed by young infants only reflect the types of general cognitive pattern analysis which can also be observed in other species, as well as in non-linguistic tasks (e.g. involving music or blinking lights; Tomasello and Akhtar 2003). Real linguistic knowledge, on the other hand, is taken to develop slowly and to crucially depend on the children's understanding the meaning of forms involved, as well as the children's intentional desire to engage in social communication (e.g. Tomasello 1992, 1998, 2000; Bloom, Margulis et al. 1996; Lieven, Behrens et al. 2003; Goldberg, Casenhiser and Sethuraman 2004). Conceptual, rather than structural, processes are taken to play a key role in initial acquisition and development. As a result, the early sensitivity to the different behavior of transitive and intransitive verbs is linked to the salience of transitivity and the prototypical scene in which an animate, causal agent visibly affects an inanimate agent (Talmy 1988; Dowty 1991; Herr-Israel and McCune 2006). Relevant evidence for this position includes a widespread preference in language development to first express grammatical structures like meaningful word order (e.g. Subject – Verb – Object) or morphological marking for the description of prototypical transitive situations (Slobin 1973, 1982; Naigles 2003; Childers and Echols 2004).

The models presented here suggest a promising middle ground between the nativist and the constructionist positions because they show that the desired cognitive phenomena can appear without innate syntactic or rich conceptual scaffolding. As we have seen, the networks managed to learn the 'who did what to whom' task quite well

for all but one of the language types we investigated. This finding was not unexpected for sentences made up of novel combinations of known words (Experiments 1 and 2), because it mirrors the corpus research which has been done using distributional analysis techniques (Burgess and Lund 1997; Cartwright and Brent 1997; Redington, Chater and Finch 1998; Mintz, Newport and Bever 2002; Mintz 2003). However, the fact that it also holds for generalization to sentences with completely novel words (nouns, verbs or both in Experiment 3) demonstrates that useful information about the basic structure of a language is available if we look at morphological markers and word order patterns. Crucially, our networks learned to solve 'who did what to whom' despite the absence of an innate theta criterion or any kind of semantic or pragmatic information.

One may wonder about the cognitive importance of this result given that the models learned the task using explicit feedback from the backpropagation algorithm – i.e. a supervised learning method in which the correct answer is always available to the networks. In this regard, it is worth looking at what the models did with the distinction between transitive and intransitive verbs, because keeping these two classes apart was not something they were trained on. This distinction was also learned by the models, as its effect can be observed quite easily in the difficult language type. The relevant data is found in transitive sentences with pro-drop (i.e. VO sequences). In the novel nouns condition of Experiment 3, only the initial verb was familiar. Despite the nouns being novel, 91% of them were parsed correctly at the end of the sentence. In the novel verbs condition, on the other hand, performance on the object nouns in VO

sentences was only 30%, despite the fact that the networks knew these nouns very well from training. This big (and somewhat counter-intuitive) difference makes sense when we take into account that in the novel nouns condition, the networks knew that some verbs in the language are followed by another word and this word is the object of the sentence – i.e. they are transitive verbs. It is also interesting that it is the presence of this additional argument which makes these verbs stand out as a class: the sequence SV can occur both as a complete intransitive sentence, or as the beginning of a transitive one. This provides complete transitive sentences with a structural salience, independent of its link to a conceptual agent-patient scene.

Recall that the transitive/intransitive distinction is not built into the model, and that it is not even required to solve the 'who did what to whom' problem. Instead, the distinction emerged spontaneously during training as the result of the learning algorithm forcing the connections between the units to develop representations that could be used to decide which words were the subject, verb or object. The representational capacity of the networks was limited so the representations for words that behaved alike also started to look alike. By storing transitive and intransitive verbs differently, the networks were able to perform their task more efficiently: i.e. they developed expectations about the possible presence of a following word which made the sentences easier to parse (Elman 1990). As a result, parsing could be done successfully even when this next word was unfamiliar.

The middle ground referred to earlier is thus the proposal that the capacity which children acquire for basic 'who did what to whom' parsing (as well as for

distinguishing transitive and intransitive verbs) needs less conceptual ability than constructivists posit, but simultaneously also less innate syntactic machinery than implied in syntactic bootstrapping accounts. If we look at what is built into the models – as opposed to behavior which emerges during training – there are two 'innate' mechanisms. First, there is the ability to segment sentences into individual words at the input layer. But this appears to be a reasonable assumption as it is something even very young infants (as well as other species such as tamarin monkeys) can do quite easily (Mandel, Kemler Nelson and Jusczyk 1996; Hauser, Newport and Aslin 2001; Mattys and Jusczyk 2001; Saffran and Wilson 2003; Weiss and Newport 2006). Second, and potentially more controversial, there is the task of assigning these words to distinct slots at the output layer. In this regard, it bears repeating that the 'subject', 'verb' and 'object labels which we have assigned to the slots imbue them with more meaning than they really have from an architectural point of view (see section 3.2 above). Especially the 'subject' and 'object' slots could also have been called 'NP1' and 'NP2' because there is no specific theoretical definition of 'subject' vs 'object' underlying them.[21] We think that the output slots in our models are sufficiently vague that they can stand in (however poorly) for the more developed cognitive structures assigned to words. It is exactly the fact that these underspecified representations can be used by our models to solve both the explicit 'who did what to whom' task as well as the implicit transitive/intransitive distinction which leads us to believe that both less syntax and

[21] In fact, a very similar output architecture was used to represent semantic roles such as agent, experiencer, and patient in Morris, Cottrell and Elman (2000).

less semantics may be involved in early language processing – at least in SVO languages – than what is generally argued for. Rather than positing 'subject'-'object' or 'agent'-'patient' representations in the first stages of language development, 'entity 1' and 'entity 2' could well be sufficient at first.

At the same time, we also want to stress that we are not denying the importance of both form-driven and meaning-driven processes in language acquisition. In this section, we have mentioned several experimental studies which have demonstrated the influence they can have on specific tasks. Moreover, it could be claimed that the reason the models make do with so little syntactic complexity is only because the input language they observe is overly limited in the types of linguistic structures it includes. However, we want to point at previous work presented in Van Everbroeck (2003) in which the artificial languages learned by the models also featured possessive phrases, locative phrases and relative clauses. Those connectionist simulations showed that additional complexity will slow learning, but even the most complex language was learnable given sufficient exposure (and the presence of word order and morphological cues). In addition, it has been shown that learning of complex languages will often be more efficient if it is at first restricted to the simple patterns in the input (Newport 1990; Elman 1993).

However, the lack of basic conceptual representations in the models is a greater deficiency, because human languages undoubtedly have communicative intent as their driving force (*pace* Chomsky), and the model lacks the cognitive abilities of even very young infants (e.g. Baillargeon 1987, 1998, 2004). So one could make a

reasonable case that it is simply incapable of displaying the semantic effects observed in children. Nonetheless, infants become sensitive to linguistic distinctions long before they can use them for their own communicative purposes, so it remains an open question as to how much semantic and pragmatic knowledge we must assume to account for all the acquisition data. Simulations like the ones we present here can thus be used to investigate how important a role meaning plays in various parts of language acquisition.[22]

6.3 Summary

In Experiment 3, we investigated how well the different language types could generalize their acquired parsing strategies to sentences containing novel words (nouns, verbs, or both). We found that this generalization task was remarkably easy (at least 97% correct) for all the language types which featured a reliable word and/or morphological markers. The 'difficult' language type, however, was impacted severely, especially when the novel words were verbs. This finding shows that the difficult

[22] The usefulness of addressing this question experimentally has been illustrated by models of language evolution such as Hare and Elman (1995) and Polinsky and Van Everbroeck (2003). Using connectionist simulations, Polinsky & Van Everbroeck modeled the evolution of the three-gender nominal system of Vulgar Latin (neuter, masculine, feminine) into the two-gender system of Old French (masculine, feminine). Their results showed that many of the gender changes which had often been described as having a semantic basis (e.g. all trees should be the same gender) could also be accounted for by learning mechanisms which only had access to the phonological shape of the nouns in the lexicon. The goal of this work was also not to discount the importance of semantic/conceptual similarities in processes of language evolution, but to investigate whether they were required or not.

type's reliance on lexical identity knowledge not only makes it slower to learn 'who did what to whom', but it is also less capable of supporting generalization.

There were two other notable findings. First, some important linguistic distinctions such as the transitivity distinction can be learned using less syntactic or conceptual knowledge than often posited. Second, the results of the models supported the hypothesis that the presence of pro-drop in a language will make its verbs more salient than in similar languages without pro-drop. This increased verb salience can in turn lead to more verbs being acquired early on.

Chapter 7. Experiment 4

The goal of this experiment is to determine how the presence of noun/verb homonymy affects language learnability. Noun/verb homonyms are words that can be used as a noun or a verb without a change in surface form; the meanings of the noun and verb can be closely related (e.g. English *to paint* vs *paint*) but may also be unrelated (e.g. English *to trip* vs *a trip*). The existence of such homonymy in English is a side-effect of the lack of morphological marking in the language. Given that the majority of the world's languages (semi-)consistently mark verbs and/or nouns in one way or another, homonymic noun/verb pairs are actually not that common. However, for English, an analysis of the CELEX database reveals that about half of the word types which are used as nouns or verbs can also be used as a member of the other category (Baayen, Piepenbrock and Gulikers 1995; see Table 18 below). This is not just a property of esoteric words, because even if we limit ourselves to the 2,000 most frequent words in the English language, 601 of them (30%) can be used as both nouns and verbs.[23]

There are two reasons why the homonomy issue is relevant for our simulations. First, noun/verb homonyms are a common phenomenon in languages without much morphology – e.g. English and Mandarin (e.g.Clark 1993; Li 1998;

[23] These percentages are somewhat lower in the spoken language because otherwise homonymous words may still have different stress or tone patterns.

Sandhofer, Smith and Luo 2000; Li, Jin and Tan 2004). However, it is also a phenomenon which has received very little attention in the literature on cross-linguistic variation. While homonymy has been studied in individual languages, we know of no studies that compare the level of homonymy in different languages or across languages types, or that try to link homonymy to other linguistic parameters. By implementing this phenomenon in our simulations, we are able to investigate its effect systematically and look for such linguistic correlations.

Second, homonymy is known to affect natural language processing. Homonymous words create inconsistencies for the language processor because they need to be disambiguated before their exact contribution to the meaning of their sentences can be determined. Recent experiments with children have also shown that assigning a second meaning to a known form – i.e. learning a homonym – is more difficult than assigning the very same meaning to a novel word. In many instances, younger children use the primary, more common meaning of the homonym even when contextually inappropriate (Campbell and Bowe 1983; Beveridge and Marsh 1991; Mazzocco 1997; Doherty 2000, 2004; Mazzocco, Myers et al. 2003). For example, when 4-year-old children are told a story about a castle wing and are then asked to draw this wing, they are more likely to draw a bird's wing, rather than part of a castle. Most of these findings are based on experiments in which children are asked to learn noun/noun homonyms, but Casenhiser (2003) has found that children learning English also have a significant dispreference for assigning a new noun

meaning to a world already known as a verb (e.g. *eat*), as opposed to assigning the same new meaning to a novel word (e.g. *mack*).

Because homonymy creates lexical ambiguities between nouns and verbs, we can expect it to make the 'who did what to whom' task more difficult for the models. With homonyms, simply recognizing the form of the word is no longer sufficient to determine its lexical category. Some other source of information is needed to help decide whether the word is a noun or a verb. Consequently, we predicted that the language type which only has access to lexical identity to distinguish nouns from verbs – i.e. the difficult type – would be especially affected by the presence of homonymy. This prediction is compatible with wide-spread homonymy in English, because the position of the homonymous word in the sentence is usually a good indicator of its lexical category. Mandarin Chinese, on the other hand, is significantly closer to the difficult type, and the existence of frequent pro-drop entails that sentential position by itself cannot disambiguate homonymous words. Because of this difference, we expected that the language type resembling English would allow for more homonymy than the one resembling Mandarin. We will see below that the network results confirm this prediction. We will also present homonymy counts from Mandarin and English corpora supporting this result.

7.1 Network results

Given the lack of frequency data on noun/verb homonymy, it is hard to establish a realistic baseline for the phenomenon. With that in mind, we designed the

experiment to compare the effects of varying levels of homonymy. The levels which we set were: 5%, 10%, 25%, and 50% noun/verb homonymy. For example, in the 10% condition, 40 of the 400 words which were previously either a noun or a verb could now belong to both categories. These homonyms were split evenly between the various possible combinations of nouns and verbs: e.g. in the 10% condition, 2.5% of the homonyms could appear as an animate noun or a transitive verb, 2.5% as an animate noun or an intransitive verb, and the other 5% was used for the corresponding inanimate noun combinations. This equal distribution of the various noun/verb combinations is unrealistic, but it is a reasonable starting point given the lack of clear generalizations about the phenomenon in human languages (Nichols, Peterson and Barnes 2004). By keeping the combination types equally frequent, we also avoided biasing the model towards any particular word class. The new lexicon files with the homonymous words were then used to generate training and test corpora with 3,000 sentences each. The networks were trained for up to 100 epochs, and we kept track of their performance on the test corpora after 10, 30, 50, 75, and 100 epochs. The performance measure remained the same as in the previous simulations: i.e. the percentage of test sentences in which each of the words was assigned to the correct output slot (S, V, or O).[24]

---

[24] There is a side-effect of an original experimental design choice that is worth mentioning here. One may recall that the words in the lexicon files were not associated with specific frequencies. For example, each animate noun had an equal chance of being used in a sentence. In the previous experiments, this design choice resulted in all animate nouns being learned at the same rate, with minor differences depending on how frequent each individual word was in a particular training set. With the introduction of homonymy into the simulations, the lack of individual word frequencies resulted in all homonymous words sharing the same

Table 16. Results from Experiment 4. Test on familiar words, 50% homonymy (10 epochs).

– PRO		N-marking	
		—	Case
V-marking	—	99.4%	99.4%
	T/A/M	98.9%	99.5%
	Agr	99.3%	99.4%

+ PRO		N-marking	
		—	Case
V-marking	—	56.8%	98.6%
	T/A/M	99.1%	99.2%
	Agr	98.9%	98.9%

We discuss the unproblematic types and the difficult language type separately. For the former, the simulations show that even the highest level of homonymy (50%) did not have much impact on learnability. As summarized in Table 16, 10 epochs of training suffice for almost perfect performance on the test corpora. (After 30 epochs, 99.9% of the test sentences were processed correctly for all these types.) This finding is not surprising if we take into account that all but one of these language types featured case-marking and/or verb markers which essentially removed any real homonymy from the languages. As we have seen before, the reliable presence of

---

noun-to-verb frequencies. Due to there being more nouns than verbs in the lexicon files, each homonym was thus less likely to appear as a noun than as a verb. This tendency became even stronger in the language types with pro-drop, because nouns could be omitted whereas verbs were always present. The homonyms in natural languages, on the other hand, display widely varying frequency differences between the noun and verb pairs. We plan to implement such differences in the next round of simulations, but an advantage of the current design is that it makes the networks' task harder. In natural languages, a noun may occur 20 times as frequently as the homonymous verb, so frequency-driven learning algorithms would quickly consider the ambiguous form to be a noun wherever it occurs. By keeping the noun-to-verb ratios relatively close, our simulations made it easier to study the effect of homonymy by itself.

morphological markers on just one of the two lexical categories is sufficient for both of them to be processed correctly. The networks learn that the absence of any markers on a word indicates that this word belongs to the other class. However, the networks learning the English-like language type, i.e. without case or head-marking (but also without pro-drop) performed at the same level as the networks which did have access to morphological cues. This result demonstrates that the structural position in which a homonymous word appears in a sentence can also be sufficient to disambiguate its lexical category. As long as some source of information is available, the presence of even rampant homonymy thus does not present any major learning difficulties.

When we look at the score for the difficult language type in Table 16, we find that only 57.6% of the test sentences in these corpora were processed correctly. This percentage is considerably worse than the 73.4% scored by the networks learning the difficult language without any homonymy. However, Table 16 also represents a worst case scenario. It is based on a very high degree of homonymy and only 10 epochs of learning. We have seen before that additional training epochs are quite beneficial for the difficult type because it provides these networks with more exposure to the problematic data. A better overall picture of the effects of noun/verb homonymy is provided in Figure 5, which summarizes the performance of networks learning the difficult type with increasing levels of homonymy (0-50%) and after increasing amounts of training (10-100 epochs).

The data in Figure 5 supports three conclusions. First, noun/verb homonymy makes the difficult language type harder to learn. Whether after 30 epochs of training

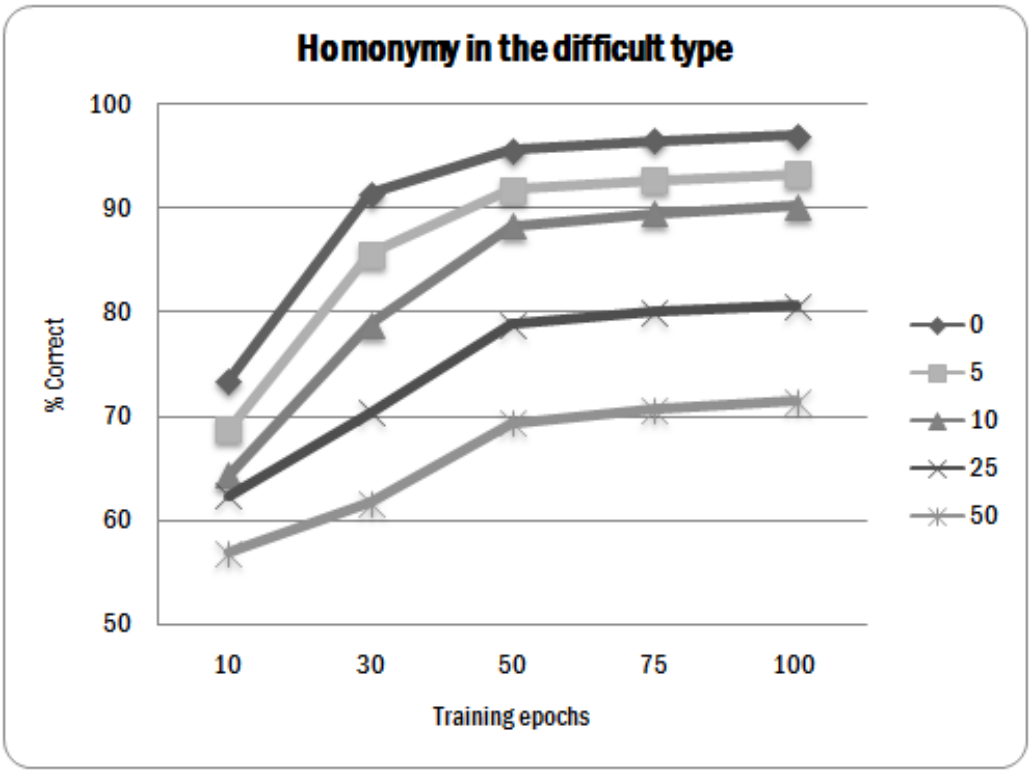

Figure 5. The interaction between amounts of training and homonymy. Percentages of test sentences processed correctly by the neural networks learning the difficult language type after varying amounts of training (10-100 epochs) and learning languages with varying amounts of noun/verb homonymy (0-50%).

or 100, the difference in performance between the language without homonymy and the one with only 5% of it is highly significant; e.g. after 30 epochs, a one-way ANOVA shows an effect of homonymy: $F(1, 38)=53.900$, $p<.0001$. Second, increasing levels of homonymy lead to reliable decreases in performance. None of the lines in Figure 5 cross, with the only close call being for the 10% and 20% levels after 10 epochs of training (64.3% and 62.3%, respectively). Third, prolonged exposure to

the homonymous words still allows for acceptable overall performance for low levels of homonymy, i.e. up to 10%. All the networks improve their scores with additional training, but only limited progress is made after 50 epochs. The improvements made by the networks learning languages with lower levels of homonymy are also much larger than those made by the networks in the 25% and 50% conditions. E.g. the 2% difference in performance between the 10% and 20% conditions after 10 epochs of training has become much larger after 50 epochs (88.3% and 78.9%, respectively). In short, despite its immediate negative impact, even the difficult type still appears to tolerate some level of homonymy. We return to this finding in the next section, when we compare the available homonymy data for English and Mandarin Chinese.

The last issue we address here is the number of errors made in the homonymy simulations. As was to be expected, the networks make most of their mistakes when they are processing homonyms which appear early in the sentence. The crucial problem with the difficult language type is that the networks need to learn the lexical category of each word before they can decide whether the first word in a sentence is the subject noun (SVO, SV) or the verb (VO, V). When there is noun/verb homonymy in the language, recognizing a familiar word may no longer be sufficient to identify its lexical category. The impact on processing is similar to what we have seen in the previous experiments when the models were not sure whether a (novel) word was a noun or a verb. I.e. when there is a homonym in sentence-initial position, the networks hedge their bets and copy the homonymous word to both the Subject and the Verb banks at the output layer. Higher levels of activation can be found in the

Verb bank, because homonyms appear more frequently as verbs than subject nouns (see above). Due to the networks' limited ability to recover from initial ambiguity, this difference in activation levels results in fewer subject nouns being processed correctly.

For example, after 10 epochs of training in the 5% homonymy condition, subject nouns in both intransitive (31.2%) and transitive (18.0%) sentences perform much worse than the verbs in the same intransitive (87.9%) and transitive (68.8%) sentences. Given additional exposure to the language, the subject nouns (intransitives: 85.1%; transitives: 75.4%) approach reasonable performance after 100 epochs, but they continue to trail the verbs (intransitives: 97.4%; transitives: 95.1%). Similarly, while the verbs in the 25% homonymy condition seem close to being learned after 100 epochs (intransitives: 90.6%; transitives: 81.1%), the subject nouns are still doing poorly (intransitives: 57.0%; transitives: 47.2%). Moreover, the limited amounts of improvement between epochs 50 and 100 suggest there is no reason to believe that additional training will lead to significant improvement. We can thus conclude here that the overall performance on a language depends on how well the subject nouns can be processed.[25]

---

[25] By comparison, in the 5% homonymy condition, all the pronouns and even the object nouns already perform above 90%. That the pronouns do well is not surprising; the subject and object ones have different forms which makes it easy for the networks to map them to the correct function. The object nouns are not case-marked, but they have two significant advantages over the subject nouns. First, object nouns are significantly more common than subject nouns, so the networks learn to expect them early on. Second, they benefit from linguistic context – unlike sentence-initial subject, object nouns follow other words which often provide clues about the function of the object word.

In summary, these simulations indicate that the presence of noun/verb homonyms has a negligible effect on the learnability of all language types but the difficult one. When neither morphological markers nor word order can reliably compensate for lexical category ambiguities, overall performance – and subject nouns in particular – suffers. However, even the difficult type can reach acceptable scores on the 'who did what to whom' task, as long as the level of homonymy is limited (up to 10%) and additional exposure to the language is available (at least 50 training epochs).

7.2 Linguistic discussion: homonymy in natural language

The main goal of Experiment 4 was to explore the impact of noun/verb homonymy on language learning. The results of our experiments suggest that the effects of noun/verb homonymy are very limited. In addition, they are not clearly correlated with any other linguistic parameters. Going by the simulations, we can conclude that the only scenario in which noun/verb homonymy will noticeably impact the processing of a language is when this language is both isolating (i.e. it lacks noun or verb morphology) and has pro-drop. That morphological markers compensate for the effects of noun/verb homonymy is not surprising, because they essentially prevent truly homonymous forms from appearing in the language – e.g. the *–ed* suffix in English unambiguously identifies *formed* as a verb, although its stem *form* could be either a noun or a verb. With the majority of the world's languages exhibiting regular morphological marking on nouns and/or verbs, it is thus not surprising that homonymy has not received much interest.

In the absence of reliable morphology, word order information can also compensate for noun/verb homonymy. For example, several studies have applied distributional analyses to large corpora of English data and found that the noun and verb forms of homonyms can easily be distinguished on the basis of the preceding words (Cartwright and Brent 1997; Redington, Chater and Finch 1998; Mintz, Newport and Bever 2002; Levy and Manning 2003). This finding is mirrored by the near-perfect performance of the models learning the language type which resembles English.

The crucial language type for us is the 'difficult' one, i.e. the type that lacks morphology but has pro-drop. In the simulations, the models experienced noticeable difficulties learning this type, even with low levels of homonymy. Interestingly, Mandarin Chinese, which is quite close to this language type (see section 4.2.2 above), has often been described as having rampant homonymy (Li 1998; Sandhofer, Smith and Luo 2000; Zhang, Wu and Yip 2006). Its presence does not appear to have much effect on adult processing. In a psycholinguistic study by Li, Shu et al. (2002), it was found that adult subjects will typically disambiguate (noun/noun) homonyms even before the end of the homonymous word. The cues which make this possible are the linguistic context (which primes the appropriate meaning), as well as frequency differences (which support the most commonly used meaning). As for noun/verb homonyms, it has been shown that with access to a large enough corpus of Mandarin sentences, a relatively simple distributional analysis using self-organizing networks is capable of distinguishing between the noun and verb forms of homonyms (Li 2002).

Table 17. Frequency of noun/verb homonymy in English and Mandarin, both for the 2,000 most frequent words in their lexica, as well as the entire lexica.

N/V HOMONYMY	English (46,133 words)		Mandarin (50,137 words)	
Words	Raw	%	Raw	%
2,000	601	30	506	25
All	2521	5	3432	7

These findings are obviously surprising given our conclusion that languages with pro-drop should suffer more from the presence of noun/verb homonymy than similar languages without pro-drop. However, a closer look at the data reveals that the results from the simulations are still relevant.

First, noun/verb homonymy being relatively common in Mandarin still leaves open the possibility that it can be more common in a language such as English, where word order can compensate for category ambiguities. To test this prediction empirically, we analyzed the word category information available in the CELEX database for English (Baayen, Piepenbrock and Gulikers 1995), and then obtained comparative numbers for Mandarin Chinese from Ping Li (personal communication). Table 17 contains the results, both for the 2,000 most frequent words in each language, as well as for all the word types in respective the databases (46,133 for English; 50137 for Mandarin). Perhaps contrary to popular belief, our counts show that English indeed features more noun/verb homonymy: for the 2,000 most frequent

Table 18. Frequency of nouns, verbs, and noun/verb homonyms in child-directed and adult English. The 968 word types analyzed were found in the Eve corpus (Brown 1973).

N/V HOMONYMY	CHILDES		CELEX	
	Word types (968)	Word tokens (18,528)	Word types (968)	Word tokens (3,531,205)
Noun	70%	40%	41%	14%
Verb	24%	53%	9%	45%
Noun/Verb	6%	7%	50%	41%

words, there is about 20% more noun/verb homonymy in English (601) than in Mandarin Chinese (506). This pattern is reversed when we look at the larger word sets, but with the more frequent words also being the ones that are generally learned first, it appears safe to assume that children learning English will face more noun/verb homonymy than children learning Mandarin. Children learning English can cope with greater lexical category ambiguity in the input because they can use a word's overt position in the sentence to help them determine whether it is being used as a noun or a verb.

Second, all the data we have so far reported is based on the adult language, which raises an interesting question about the difference in homonymy in adult vs child-directed speech. To explore this issue, we analyzed the Eve corpus from CHILDES (Brown 1973; MacWhinney 1995) and calculated type and token

frequencies for how often words occurred as a noun, a verb, or both.[26] We then looked up how the same words were coded in the CELEX database for English. The data for the 968 relevant words which occurred in both corpora are summarized in Table 18. In percentages, the Eve corpus from CHILDES has 6.1% homonymous types and 6.8% homonymous tokens. For the very same words, the CELEX database shows no less than 50.2% noun/verb type ambiguity and 41.5% token ambiguity! Obviously, the large difference in the absolute number of tokens between the two corpora means more research is needed, but the comparisons in Table 18 suggest strongly that there is considerably less noun/verb homonymy in child-directed speech than counts based on adult corpora might lead one to believe. As our simulations have shown, the level of noun/verb homonymy affects learnability when compensatory cues are scarce, so the presence of less ambiguity in the input could presumably benefit language learning children as well.

Whether the disparity in English between child-directed and adult corpora holds for other languages remains to be determined. Still, the fact that children learning Mandarin rarely use nouns as verbs (or vice versa) indicates that they are well aware of how each word is used in their input (see section 5.2.2 above; Erbaugh 1982). During the early stages of language acquisition, they also do not have access yet to the same kinds of pragmatic and contextual information that adults speakers of Mandarin

[26] We initially used the morphosyntactic tier that is already available for the Eve corpus. However, we found several coding mistakes (e.g. *wear* as a noun; *shoe* as a verb) so we manually checked all words where the verb usage or the noun usage was low to verify that these had been coded correctly.

use to disambiguate between homonyms. On the other hand, one difference between English and Mandarin which may work to the advantage of children learning Mandarin is that it appears more common for noun/verb homonyms in Mandarin to have completely unrelated meanings. As a result, even broad context knowledge would be sufficient to determine the right meaning – and thus also the category of homonymous word. In English, on the other hand, most noun/verb homonyms are semantically closely related with the verb typically expressing an activity which involves the noun (e.g. *jump, water, nap*). Finally, we need to point out that the amount of ambiguity in Mandarin is further reduced by the common use of compound words, in which each syllable may be a noun/verb homonym, but the combination of the two is not.

7.3 Summary

In Experiment 4, we investigated how well the different language types could cope with varying degrees of noun/verb homonymy in the input. We found that this task was remarkably easy (99+% correct) for all the language types which featured a reliable word and/or morphological markers. The 'difficult' language type, however, was impacted severely, with performance dropping steadily as the amount of homonymy increased. Additional exposure to the language helped, but the networks never reached the level of performance found in the other language types.

We also examined the frequencies of noun/verb homonyms in English and Mandarin Chinese, and found that English – where word order is a more reliable cue – has more such homonyms in the 2,000 most commonly used words than Mandarin. A

comparison between data based on adult and child-directed English corpora furthermore revealed that noun/verb homonymy is less common in the input directed to children. While comparable data is not yet available for Mandarin, other properties of the language appear to compensate for the ambiguities created by noun/verb homonyms. Consequently, the actual level of homonymy faced by children learning Mandarin is likely to be in the range (5-10%) that had a limited impact on the learnability of the difficult language type in our simulations.

Chapter 8. General Discussion

In this section we first take a closer look at the strengths and weaknesses of the different sources of information that were available to the networks. We then compare the results of our modeling experiments to the findings of several lines of research which share interests with ours. These comparisons allow us to highlight what is novel in our work by putting it in the context of related projects. We discuss the Competition Model by MacWhinney and Bates (e.g. MacWhinney 1987) as it has also investigated cross-linguistic differences in the acquisition of 'who did what to whom' – though largely without the use of computational models. We will also contrast our connectionist simulations with other probabilistic language learning methodologies such as statistical NLP and stochastic OT.

8.1 Word order vs. morphological marking vs. lexical identity

The generalization ability tested in the experiment with novel words (see Chapter 6) gives us an opportunity to compare how the different sources of linguistic information help in solving the 'who did what to whom' task. The results from the first two experiments were somewhat inconclusive because they only showed that lexical identity knowledge could not be learned as quickly as the other two sources. But in the absence of a good understanding of how many network training cycles correspond to a plausible amount of language exposure, the importance of this

difference in speed of acquisition is hard to gauge. Similarly, the finding that word order and morphology are equally useful also warrants a closer look, because it suggests that the presence of either one could remove any need for the other.

Let's begin with the problems experienced by networks that only had access to lexical identity knowledge, i.e. the networks learning the difficult language type. The generalization tests show clearly that lexical identity knowledge by itself is often insufficient for interpreting novel content words. These networks still got most of the sentences with novel nouns correct (80.2%), but their performance on sentences with novel verbs was very poor (52.8%). This result is not entirely surprising: when an unfamiliar word is encountered for the first time, it doesn't provide any lexical cues of its own as to how it should be interpreted. Instead, the listener must make inferences about the novel word by combining non-linguistic situational information with linguistic cues such as the novel word's position in the sentence, as well as the lexical properties of the surrounding known words. Under certain conditions, neighboring words can provide valuable clues about the properties of the novel word: e.g. a novel word following a transitive verb is likely to be a noun. However, until the novel word has been observed at least a few times, the listener had better be careful about inferring too much.

An important difference between our models and natural language use is that human beings are remarkably adept at learning novel words on the basis of limited input; this is the basis of the 'fast mapping' phenomenon (Carey and Bartlett 1978). In experimental settings, children as young as 15 months old could learn novel nouns

after very limited exposure (Schafer and Plunkett 1996; Houston-Price, Plunkett and Harris 2005); fast mapping of novel verbs starts somewhat later (e.g. Golinkoff, Jacquet et al. 1996). Because people can learn new words so quickly, their performance on a generalization task like the one we gave to the models would have been different. In the models, the weights on the connections were frozen for the generalization experiment, so they could not build up any knowledge about the behavior of the novel words. Every time they encountered a specific novel word, they had to interpret it as if they had never seen it before. Human subjects, on the other hand, could have noticed that certain novel words appeared consistently in certain contexts and their fast mapping ability would have allowed them to change their sentence interpretations accordingly. Consequently, human performance on the last part of the test corpora would likely have been much better than their performance on the first part. If we had implemented a similar ability in our simulations (see Milostan (1995) for a discussion of connectionist modeling and fast mapping), the difference between lexical identity knowledge and morphology or word order would have been less striking.

The second modeling result we want to discuss here is the result that word order and morphology (whether marked on the nouns or the verbs) appear to be interchangeable (compare Sapir 1921; Comrie 1989; Croft 2000). We have found that the presence of either one in the artificial language allows for absolutely perfect generalization to novel sentences containing familiar words (Experiments 1 and 2), and near-perfect generalization (97.3% - 99.9%) to novel sentences containing unfamiliar words (Experiment 3). We have also seen that combining a fixed word order with

morphological markers in a single language only provides limited benefit over the presence of just one of them. These findings indicate that both sources of information provide very similar kinds of cues to the models, and suggest – when taken at face value – that natural languages might also be able to survive with just word order or morphological markers.

The reason that both these types of information hold up so well in the generalization task is quite simple: unlike lexical identity knowledge, word order and morphology are 'robust' in the sense that they can appear with novel words in the very same way as they do with familiar words. A sentence with novel nouns and verbs can display the very same word order and morphological markers found in a sentence with known words. It is exactly because these patterns are general in nature that they generalize well to new data. The notion that a complete nonsense sentence could be easy to process might be counter-intuitive, but for an 'English' sentence like *the wugs dag the rits* most, if not all, speakers will agree that *the wugs* is the subject of the *dag*-ing action, and *the rits* the object – whatever these words may mean (Berko-Gleason 1958). The ability of our models to parse 'blindly' thus mimics a similar ability in human language processing.

Morphological markers and word order patterns are robust in another sense as well. They can be learned despite significant variation in the input data. For example, even in as fixed a word order language as English, careful analyses of natural language corpora invariably reveal the presence of sentences with different word orders. Similarly, there are no natural languages in which the case or agreement markers occur

as reliably as they do in our simulations, if only because the markers often take different morpho-phonological shapes depending on the grammatical class or number of the noun they attach to (case markers) or agree with (agreement markers). However, most of this variation appears to have little impact on how children learn the language. Given enough data, they always manage to pick up the more salient and frequent patterns (see section 5.2.1 above). These acquisition data have been modeled using relatively simple statistical methods that look for recurring patterns in the input (e.g. Cartwright and Brent 1997; Li, Burgess and Lund 2000; Mintz 2003). The success of these methods suggests that it should be possible to extend our current modeling setup to include less common constructions, without significantly affecting the results we have described.

However, there are also important differences between word order and morphology. Of the three sources of information we have looked at, word order is by far the most abstract. It cannot be associated with a salient phonological shape: i.e. a structural sequence such as SVO can be instantiated by many different combinations of words, and each word can also appear in different structures (see Akhtar and Tomasello 1997). Unlike morphological markers which typically take an easily recognizable phonological shape (e.g. English *–ing*), the closest counterpart in a word order pattern is its prosodic shape. These shapes are less reliable than markers and require the ability to pay attention to sequences of many words. Still, the presence of stress contours and pauses in the input data can be used quite successfully to identify

the kinds of units which form clauses and even phrases in a language (e.g. Fisher and Tokura 1996; Mandel, Kemler Nelson and Jusczyk 1996).[27]

The abstractness of word order also entails that it is always present – any sequence of words has an order – and this gives it the potential to become a sufficient source of information for solving the 'who did what to whom' task. All our models which learned language types with a reliable word order did extremely well, and the existence of natural languages with plenty of word order but poor morphology, such as English or Vietnamese, points in the same direction. However, our simulations have also exposed a major risk in relying exclusively on word order: if it is common in a language for subjects to remain unexpressed, word order alone can not be relied on to parse a sentence. This doesn't mean that the presence of pro-drop negates the value of word order, but it does lead to more possible structural sequences (compare SV, SVO to SV, SVO, V and VO), and thus also to potential ambiguities (e.g. NV vs. VN in a two-word sentence) which need to be resolved using another source of linguistic information. That is where morphological markers can come into the picture as a solution for the problems created by pro-drop. Whether the verb markers take the form of a rich agreement system or T/A/M morphemes is not as important in SVO languages, as either one suffices to distinguish the nouns from the verbs.

---

[27] In our models, all morphological markers had a unique and salient shape in the form of a specific pattern of 1's and 0'. Prosodic information was implemented in a simple way through the end-of-sentence symbol which followed the last word of each sentence, effectively telling the networks the sentence was over, just like a falling intonation pattern accompanies the end of most spoken declarative sentences.

In short, the relationship between morphological markers (or their closed class equivalents in isolating languages) and word order is more complex than a cursory reading of the network results might lead one to believe. Both encoding strategies are indeed quite capable of marking subjects and objects by themselves, but this result is only valid for fully reliable languages such as the ones we modeled. Natural languages have 'noisy' systems, complicated by the inconsistencies of optional/irregular marking or pro-drop. This is why both sources of information are present to some extent in all natural languages; they are often redundant in simple sentences, but can be the only cue present in more complex or incomplete sentences.

Finally, word order and morphological markers typically occur with words which are known to the hearers, so they can also draw upon their lexical category knowledge to provide them with useful cues about who did what to whom. The resulting picture is much more complex than the contrast of lexical identity vs. word order vs. morphology presented here. All three play a role, though their default importance differs in each language – an issue we return to in the next section. For example, word order has been found to be the most important cue in English, but it actually depends on morphology as well when lexical identity knowledge is unavailable. Compare how *the wugs dax the rits* makes considerably more sense (at least in terms of 'who did what to whom') than just *wug dax rit* which lacks the (noun identifying) article *the* and the plural marker *–s*. These sentences are obviously extreme in their use of unfamiliar words, but it is important to keep in mind that a young child learning language at times hears sequences of novel words as well. Any cue (or combination

thereof) which is available can make the difference between gibberish and being able to learn something about the words in the sentence. It is one of the strengths of our simulations that we can explore these interactions between the different sources of information under controlled experimental conditions.

8.2 The Competition Model

The Competition Model is a well-known framework for studying sentence processing strategies using cross-linguistic and developmental data (Bates, McNew et al. 1982; Bates and MacWhinney 1987; MacWhinney 1987; MacWhinney and Bates 1989; Li, Bates and MacWhinney 1993). Within this research program the main focus has been on how different languages express 'who did what to whom', and especially which linguistic cues (e.g. word order, agreement, stress, animacy) subjects rely on most when cues are in competition – i.e. they suggest different interpretations. A recent survey of these studies lists the results for 15 different languages, and no two languages use the same set of cues in the same order (Year 2003). It has also been found that these cross-linguistic differences in adult processing strategies are not always mirrored in the developmental data because more abstract cues (e.g. discourse) take longer to learn (Bates, MacWhinney et al. 1984; Döpke 1998; Thal and Flores 2001; Reyes 2003; Dick, Wulfeck et al. 2004).

More specifically, Competition Model studies show that the strength of a particular cue in a given language depends on its language-specific salience and availability. Cues that appear frequently, always have the same function, and are easier

to detect are thus the ones that are learned first and are strongest when in competition with other cues. In most languages, word order patterns or morphological markers on nouns or verbs fit this description and become the dominant cues. In the case of Mandarin Chinese, we have seen that these traditional cues are not sufficiently reliable. Unsurprisingly, Competition Model studies of the language (Li, Bates and MacWhinney 1993; Li 1998) have found that determining 'who did what to whom' in Mandarin can be a complex affair. However, the following cue strength hierarchy captures most of the experimental data: *bei* (passive marker) > animacy > word order > *ba* (object marker) > *yi* (indefiniteness marker). While this hierarchy can't be mapped directly onto our model of the corresponding language type, the finding that grammatical markers play a very important role in determining the subject and object of the sentence, even in a discourse-oriented language such as Mandarin Chinese, is entirely compatible with the results of our simulations.

There are obviously many similarities between the Competition Model studies and our own simulations. The potential for connectionist networks to add to the insights of the Competition Model has been recognized for a long time (see e.g. MacWhinney 1987, 2001, 2004; Hernandez, Li and MacWhinney 2005), but the only 'official' Competition Model implementation so far has been a comparison of the importance of morphological and semantic cues in the processing of German and Russian (Kempe and MacWhinney 1999). In this study, Kempe & MacWhinney found that a simple recurrent network could capture most of the human subjects data, both

with respect to the respective strength of the cues in both languages, as well as the reaction times for different kinds of sentences.

The success of Kempe & MacWhinney's model at modeling quite specific language data suggests an interesting avenue for further research. However, we believe there is more use for connectionist simulations that complement the Competition Model's core research rather than duplicate it. Due to the amount of labor involved in working with human subjects, Competition Model experiments are inherently limited to a (relatively) small number of languages and a (relatively) small number of linguistic cues to investigate. In addition, because they use human subjects, these studies can only investigate language types which are currently attested. Models such as ours are not restricted in these two areas and thus make it possible to include many more (possible) languages as well as (possible) cues in the experiments. For example, at the language level, the simulations we have described above looked at all combinations (rather than a subset) of several different kinds of morphological marking and their interaction with SVO word order and pro-drop. With respect to linguistic cues, we investigated the usefulness of T/A/M marking as a cue for determining 'who did what to whom', and determined that it could be just as effective as rich agreement marking.

In short, we suggest that computer simulations should be used to explore the space of possible languages and cues in a systematic fashion. When and where the data shows interesting differences or correlations, Competition Model experiments can be run to see whether they also appear in human subjects. The relative ease of setting up simulations compared to experiments with humans becomes especially important

when we expand the scope of research to include phenomena from language contact (e.g. multilingualism; second/third language acquisition; language loss) and aphasia. For the latter, models of varying amounts of damage to trained networks can give insight into 'cue robustness', while varying the location of the damage could conceivably mirror the symptoms of fluent and non-fluent aphasia (Bates, Ostrin et al. 1991; Packard 2006). For the former, models make it possible to exhaustively explore the interaction between all kinds of different language types to make predictions about which cues are likely to interfere or support each other in environments where more than one language must be learned or maintained (Döpke 1998; Su 2001; Hernandez, Li and MacWhinney 2005).

Finally, there is another reason why Competition Model experiments and (connectionist) simulations are best seen as complementary research – i.e. there are processing cues such as animacy, discourse information and sentence intonation which play a crucial role in some natural languages but which have so far resisted computational implementation. A language like Italian in which animacy and stress play far greater roles than word order is thus hard to capture well in models like ours that have no lexical semantics. More generally, we haven seen that lexical knowledge in the models is both slower to learn and less useful for generalization purposes. However, its importance for human language processing does not need much motivation. The primary function of lexical knowledge in natural languages is not to encode 'who did what to whom' but to express meaning (*pace* Chomsky). Where the models are learning meaningless strings of symbols such as *X is Ying Z*, the child is

learning something about the world such as *Sandy is preparing the food*. Note that the former is essentially interchangeable with *A is Bing C* but the latter is quite different from *Morgan is drawing a picture*. At least until we have models which can ground their symbols in sensory-motor data from the real world, running experiments with human subjects will reveal phenomena which these simulations cannot capture.

8.3 Probabilistic linguistics

In recent years, linguists have shown increasing interest in the use of probabilistic models to support linguistic analyses in many of its subfields (e.g. Seidenberg and MacDonald 1999; Boersma and Hayes 2001; Bybee and Hopper 2001; Jurafsky, Bell et al. 2001; Bod, Hay and Jannedy 2003; Polinsky and Van Everbroeck 2003; Van Everbroeck 2003; Chater and Manning 2006; Crocker and Keller 2006; Bresnan, Cueni et al. 2007). The common theme in these different lines of research is that observed frequency differences between linguistic patterns correlate with linguistic generalizations over the same patterns. What's more, it has also been demonstrated that these generalizations can often be learned by computational models that pay attention to the frequency data. The growing awareness among linguists of the importance of probabilistic models stands in stark contrast to the original tenets of formal linguistics, where neither frequency data nor learning were considered important (e.g. Chomsky 1988; Piatelli-Palmarini 1989), but it is bringing the field closer to well established findings in cognitive linguistics (e.g. cognitive grammar has long been described as 'usage based' – Langacker 1987; Barlow and Kemmer 2000)

and cognitive science (e.g. how infants, toddlers and adults learn both natural and artificial languages – Saffran, Aslin and Newport 1996; Seidenberg, MacDonald and Saffran 2002; Saffran and Wilson 2003; Gerken, Wilson and Lewis 2005; Hudson Kam and Newport 2005; Matthews, Lieven et al. 2005).

The connectionist models that we have described above are quite compatible with the overall trend towards more probabilistic data. Neural networks like ours learn through detecting regularities in the input and adjusting their weights to classify them appropriately (Rumelhart, Hinton and Williams 1986). However, most of the recent work by linguists uses more straightforward statistical approaches (see below), possibly because there are still concerns about the validity of connectionist modeling for the study of natural language.

One general criticism applies to the small sizes of the grammars and lexica used in most connectionist models – i.e. the networks are only asked to process 'toy languages' (Tepper, Powell and Palmer-Brown 2002; Newmeyer 2003). It indeed remains to be shown that connectionist networks can handle the full complexity of an adult lexicon with tens of thousands of words, but our simulations have demonstrated that learning several hundred words (300 nouns; 100 verbs) is definitely feasible, even in a small network which lacks the neural capacity of a 1-year-old infant.

A more specific criticism about neural nets is the claim by Marcus, Vijayan et al. (1999) that "[in connectionist models], there is no generalization to novel words. Such networks can simulate knowledge of grammatical rules only by being trained on all items to which they apply; consequently, such mechanisms cannot account for how

humans generalize rules to new items that do not overlap with the items that appeared in training" (79). This claim was addressed by Elman (1998), who demonstrated that neural networks are capable of generalizing a known word to a novel position in the sentence – e.g. even if a network has never seen the word *dog* in subject position, it may still predict its possible occurrence there on the basis of its appearance in object position as well as the behavior of other nouns like *dog*. The generalization abilities which we have described here are considerably more powerful, because we tested our networks on words which they had not seen in any position before and which were thus completely novel – more akin to testing a young child on an uncommon word such as *doberman* rather than *dog*. We have seen above that even sentences containing nothing but novel words can be processed correctly in all language types except for the 'difficult' one.

Finally, our results also contradict the recent conclusion by van der Velde, van der Voort van der Kleij and de Kamps (2004) that recurrent networks are incapable of determining 'who did what to whom' in Noun Verb Noun sentences. As they put it: "Consider, for instance, the sentences *cat chases mouse* and *mouse chases cat*. Both sentences are *N V N* sentences and they are thus indistinguishable for the S[imple] R[ecurrent] N[etworks]s" (42). Van der Velde et al. base this conclusion on the failure of their own model to solve the task at hand. We suggest here that they have severely underestimated the space of possible connectionist models as our simulations demonstrate quite clearly that networks can be trained to construct distinct 'who did what to whom' representations for similar sentences.

As mentioned already, most of the recent work on probabilistic linguistics does not use connectionist networks. Instead, we find models which are grounded in the precise frequencies with which the relevant linguistic items occur in natural language corpora (e.g. Manning and Schütze 1999; Bod, Hay and Jannedy 2003). For example, in a typical Bayesian model, one might count exactly how often the word *said* is followed in the corpus by a sentential complement versus how often it is followed by a noun phrase. After scanning the entire corpus, we can then use the calculated probabilities for determining the likeliest interpretation for novel sentences that contain *said*. Compared to connectionist networks, these frequency based 'statistical' models have numerous advantages.

First, they can be trained faster because there is no need to experiment with architectures or to calculate weight updates for (tens of) thousands of connections between units. Second, their behavior is more predictable because the learning algorithms in multi-layer neural networks are not guaranteed to find a good (let alone the best) solution. Third, statistical models can easily work with the kinds of abstract concepts which are used by linguists – e.g. 'sentential complement', 'relative clause' or 'verb phrase'. Such high-level concepts have historically been avoided in connectionist models in favor of letting the networks develop their own complex representations in the hidden layers. As a result, it is far easier to interpret the output of statistical models than it is to make sense of what a neural network is doing internally. For all these reasons, when the goal is to create a working text analysis system with the best possible performance, statistical models are far more likely to produce a useful product than

neural networks and are thus used for the analysis of (very) large natural language corpora.

Despite growing interest in the use of Bayesian learning methods for modeling developmental data (e.g. Narayanan and Jurafsky 2002; Goldwater 2007; Gopnik and Tenenbaum 2007; Xu and Tenenbaum 2007), the long-term cognitive plausibility of these successful statistical models remains to be determined. Not only is it unlikely that young children learning their native language(s) would be keeping track of exact frequencies and manipulating these according to Bayes' rule, it has also been demonstrated that early language acquisition involves a certain disregard for observed frequencies in favor of generalizing common language patterns to inconsistent forms (e.g. Hudson Kam and Newport 2005; Zhu and Gigerenzer 2006). The cognitive implementation of e.g. Bayesian learning is unclear at best, leading to the suggestion that statistical models may have to be translated into connectionist ones for cognitive research (Jurafsky 2001; Shultz 2007).

Connectionist modeling, on the other hand, has long been interested in cognitive language phenomena, both relating to developmental profiles and online processing (Rumelhart and McClelland 1986; Elman 1993; Plunkett and Marchman 1996; Christiansen and Chater 1999; Seidenberg and MacDonald 1999; Munakata and McClelland 2003; Cangelosi 2005; Elman 2005). The advantages of statistical models just mentioned are best thought of as relative to the engineering goal of best performance. Language learning in children is a slow process, taking many years and exposure to millions of words before an approximation of the adult language is

mastered. There is also noticeable variation between subjects in both acquisition as well as usage. Such differences are quite naturally expressed in neural network terms: e.g. the number of units in each layer as well as the connectivity patterns between layers play an important role in learning, as do the initial weights and the parameters for updating them. Finally, the use of high-level linguistic concepts to describe language acquisition often fails to do justice to the very specific categories which are often observed in the very early stages of language learning (Tomasello 2000; Gerken, Wilson and Lewis 2005; Naigles, Bavin and Smith 2005; Uziel-Karl 2006).

In the experiments described earlier, we have also been far more concerned with their cognitive implications – i.e. how the results compare to what is known about human languages – than with the networks' level of performance. For example, rather than optimizing the learning rate, we have compared the networks' behavior to what is known about the acquisition of Mandarin Chinese, and to how children learn homonymous words. Another productive area of study concerns the robustness of different ways of encoding 'who did what to whom', both in terms of acquisition (how many network 'subjects' learn the language without problems?) and catastrophic language loss (how do varying amounts of damage to the trained connections affect performance?; Van Everbroeck (in preparation)). Neither question can easily be asked from the standpoint of a more formal statistical model.

Finally, it is worth pointing out that a statistical implementation of our simulations is neither trivial to set up nor likely to perform much better. With respect to the latter, the networks typically managed to get at least 99% of the test sentences

correct with limited training, leaving very little room for significant improvement by another class of learning mechanism. As for the former, how to do the re-implementation using regular probabilities is not entirely obvious, given that we also tested our networks on sentences which contained nothing but novel words. Such words would not have had any frequencies associated with them, essentially short-circuiting the formulas which depend on them. One possible solution might be to ignore the lexical identities of the words and to calculate the probabilities of specific sentence structures (e.g. SV or SVO) on the basis of the number of words in a sentence and the presence of any markers. However, such an implementation could not be used to explore the issue of noun/verb homonymy, as it crucially depends on the noun and the verb having the same lexical form.

In short, both the Bayesian and the connectionist approach have a lot to offer and it is crucial to use them differentially depending on the nature of the research task (compare Chater and Manning 2006; Shultz 2007).

8.4 Optimality Theory

Optimality Theory (OT), first introduced by Prince and Smolensky (1993), has been applied mainly to phonology. Its core apparatus consists of two elements: *Gen*, which generates a wide range of possible phonological output forms for a given underlying input form; and *Eval*, which uses a set of constraints on the relationship between the input and the output forms to determine which output form is optimal in a language. Crucially, the set of constraints used by *Eval* is assumed to be strictly

ranked in a hierarchy, and the optimal form can only violate lower-ranked constraints than the other candidate forms. These constraint rankings differ between languages and this in turn accounts for the fact that for example particular stress patterns or syllable shapes are well-formed in one language but not acceptable in another. What makes OT potentially relevant for our experiments is the hypothesis that such constraint rankings can be learned automatically. However, a brief survey of this work will show that it is not feasible to recast our experiments in the OT framework.

Outside phonology, the constraint ranking mechanisms of OT have been applied to a variety of non-phonological phenomena, including word order (Costa 2001; Samek-Lodovici 2001; Flack 2005), case marking (Müller 2001; Aissen 2003), agreement systems (Morimoto 2002; Samek-Lodovici 2003), as well as pro-drop (Bresnan 2001; Speas 2001). In general, these analyses account for the cross-linguistic differences between languages by combining a few typical OT constraints – to keep the output form looking at least somewhat like the input form – with a number of specific constraints related to the area under investigation: e.g. SUBJECT (to require expressed subjects), TOPICFIRST (to make topics sentence-initial), or NOFEATURES (to prevent overt agreement markers). While such analyses work from a descriptive point of view, we feel they are explanatory deficient in two important areas. First, some of the constraints lack independent motivation and become mere labels for different language types – e.g. Samek-Lodovici (2003) accounts for three types of agreement phenomena by using three constraints, each of which neatly defines the behavior of a type (and is violated by the other two types). Second, the set of available constraints is

itself unconstrained, leaving it an open question whether – and how – all of them could interact (compare Newmeyer 2005). At least with regards to 'who did what to whom', there is good reason to believe that the constraints on word order, case, agreement and pro-drop should all affect one another. OT Syntax, at least in its current state, is simply not sufficiently developed to address questions of this complexity.

One promising line of OT research investigates how algorithms can automatically determine the correct ranking for a set of constraints on the basis of a corpus of input and output forms. For example, it has been shown that the Constraint Demotion model of Tesar (2004) is guaranteed to converge on the correct ranking. However, its cognitive plausibility suffers because it cannot deal well with noise in the corpus, such as the occasional incorrect output form. The Gradual Learning Algorithm of Boersma and Hayes (2001) is more promising in this regard. It's a stochastic model in which the constraints are ranked by their (learned) weights, rather than by strict dominance, allowing the model both to overcome some noise in the input, and to account for language variation phenomena as well as gradient grammaticality judgments. However, it is still unclear whether even the Gradual Learning Algorithm can scale to more complex language data (see also Keller 2000; Kuhn 2003). First, it has so far only been used to account for phonological phenomena and it remains to be seen how it will scale to syntactic, semantic and pragmatic questions. For example, lexically specified information plays only a limited role in phonology beyond the shape of the word involved, but the same word will likely have a much richer representation

for its syntactic and semantic properties. Second, the literature on learning in OT still hasn't connected to what is known about language acquisition in children. It would obviously be desirable for the constraint ranking process of the Gradual Learning Algorithm to mirror the human developmental data for the same language items.

For the sake of the argument, however, let's assume that we know which constraints to use for the 'who did what to whom' question, and also that we have a learning mechanism for automatically ranking them. Could we now rephrase our experiments in an OT fashion? We still think the answer is no. One major obstacle is that language production is the strength of OT, while our models perform a comprehension task by going from a sequence of words to a representation of the entire sentence. The OT version of this task would be to have the representation of the sentence as the underlying form and then have *Gen* produce a number of possible output sentences which are checked by *Eval* against the ranked list of constraints for the relevant language type (compare Gibson and Broihier 1998). We can then verify whether the winning output sentence is indeed grammatical for the language type. If it is not, the ranking of the constraints would be updated by the (hypothetical) learning algorithm until the correct sentence is selected for every underlying form. There are several problems with this scenario. For example, it is not clear how we would investigate the effect of novel words or noun/verb homonymy in a model like this. Also, learnability has not played a role in OT's view of cross-linguistic variation. Instead, the space of possible languages has been defined as a 'factorial typology', i.e. all possible rankings of all constraints. The goal of OT has been to find rankings that

allow each and every attested language, while simultaneously excluding the unattested ones. As a result, the theory does not distinguish between plausible versus implausible unattested languages, while exploring this distinction has exactly been the goal of our experiments. It is doubtful that the model would have any more problems learning the constraint rankings for the difficult language type – i.e. with pro-drop, but without nominal or verbal marking – than for any of the other types. From a production point of view, not having to express markers only makes things simpler. Pro-drop would not be problematic either, because we know there are constraints available to make it possible in other language types. In short, we would not have a good explanation for why the difficult type is harder to learn.

Chapter 9. Conclusion

In this last section we first give summaries of the four experiments which we have presented. We then return to the 'big picture' issues we first mentioned in the introduction to assess what our simulations have contributed to them. To wrap up, we describe various options for extending the current models to address related cross-linguistic and acquisition phenomena.

9.1 Summary of the experiments

9.1.1 Experiment 1

The goal of Experiment 1 was to establish a baseline for the impact of pro-drop on SVO language types with varying amounts of morphological marking. The marking parameters were case-marking on nouns and two kinds of head-marking on verbs (Tense/Aspect/Modality, and rich agreement). The main result of was that pro-drop only makes a language type harder to learn when there is no morphological information available – the type which combined pro-drop with no marking only got 73.4% of the test corpus correct, whereas the worst score of all the other types was 98.7%. This finding suggests that the presence of pro-drop need not correlate with the presence of a rich agreement system. Instead, any reliable and consistent strategy for

marking the nouns or verbs in a language can be sufficient to determine 'who did what to whom'.

From a cross-linguistic perspective, the results of Experiment 1 suggest that the types which the models could learn may be attested in the real world, while the difficult type should be unattested. Our survey of the world's SVO languages found two interesting deviations from these predictions. First, we didn't find any attested SVO language which has nominal case-marking but no verbal agreement system. Second, we found two groups of languages which have been described as having the features of the type which the networks found hard to learn. However, a closer look at the world's creole languages revealed that they don't have any interesting referential pro-drop without agreement. Our investigation of Mandarin Chinese, the best studied example of the South-East Asian Sprachbund, revealed that it has wide-spread pro-drop but also a number of grammatical cues for identifying both nouns and verbs.

9.1.2 Experiment 2

In Experiment 2, we explored whether additional training epochs would make the neural networks perform better on the difficult language type. The hypothesis was that better lexical knowledge of the words in the language would allow the networks to decide whether a word in the input was a noun or a verb, even if it occurred as a bare stem without morphological cues. After 30 training epochs (vs 10 in Experiment 1), performance on the difficult type had increased to 91.4%, with all the other types at

99.9%. Additional training on the difficult type showed decreasing improvements with performance on the test corpus not surpassing 97.8% even after 200 training epochs.

These results are compatible with two findings from child language acquisition. First, the acquisition data for Mandarin Chinese show that children can learn early on exactly how each word in a language must be used. In Mandarin, the presence of pro-drop entails that the structural position a word appears in is not a reliable cue and thus cannot be used to decide whether this word is a noun or a verb. Second, the developmental data also demonstrate that children are sensitive to reliable morphological (or other grammatical) cues from a very early age. The acquisition of 'who did what to whom' in languages with consistent morphology takes place considerably faster than in languages where such cues are not available, or not reliable.

9.1.3 Experiment 3

The goal of Experiment 3 was to determine how our networks generalized their parsing strategies to sentences containing novel nouns and/or verbs. Success in this task is a requirement for cognitive models because there is plenty of experimental evidence for (over)generalization in children. For example, children can deduce a novel word's lexical category on the basis of its position in a sentence as well as any morphological markers. To test our models, we presented test corpora which contained all novel nouns, all novel verbs, or both novel nouns and verbs. We found that the presence of novel content words only created problems for the networks learning the difficult type; all the other networks scored more than 97% correct under

every condition. We also found that the presence of novel verbs (52.8% correct) had a much bigger impact on the difficult type than novel nouns (80.2%).

The results of Experiment 3 touch on several issues in language acquisition. First, they suggest that the distinction between transitive and intransitive verbs can be learned without access to rich conceptual or syntactic knowledge. Second, generalization of lexical categories can be done equally well using word order or morphological markers, but the difference in abstractness between the two means they complement each other rather than overlap. Finally, the models provide experimental support for the contentious hypothesis that verbs play a more important role in languages with pro-drop than in those without.

9.1.4 Experiment 4

In Experiment 4, we investigated the impact of Noun/Verb homonymy on the learnability of our languages. Such homonyms are attested in many languages without (much) overt morphology and they potentially make it harder for a child (or a neural network) to determine the correct range of use of the homonymous words. For our simulations, we tested increasing levels of N/V homonymy (5% to 50%) with various amounts of training (10 to 100 epochs). We found that only the networks learning the difficult type experienced problems; all the others got at least 98% of the test sentences correct. With the difficult type, however, even low levels of homonymy had a serious impact on the learnability of the language. Additional training cycles helped,

but reasonable performance (90%) could only be reached with at most 10% N/V homonymy.

When we examined the frequency of N/V homonymy in English and Mandarin Chinese, we found that the two languages have very similar numbers of N/V homonyms, both in their most frequent words and in their larger vocabularies. Moreover, a comparison of adult and child-directed varieties of English showed that N/V homonymy appears to be considerably less common in child-directed speech. If the same pattern also holds for Mandarin, the degree of homonymy which children learning the language have to cope with is likely to be no more than the 10% learnability limit we observed in the models.

9.2 The big picture

We have presented four experiments which used connectionist modeling to study the effect of pro-drop on SVO languages with varying amounts of morphological marking. We wanted to find out whether the traditional linguistic generalizations about pro-drop would be borne out – i.e. can pro-drop only occur when a language features either rich subject-verb agreement, or very little morphology at all (Huang 1984, 1989; Jaeggli and Safir 1989; Nicolis 2005)? In our experiments, we tested how the presence of pro-drop affected each of the simulated language types by determining whether a type with pro-drop was significantly harder to learn than the otherwise identical type without pro-drop. Learnability was defined in terms of how well the models performed on solving 'who did what to whom' for sentences of the

relevant language type. Our results show that pro-drop in the models had two quite distinct faces.

On the one hand, pro-drop barely had any effect at all on language type learnability as long as there was some reliable morphological marking available in the language (Experiment 1). Interestingly, our experiments suggest that for SVO languages the kind of morphological markers present was of relatively little importance. Both nominal case markers and verbal Tense/Aspect/Modality markers were equally effective in compensating for pro-drop as rich agreement markers. In SVO languages, pro-drop led to potential ambiguities between nouns and verbs, but not to uncertainty about whether a noun was the subject or the object of the sentence. Consequently, any cues that reliably distinguished the nouns from the verbs were sufficient to make an SVO language with pro-drop learnable. This was true even when the sentences of the language contained many novel words (Experiment 3) or words that were Noun/Verb homonyms (Experiment 4).

On the other hand, pro-drop had a major impact on language learnability when no reliable morphological markers were present in the language type (Experiment 1). Prolonged exposure to such a language improved performance on the 'who did what to whom' task, but it never reached the level of the other types without pro-drop (Experiment 2). Moreover, when tested on sentences containing novel words (especially verbs; Experiment 3) or N/V homonyms (Experiment 4), the models trying to learn this 'difficult' type with pro-drop but without morphological markers

experienced severe difficulties. These problems were caused by the inability to distinguish between nouns and verbs in the language.

These network results are obviously at odds with the two linguistic generalizations we mentioned about where pro-drop can occur. How should we interpret the discrepancies? First, the modeling result that any kind of regular morphological marking – not just rich agreement – can suffice to determine 'who did what to whom' is somewhat misleading. In the experiments with pro-drop, the networks did not have to provide a representation at the output layer for the unexpressed subject. Consequently, any unexpressed subject was referentially identical to every other one. While this makes modeling sense (in the absence of an implementation of discourse processing), it does not do justice to the conceptual representations built while processing natural languages. When a speaker produces a pro-drop sentence, she typically knows quite well who the unexpressed subject refers to. The hearer similarly is expected to figure out who the unexpressed subject is and he can typically do so without any problems. While context (both linguistic and non-linguistic) can provide numerous clues to help with this, the presence of rich verb agreement in a language can really help a hearer narrow down the set of potential referents by excluding all the options that don't match the features contained in the agreement markers. In contrast, case markers on nouns and/or Tense/Aspect/Modality markers on verbs don't carry any referential information about the unexpressed subject. We have no doubt that if we had asked our networks to produce a likely output representation for the subject in pro-drop sentences, the

language types with agreement would also have been able to do this task in a way the types without it could not have.

Still, the correlation between languages with pro-drop and those with rich verb agreement should be considered an important cross-linguistic observation, but not one that should be required by linguistic theories. Even ignoring the referential advantage granted by agreement markers, the preference for agreement over case-marking in SVO languages with pro-drop can also be explained from an acquisition perspective. As we have discussed earlier (see section 6.2.3), verbs in languages with pro-drop are relatively more salient because they are never left unexpressed. As a result, they also become relatively easier to acquire. Moreover, large typological samples show that agreement is simply more common in SVO languages than case-marking. In the database of Siewierska (1996), agreement is present in 61% of the SVO languages, versus 32% for case-marking on nouns and 47% on pronouns. (Interestingly, she also found case-marking to be less common in SVO languages than in those with other basic word orders.) The World Atlas of Language Structures database (Haspelmath, Dryer et al. 2005) similarly reports more VO languages with agreement (246) than without (66). In the latter group, pro-drop is absent (43 languages) more often than present (23). This last number is important because it suggests there may be tens of SVO languages with pro-drop but no rich agreement. As our experiments have shown, such languages are still learnable as long as some other type of reliable morphological marking is available. A detailed linguistic analysis of the relevant languages in the

WALS database is called for, but our modeling results appear to be a better match for the typological data than a linguistic rule requiring rich agreement with pro-drop.

This brings us to the second discrepancy between linguistic theory and our modeling results. Some linguists have argued, typically on the basis of data from Mandarin Chinese (e.g. Huang 1984, 1989), that pro-drop is possible when there is no morphology in the language at all. All of our experiments suggest strongly that this should not be the case. The language type with pro-drop but without morphological marking was consistently harder to learn than the other types. The fundamental problem with this type was that there was often insufficient information available to decide what lexical category a novel word belonged to. This uncertainty created structural ambiguity about the correct interpretation of the sentence, and led to processing mistakes by the networks.

This modeling result may appear undesirable at first, because human language processors are remarkably adept at dealing with ambiguities in their input. Still, there is plenty of research demonstrating that structural ambiguities can be problematic for people as well, and cross-linguistic work also suggests that such structures are generally dispreferred (Hawkins 2004; Levy 2006). Moreover, ambiguity resolution is a process which adults are much better at than children, precisely because it depends upon using information which is not available in the immediate linguistic context. Similarly, the ability to recover from a garden path in processing is not found in young children. On the contrary, once the initial interpretation stops making sense for the sentence they're parsing, they tend to give up on processing the sentence altogether (Trueswell,

Sekerina et al. 1999; Traxler 2002; Felser, Marinis and Clahsen 2003). This behavior is qualitatively comparable to how the networks performed on ambiguous sentences.

With respect to Mandarin Chinese, we have argued that while the language is indeed impoverished from a morphological perspective, there are still numerous cues available to help identify nouns and verbs. For example, one can use classifiers or the object preceding *ba* form for predicting nouns and aspect markers and auxiliaries for predicting verbs. Crucially, the acquisition data for Mandarin shows that children learning the language quickly start paying attention to these cues (compare Tardif 2006). Unlike the adult language which is characterized by a great deal of structural variation, the sentences produced by children follow far more rigid word order patterns. Finally, we pointed out that children learning Mandarin hardly ever use nouns as verbs (or vice versa), whereas this is a very common occurrence in children learning English. The absence of this phenomenon in Mandarin is exactly what the results of the models lead us to expect: unlike English, Mandarin has pro-drop, so there is far more potential for confusion between nouns and verbs; avoiding this confusion prevents structural ambiguities and processing errors.

More generally, we suggest that natural languages also avoid combining pro-drop with a complete lack of morphology. Isolating languages with pro-drop such as Mandarin and Thai probably get as close to this type as one can get while still being learnable by the relatively limited processing abilities of young children. When it comes to understanding 'who did what to whom', language learners will acquire the relevant grammatical elements that are available to them in their input. In most languages, these

elements are relatively conspicuous morphological markers on the nouns or the verbs. In languages which lack abundant morphology, other cues will be used even if their primary meaning is quite different. We believe that studying the early acquisition of syntactic structures can reveal processing strategies which are not what the adult language (or its grammars) would lead one to expect.

In short, modeling experiments like ours can be a useful heuristic for identifying both language types and linguistic strategies which require further investigation.

9.3 Future work

The models we have presented in this dissertation draw upon findings from various areas of linguistics and cognitive science. In this section, we briefly sketch how one could expand on the current work in different ways to investigate particular issues in greater depth.

Let us first consider the cross-linguistic perspective. The obvious place to start is to flesh out the table in which we map modeling types onto natural languages. In particular, it would be desirable to find natural languages which combine an SVO word order with both unexpressed subjects and case-marking, but no rich agreement. As mentioned earlier, the WALS database contains some likely candidates, and it has also been suggested to us that Ambonese Malay may well substantiate this type (Mark Donohue, personal communication). On the other hand, we should also continue looking for counter-examples to the models' finding that there is an unlearnable type.

In this regard, two languages which deserve further study are Riau Indonesian and Singapore English. The former is a lingua franca spoken in parts of Sumatra . It may be a creole (McWhorter 2001), but it has also been described as a regular – i.e. not pidgin based – language which has evolved creole features (Gil 1994). As described by Gil (1994, personal communication.), Riau Indonesian combines a basic SVO word order with considerable pro-drop, very limited inflectional morphology and widespread zero-conversion between word classes – supposedly to the extent that any imaginable sequence of words constitutes a valid sentence in the language! As one may recall from section 4.2.1, the situation with Singapore English is fairly similar. It is a colloquial register spoken in Singapore which combines linguistic structures of several nearby Chinese dialects as well as Malay and Tamil (all languages with pro-drop) and then adds English words (Zhiming 2001, 2005). The resulting mixture can be very elliptic and thus hard to make sense of when taken out of the original discourse context (Mark Donohue, p.c.). If these descriptions are correct, Riau Indonesian and Singapore English would likely be unlearnable by our models. However, it is not obvious to us that we are dealing here with fully fledged languages – i.e. due to the amount of language contact in this area of the world, neither is ever learned as the only language; they are not used in more formal contexts; and there appears to be more variation between speakers than what would typically expect to find in a regular language. Still, it will be extremely interesting to look at acquisition data for these languages when it becomes available.

Another typological question is how our models fare on languages with other basic word orders. There are certain observed cross-linguistic patterns which one would hope to find in the simulations as well, such as the combination of SOV word order with case-marking, or the presence of rich agreement in verb-initial languages. We have already run our model on all six basic word orders; it is simply a matter of analyzing the computational results and comparing them to the cross-linguistic data. It would not surprise us if each word order requires the kind of analysis we have had to give to the SVO languages here.

A second area in which there is need for further study is language acquisition. We have mentioned relevant data in our discussions of how children learn Mandarin, whether they acquire nouns or verbs first, and how they cope with homonyms. But there are several important questions which we haven't touched upon. For example, does the developmental profile of the models match what is observed in children, and are there similarities in the order in which different constructions and types of sentences are learned? A similar issue to examine is how the acquisition of pro-drop varies with language features. Pro-drop is a phenomenon which has been studied quite extensively in several languages (see e.g. Kim (2000) for a summary of pro-drop acquisition in seven languages), so there is reasonable amount of human data available to evaluate the model. Finally, another avenue would be to test the model's prediction that novel words occurring at the beginning of a sentence will impair comprehension more in Mandarin than in English.

The third major area where our simulations need to be extended is the linguistic realism of the model. There are the standard options for neural networks such as the use of other architectures and learning methods, but the main interest should be in improving the linguistic plausibility of the models. Promising avenues which we have mentioned in various places are the addition of lexical semantics to help disambiguate words (Waskan 2001); the use of more linguistically complex structures such as the relative clauses and locative phrases we included in earlier work (Van Everbroeck 2003); and changing the corpora so there is a more realistic frequency distribution for how often each noun and verb is used in a sentence, as well as how often particular nouns and verbs are used together. Making the models more complex along these lines might make it possible to have the output representation be more conceptual in nature, so the words in input sentences are mapped onto meaningful semantic roles rather than generic banks of output units (Morris, Cottrell and Elman 2000). The major concern about implementing most of these changes in our models is that they will increase the overall complexity and create a combinatorial explosion of possible language types. Modeling and analyzing all possible types thus becomes unwieldy, if not impossible. It will likely be necessary to simplify the models in some areas before increasing their complexity in others.

A final area in which we think simulations like ours could be useful is the systematic study of multilingualism, and, more generally, language contact situations. It is very common for children to grow up in an environment in which more than one

language is spoken, but systematic studies of how each of the languages is acquired are quite rare due to the difficulties in finding multiple subjects with similar amounts of exposure to the various languages. With computational models, it is relatively simple to train them on multiple languages and study how each is learned. What's more, the models also enable us to look at combinations of language types which may not occur in contact situations in the real world. In all these cases, simulations may be able to provide us with valuable data on how the structures of different languages interact during acquisition – e.g. which patterns overlap and are learned quickly, and which ones are problematic because they conflict with others. The results of such experiments could be used to guide studies that look at the acquisition of specific structures in children. With respect to second language acquisition, we can see how detailed analyses of problematic structures would inform language learning textbooks, or at least make it easier to predict which structures are more problematic depending on the native language.

By raising all these questions, we have shown that our models tie together data from various fields in cognitive science. Whether ultimately connectionist in nature or not, computational models that look at the acquisition of 'who did what to whom' strategies in different language types should continue to deliver interesting results.

References

Abbot-Smith, K., Lieven, E. V. M., & Tomasello, M. (2001). What preschool children do and do not do with ungrammatical word orders. *Cognitive Development, 16*, 679-692.

Abbot-Smith, K., Lieven, E. V. M., & Tomasello, M. (2004). Training 2;6-year-olds to produce the transitive construction: The role of frequency, semantic similarity and shared syntactic distribution. *Developmental Science, 7*(1), 48-55.

Aissen, J. (2003). Differential object marking: Iconicity vs. economy. *Natural Language & Linguistics Theory, 21*, 435-483.

Akhtar, N., & Tomasello, M. (1997). Young children's productivity with word order and verb morphology. *Developmental Psychology, 33*(6), 952-965.

Anderson, K. L., & Cottrell, G. W. (2001). Age of Acquisition in connectionist networks. In *Proceedings of the 23rd Annual Cognitive Science Conference* (pp. 27-32). Mahwah, NJ: Lawrence Erlbaum.

Armon-Lotem, S., & Berman, R. A. (2003). The emergence of grammar: Early verbs and beyond. *Journal of Child Language, 30*, 845-877.

Aronoff, J. M. (2003). Null subjects in child language: Evidence for a performance account. In G. Garding & M. Tsujimura (Eds.), *WCCFL 22 Proceedings* (pp. 1-14). Somerville, MA: Cascadilla Press.

Aroonmanakun, W. (1999). *Extending focusing for zero pronoun resolution in Thai.* Unpublished Doctoral dissertation, Georgetown University.

Aroonmanakun, W. (2000). Zero pronoun resolution in Thai: A Centering approach. In D. Burnham (Ed.), *Interdisciplinary approaches to language processing: The international conference on human and machine processing of language and speech* (pp. 127-147). Bangkok: NECTEC.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9*(4), 321-324.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1999). Statistical learning in linguistic and nonlinguistic domains. In B. MacWhinney (Ed.), *The Emergence of Language* (pp. 359-380). Hillsdale, NJ: Lawrence Erlbaum.

Baayen, H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database. Release 2 (CD-ROM)*. Philadelphia, PA: University of Pennsylvania.

Baillargeon, R. (1987). Young infants' reasoning about the physical and spatial properties of a hidden object. *Cognitive Development, 2*, 179-200.

Baillargeon, R. (1998). Infants' understanding of the physical world. In M. Sabourin, F. Craik & M. Robert (Eds.), *Advances in Psychological Science. Volume 2* (pp. 503-529). East Sussex: Psychology Press.

Baillargeon, R. (2004). Infants' reasoning about hidden objects: Evidence for event-general and event-specific expectations. *Developmental Science, 7*(4), 391-424.

Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition, 20*, 191-208.

Baker, M. C. (2003). Linguistic differences and language design. *TRENDS in Cognitive Sciences, 7*(8), 349-353.

Baptista, M. (1995). On the nature of Pro-drop in Capeverdean Creole. *Harvard Working Papers in Linguistics, 5*, 3-17.

Barbosa, P., & Torres Morais, M. A. (2001). Review of *Brazilian Portuguese and the Null Subject Parameter*. *Probus, 13*, 277-284.

Barlow, M., & Kemmer, S. (Eds.). (2000). *Usage Based Models of Language*. Stanford, CA: CSLI Publications.

Bates, E., Chen, S., Tzeng, O. J. L., Li, P., & Opie, M. (1991). The noun-verb problem in Chinese aphasia. *Brain and Language, 41*, 203-233.

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition* (pp. 157-193). Hillsdale, NJ: Lawrence Erlbaum.

Bates, E., MacWhinney, B., Caselli, M. C., Devescovi, A., Natale, F., & Venza, V. (1984). A cross-linguistic study of the development of sentence interpretation strategies. *Child Development, 55*, 341-354.

Bates, E., McNew, S., MacWhinney, B., Devescovi, A., & Smith, S. (1982). Functional constraints on sentence processing: A cross-linguistic study. *Cognition, 11*(3), 245-299.

Bechtel, W., & Abrahamsen, A. (1991). *Connectionism and the Mind: An Introduction to Parallel Processing in Networks*. Cambridge, MA: Basil Blackwell.

Behrend, D. A., Harris, L. L., & Cartwright, K. B. (1995). Morphological cues to verb mearning: Verb inflections and the initial mapping of verb meanings. *Journal of Child Language, 22*, 89-106.

Bender, E. (2000). The syntax of Mandarin *ba:* Reconsidering the verbal analysis. *Journal of East Asian Linguistics, 9*(2), 105-145.

Bentivoglio, P. (1992). Linguistic correlations between subjects of one-argument verbs and subjects of more-than-one-argument verbs in spoken Spanish. In P. Hirschbühler & E. F. K. Koerner (Eds.), *Romance Languages and Modern Linguistic Theory. Selected Papers from the XX Linguistic Symposium on Romance Languages* (pp. 11-24). Amsterdam: John Benjamins.

Berko-Gleason, J. (1958). The child's learning of English morphology. *Word, 14*, 150-177.

Berwick, R. C., & Niyogi, P. (1996). Learning from triggers. *Linguistic Inquiry, 27*(4), 605-622.

Beveridge, M., & Marsh, L. (1991). The influence of linguistic context on young children's understanding of homophonic words. *Journal of Child Language, 18*, 459-467.

Bickerton, D. (1981). *Roots of Language*. Ann Harbor: Karoma.

Bickerton, D. (1984). The language bioprogram hypothesis. *Brain and Behavioral Sciences, 7*, 123-221.

Bisang, W. (1996). Areal typology and grammaticalization: processes of grammaticalization based on nouns and verbs in East and Mainland South East Asian languages. *Studies in Language, 20*(3), 519-597.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Black, M., & Chiat, S. (2003). Noun-verb dissociations: A multi-faceted phenomenon. *Journal of Neurolinguistics, 16*, 231-250.

Blake, B. J. (1994). *Case*. Cambridge: Cambridge University Press.

Bloom, L., Margulis, C., Tinker, E., & Fujita, N. (1996). Early conversations and word learning: Contributions from child and adult. *Child Development, 67*(6), 3154-3175.

Bod, R., Hay, J., & Jannedy, S. (Eds.). (2003). *Probabilistic Linguistics*. Cambridge, MA: MIT Press.

Boersma, P., & Hayes, B. (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry, 32*(1), 45-86.

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., Pascual, L., et al. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child Development, 75*(4), 1115-1139.

Bresnan, J. (2001). The emergence of the unmarked pronoun. In G. Legendre, J. Grimshaw & S. Vikner (Eds.), *Optimality-Theoretic Syntax* (pp. 113-142). Cambridge, MA: MIT Press.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, H. (2007). Predicting the dative alternation. In G. Boume, I. Kraemer & J. Zwarts (Eds.), *Cognitive Foundations of Interpretation* (pp. 69-94). Amsterdam: Royal Netherlands Academy of Science.

Briscoe, T. (2000). Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device. *Language, 76*(2), 245-296.

Brown, P. (1998). Children's first verbs in Tzeltal: Evidence for an early verb category. *Linguistics, 36*(4), 713-753.

Brown, R. (1973). *A First Language. The Early Stages*. Cambridge, MA: Harvard.

Brown, W. S., Marsh, J. T., & Smith, J. C. (1979). Principal component analysis of ERP differences related to the meaning of an ambiguous word. *Electroencephalography and Clinical Neurophysiology, 46*, 706-714.

Burgess, C., & Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes, 12*(2/3), 177-210.

Bybee, J. L. (1997). Semantic aspects of morphological typology. In J. L. Bybee, J. Haiman & S. A. Thompson (Eds.), *Essays on Language Function and Language Type: Dedicated to T. Givón* (pp. 25-37). Amsterdam: John Benjamins.

Bybee, J. L., & Hopper, P. J. (2001). Introduction to frequency and the emergence of linguistic structure. In J. L. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 1-24). Amsterdam: John Benjamins.

Bybee, J. L., Perkins, R., & Pagliuca, W. (1994). *The Evolution of Grammar. Tense, Aspect, and Modality in the Languages of the World*. Chicago, IL: University of Chicago Press.

Byrne, F. (1987). *Grammatical Relations in a Radical Creole*. Amsterdam: John Benjamins.

Camaioni, L., & Longobardi, E. (2001). Noun versus verb emphasis in Italian mother-to-child speech. *Journal of Child Language, 28*, 773-785.

Cameron-Faulkner, T., Lieven, E. V. M., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science, 27*, 843-873.

Campbell, R. N. (1969). *Noun substitutes in modern Thai. A study in pronominality*. The Hague: Mouton.

Campbell, R. N., & Bowe, T. B. (1983). Text and context in early language comprehension. In M. Donaldson, R. Grieve & C. Pratt (Eds.), *Early Childhood Developemtn and Education: Readings in Psychology* (pp. 115-126). Oxford: Basil Blackwell.

Cangelosi, A. (2005). The emergence of language: Neural and adaptive agent models. *Connection Science, 17*(3-4), 185-190.

Caramazza, A., & Hillis, A. E. (1991). Lexical organization of nouns and verbs in the brain. *Nature Neuroscience, 349*, 788-790.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development, 15*, 17-29.

Carroll, J. B., & White, M. N. (1973). Age of acquisition norms for 220 picturable nouns. *Journal of Verbal Learning Verbal Behavior, 12*, 563-576.

Cartwright, T. A., & Brent, M. R. (1997). Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. *Cognition, 63*, 121-170.

Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., et al. (1995). A cross-linguistic study of early lexical development. *Cognitive Development, 10*, 159-199.

Casenhiser, D. M. (2003). Children's resistance to homonymy: An experimental study of pseudohomonyms. *Journal of Child Language*.

Chang, H.-h. (1991). *Interaction between Syntax and Morphology: A Case Study of Mandarin Chinese*. Unpublished Doctoral dissertation, University of Hawaii.

Chang, H.-W. (1992). The acquisition of Chinese syntax. In H.-C. Chen & O. J. L. Tzeng (Eds.), *Language Processing in Chinese* (pp. 277-311). Amsterdam: North-Holland.

Chao, Y.-r. (1968). *A Grammar of Spoken Chinese*. Berkeley, CA: University of California Press.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *TRENDS in Cognitive Sciences, 10*(7), 335-344.

Chen, P. (1989). *The Distribution and Referential Interpretation of Empty Categories in Chinese.* Unpublished Doctoral dissertation, University of Texas at Austin.

Chien, Y.-C., & Lust, B. (1983). Topic-comment structure and grammatical subject in first language acquisition of Mandarin Chinese: A study of Equi-constructions. *Papers and Reports on Child Language Development, 22*, 74-82.

Chien, Y.-C., Lust, B., & Chiang, C.-P. (2003). Chinese children's comprehension of count-classifiers and mass-classifiers. *Journal of East Asian Linguistics, 12*, 91-120.

Childers, J. B., & Echols, C. H. (2004). 2 1/2-year-old children use animacy and syntax to learn a new noun. *Infancy, 5*(1), 109-125.

Childers, J. B., & Tomasello, M. (2002). Two-year olds learn novel nouns, verbs, and conventional actions from masses or distributed exposures. *Developmental Psychology, 38*(6), 967-978.

Choi, S. (2000). Caregiver input in English and Korean: Use of nouns and verbs in book-reading and toy-play contexts. *Journal of Child Language, 27*, 69-96.

Choi, S., & Gopnik, A. (1995). Early acquisition of verbs in Korean: A cross-linguistic study. *Journal of Child Language, 22*, 497-529.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, N. (1988). *Language and Problems of Knowledge*. Cambridge, MA: MIT Press.

Christiansen, M. H., & Chater, N. (1999). Connectionist natural language processing: The state of the art. *Cognitive Science, 23*(4), 417-437.

Christiansen, M. H., & Devlin, J. (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* (pp. 113-118). Mahwah, NJ: Lawrence Erlbaum.

Christiansen, M. H., & Monaghan, P. (2006). Discovering verbs through multiple-cue integration. In K. Hirsh-Pasek & R. M. Golinkoff (Eds.), *Action Meets Word: How Children Learn Verbs* (pp. 88-107). Oxford: Oxford University Press.

Chui, K.-W. (1992). Preferred argument structure for discourse understanding. In *Proceedings of COLING-92* (pp. 1142-1146).

Chung, S., & Timberlake, A. (1985). Tense, aspect, and mood. In T. Shopen (Ed.), *Language Typology and Syntactic Description. Volume 3: Grammatical Categories and the Lexicon*. Cambridge: Cambridge University Press.

Clark, E. V. (1982). The young word maker: A case study of innovation in the child's lexicon. In E. Wanner & L. R. Gleitman (Eds.), *Language Acquisition. The State of the Art* (pp. 390-425). Cambridge: Cambridge University Press.

Clark, E. V. (1993). *The Lexicon in Acquisition*. Cambridge: Cambridge University Press.

Clark, E. V. (2001). Morphology in language acquisition. In A. Spencer & A. M. Zwicky (Eds.), *The Handbook of Morphology* (pp. 374-389). Oxford: Blackwell.

Cole, P. (1987). Null objects in Universal Grammar. *Linguistic Inquiry, 18*(4), 597-612.

Comrie, B. (1976). *Aspect. An Introduction to the Study of Verbal Aspect and Related Problems*. Cambridge: Cambridge University Press.

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

Comrie, B., & Polinsky, M. S. (1998). The great Daghestanian case hoax. In A. Siewierska & J. J. Song (Eds.), *Case, Typology and Grammar. In Honor of Barry J. Blake*. Amsterdam: John Benjamins.

Cooke, J. R. (1968). *Pronominal reference in Thai, Burmese, and Vietnamese*. Berkeley: University of California Press.

Corbett, G. G. (2003). Agreement: Canonical instances and the extent of the phenomenon. In G. Booij, J. DeCesaris, A. Ralli & S. Scalise (Eds.), *Proceedings of the Third Mediterranean Morphology Meeting* (pp. 109-128). Barcelona: Pompeu Fabra.

Costa, J. (2001). The emergence of unmarked word order. In G. Legendre, J. Grimshaw & S. Vikner (Eds.), *Optimality-Theoretic Syntax* (pp. 171-204). Cambridge, MA: MIT Press.

Crocker, M. W., & Keller, F. (2006). Probabilistic grammars as models of gradience in language processing. In G. Fanselow, C. Féry, R. Vogel & Schlesewsky (Eds.), *Gradience in Grammar. Generative Perspectives* (pp. 227-245). Oxford: Oxford University Press.

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

De Bleser, R., & Kauschke, C. (2003). Acquisition and loss of nouns and verbs: Parallel or divergent patterns? *Journal of Neurolinguistics, 16*, 213-229.

DeGraff, M. F. (1993). Is Haitian Creole a Pro-drop language? In F. Byrne & J. Holm (Eds.), *Atlantic meets Pacific. A global view of pidginization and creolization* (pp. 71-90). Amsterdam: John Benjamins.

Déprez, V. (1994). Haitian Creole: A Pro-drop language? *Journal of Pidgin and Creole Languages, 9*(1), 1-24.

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

Dick, F., Wulfeck, B. B., Krupa-Kwiatkowski, M., & Bates, E. (2004). The development of complex sentence interpretation in typically developing children compared with children with specific language impairments or early unilateral focal lesions. *Developmental Science, 7*(3), 360-377.

Diessel, H., & Tomasello, M. (2001). The acquisition of finite complement clauses in English: A corpus-based analysis. *Cognitive Linguistics, 12*(2), 97-141.

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

Doherty, M. J. (2000). Children's understanding of homonymy. *Journal of Child Language, 27*, 367-392.

Doherty, M. J. (2004). Children's difficulty in learning homonyms. *Journal of Child Language, 31*, 203-214.

Döpke, S. (1998). Competing language structures: The acquisition of verb placement by bilingual German-English children. *Journal of Child Language, 25*, 555-584.

Dowty, D. R. (1991). Thematic proto-roles and argument selection. *Language, 67*(3), 547-619.

Dryer, M. S. (1988). Object-Verb order and Adjective-Noun order: Dispelling a myth. *Lingua, 74*, 185-217.

Dryer, M. S. (1989). Discourse-governed word order and word order typology. *Belgian Journal of Linguistics, 4*, 69-90.

Dryer, M. S. (1991). SVO languages and the OV:VO typology. *Journal of Linguistics, 27*, 443-482.

Dryer, M. S. (1992). The Greenbergian word order correlations. *Language, 68*(1), 81-138.

Dryer, M. S. (1997). On the six-way word order typology. *Studies in Language, 21*(1), 69-103.

Du Bois, J. W. (1987). The discourse basis of ergativity. *Language, 63*(4), 805-855.

Ellis, A. W., & Morrison, C. M. (1998). Real age-of acquisition effects in lexical retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 513-523.

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

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition, 48*, 71-99.

Elman, J. L. (1998). Generalization, simple recurrent networks and the emergence of structure. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the 20th Annual Conference of the Cognitive Science Society*. Mahway, NJ: Lawrence Erlbaum.

Elman, J. L. (2005). Connectionist models of cognitive development: Where next? *TRENDS in Cognitive Sciences, 9*(3), 111-117.

Erbaugh, M. S. (1982). *Coming to Order: Natural Selection and the Origin of Syntax in the Mandarin Speaking Child.* Unpublished Doctoral dissertation, UC Berkeley.

Erbaugh, M. S. (1986). Taking stock: The development of Chinese noun classifiers historically and in young children. In C. G. Craig (Ed.), *Noun Classes and Categorization* (pp. 399-436). Amsterdam: John Benjamins.

Erbaugh, M. S. (1992). The acquisition of Mandarin. In D. I. Slobin (Ed.), *The Crosslinguistic Study of Language Acquisition. Volume 3* (pp. 373-455). Hillsdale, NJ: Lawrence Erlbaum.

Erbaugh, M. S. (2006). Chinese classifiers: Their use and acquisition. In P. Li, H. Tan, E. Bates & O. J. L. Tzeng (Eds.), *Handbook of East Asian Psycholinguistics* (Vol. Volume 1. Chinese, pp. 39-51). Cambridge: Cambridge University Press.

Falk, Y. N. (2006). *Subjects and Universal Grammar. An Explanatory Theory*. Cambridge: Cambridge University Press.

Federmeier, K. D., Segal, J. B., Lombrozo, T., & Kutas, M. (2000). Brain responses to nouns, verbs and class-ambiguous words in context. *Brain, 12*, 2552-2566.

Felser, C., Marinis, T., & Clahsen, H. (2003). Children's processing of ambiguous sentences: A study of relative clause attachment. *Language Acquisition, 11*(3), 127-163.

Ferraz, L. I. (1987). Portuguese creoles of West Africa and Asia. In G. G. Gilbert (Ed.), *Pidgin and Creole Languages. Essays in Memory of John R. Reinecke.* Honololu, HI: University of Hawaii Press.

Fisher, C. (2002). Structural limits on verb mapping: The role of abstract structure in 2.5-year-olds' interpretations of novel verbs. *Developmental Science, 5*(1), 55-64.

Fisher, C., & Tokura, H. (1996). Acoustic cues to grammatical structure in infant-directed speech: Cross-linguistic evidence. *Child Development, 67*(6), 3192-3218.

Flack, K. (2005). *Ambiguity avoidance as contrast preservation: Case and word order freezing in Japanese.*Unpublished manuscript.

Foley, W. A. (1991). *The Yimas Language of New Guinea.* Stanford, CA: Stanford University Press.

Forbes, J. N., & Farrar, M. J. (1995). Learning to represent word meaning: What initial training events reveal about children's developing action verb concepts. *Cognitive Development, 10*, 1-20.

Freudenthal, D., Pine, J. M., & Gobet, F. (2002). Subject omission in children's language: The case for performance limitations in learning. In W. Gray & C. Schunn (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 334-339).

Friedemann, N. S. d., & Patiño, C. (1983). *Lengua y sociedad en el Palenque de San Basilio.* Bogotá: Instituto Caro y Cuervo.

Gentner, D. (1981). Some interesting differences between verbs and nouns. *Cognition and Brain Theory, 4*(2), 161-178.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. A. Kuczay (Ed.), *Language Development. Volume 2: Language, Thought and Culture* (pp. 310-334). Hillsdale, NJ: Lawrence Erlbaum.

Gentner, D. (2006). Discovering verbs through multiple-cue integration. In K. Hirsh-Pasek & R. M. Golinkoff (Eds.), *Action Meets Word: How Children Learn Verbs* (pp. 544-564). Oxford: Oxford University Press.

Gentner, D. (2006). Why verbs are hard to learn. In K. Hirsh-Pasek & R. M. Golinkoff (Eds.), *Action Meets Word: How Children Learn Verbs* (pp. 544-564). Oxford: Oxford University Press.

Gentner, D., & Boroditsky, L. (2001). Individuation, relativity, and early word learning. In M. Bowerman & S. C. Levinson (Eds.), *Language Acquisition and Conceptual Development* (pp. 215-256). Cambridge: Cambridge University Press.

Gerken, L. (2004). Nine-month-olds extract structural principles required for language. *Cognition, 93*, B89-B96.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language, 32*, 249-268.

Gibson, E., & Broihier, K. (1998). Optimality Theory and human sentence processing. In P. Barbosa, D. Fox, P. Hagstrom, M. McGinnis & D. Pesetsky (Eds.), *Is the Best Good Enough? Optimality and Competition in Syntax* (pp. 157-192). Cambridge, MA: MIT Press.

Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry, 25*, 407-454.

Gil, D. (1994). The structure of Riau Indonesian. *Nordic Journal of Linguistics, 17*, 179-200.

Gil, D. (2006). The acquisition of voice morphology in Jakarta Indonesian. In N. Gagarina & I. Gülzow (Eds.), *The Acquisition of Verbs and their Grammar. The Effect of Particular Languages* (pp. 201-228). Dordrecht: Springer Verlag.

Gillette, J., Gleitman, H., Gleitman, L. R., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition, 73*, 135-176.

Gilligan, G. M. (1987). *A Cross-Linguistic Approach to the Pro-drop Parameter*. Unpublished Doctoral dissertation, University of Southern California.

Givón, T. (1979). From discourse to syntax: Grammar as a processing strategy. In T. Givón (Ed.), *Syntax and Semantics 12. Discourse and Syntax* (pp. 81-112). New York: Academic Press.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language Learning and Development, 1*(1), 23-64.

Goldberg, A. E. (2002). Surface generalizations: An alternative to alternations. *Cognitive Linguistics, 13*(4), 327-356.

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics, 15*(3), 289-316.

Goldfield, B. A. (2000). Nouns before verbs in comprehension vs. production: The view from pragmatics. *Journal of Child Language, 27*, 501-520.

Goldwater, S. J. (2007). *Nonparametric Bayesian Models of Lexical Acquisition.* Unpublished Ph.D. Dissertation, Brown University.

Golinkoff, R. M., Jacquet, R. C., Hirsh-Pasek, K., & Nandakumar, R. (1996). Lexical principles may underlie the learning of verbs. *Child Development, 67*(6), 3101-3119.

Gomez, R. L. (2002). Variability and dectection of invariant structure. *Psychological Science, 13*(5), 431-436.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70*, 109-135.

Gomez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *TRENDS in Cognitive Sciences, 4*(5), 178-186.

Gopnik, A., Choi, S., & Baumberger, T. (1996). Cross-linguistic differences in early semantic and cognitive development. *Cognitive Development, 11*, 197-227.

Gopnik, A., & Tenenbaum, J. B. (2007). Bayesian networks, Bayesian learning and cognitive development. *Developmental Science, 10*(3), 281-287.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (Ed.), *Universals of Language* (pp. 73-113). Cambridge, MA: MIT Press.

Grieser, D. L., & Kuhl, P. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in Motherese. *Developmental Psychology, 24*(1), 14-20.

Grima, J. A. (1986). Discourse factors contributing to the understanding of a zero pronoun in a passage from the Phrârâatchawícaan. In R. J. Bickner, T. J. Hudak & P. Peyasantiwong (Eds.), *Papers from a conference on Thai studies in honor of William J. Gedney* (pp. 159-169). Ann Arbor, Michigan: University of Michigan.

Grinstead, J. (2000). Case, inflection and subject licensing in child Catalan and Spanish. *Journal of Child Language, 27*, 119-155.

Haegeman, L. (1999). Adult null subjects in non pro-drop languages. In M.-A. Friedemann & L. Rizzi (Eds.), *The Acquisition of Syntax* (pp. 129-169). Harlow: Longman.

Haegeman, L., & Ihsane, T. (2001). Adult null subjects in the non-pro-drop languages: Two diary dialects. *Language Acquisition, 9*(4), 329-346.

Hare, M., & Elman, J. L. (1995). Learning and morphological change. *Cognition, 56*, 61-98.

Haspelmath, M., Dryer, M. S., Gil, D., & Comrie, B. (Eds.). (2005). *The World Atlas of Language Structures*. Oxford: Oxford University Press.

Hatton, H. (1975). A Thai discourse pattern. In J. G. Harris & J. R. Chamberlain (Eds.), *Studies in Thai linguistics. In honor of William J. Gedney* (pp. 231-251). Bangkok: Central Institute of English Language. Office of State Universities.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition, 78*, B53-B64.

Hawkins, J. A. (1983). *Word Order Universals*. New York, NY: Academic Press.

Hawkins, J. A. (1988). Explaining Language Universals. In J. A. Hawkins (Ed.), *Explaining Language Universals* (pp. 3-28). Oxford: Basil Blackwell.

Hawkins, J. A. (1993). Heads, parsing and word-order universals. In G. G. Corbett, N. M. Fraser & S. McGlashan (Eds.), *Heads in Grammatical Theory* (pp. 231-265). Cambridge: Cambridge University Press.

Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.

Hawkins, J. A. (2002). Symmetries and asymmetries: Their grammar, typology and parsing. *Theoretical Linguistics, 28*, 95-150.

Hawkins, J. A. (2004). Efficiency and complexity in grammars: Three general principles. In J. Moore & M. S. Polinsky (Eds.), *The Nature of Explanation in Linguistic Theory* (pp. 121-152). Stanford, CA: CSLI Publications.

Hernandez, A., Li, P., & MacWhinney, B. (2005). The emergence of competing modules in bilingualism. *TRENDS in Cognitive Sciences, 9*(5), 220-225.

Herr-Israel, E., & McCune, L. (2006). Dynamic event words, motion events and the transition to verb meanings. In N. Gagarina & I. Gülzow (Eds.), *The Acquisition of Verbs and their Grammar. The Effect of Particular Languages* (pp. 125-150). Dordrecht: Springer Verlag.

Höhle, B., Weissenborn, J., Kiefer, D., Schulz, A., & Schmitz, M. (2004). Functional elements in infants' speech processing: The role of determiners in the syntactic categorization of lexical elements. *Infancy, 5*(3), 341-353.

Holm, J. (2000). *An introduction to pidgins and creoles*. Cambridge: Cambridge University Press.

Houston-Price, C., Plunkett, K., & Harris, P. (2005). 'Word-learning wizardy' at 1;6. *Journal of Child Language, 32*, 175-189.

Hu, M. (1991). *Function of Word Order in Mandarin Chinese*. Unpublished Doctoral dissertation, University of Florida.

Hu, Q. (1993). *The Acquistion of Chinese Classifiers by Young Mandarin-speaking Children*. Unpublished Ph.D. Dissertation, Boston University.

Huang, C.-c. (2006). Child language acquisition of temporality in Mandarin Chinese. In P. Li, H. Tan, E. Bates & O. J. L. Tzeng (Eds.), *Handbook of East Asian Psycholinguistics* (Vol. Volume 1. Chinese, pp. 52-60). Cambridge: Cambridge University Press.

Huang, C.-T. J. (1984). On the distribution and reference of empty pronouns. *Linguistic Inquiry, 15*(4), 531-574.

Huang, C.-T. J. (1989). Pro-drop in Chinese: A generalized control theory. In O. Jaeggli & K. J. Safir (Eds.), *The Null Subject Parameter* (pp. 185-214). Dordrecht: Kluwer.

Huang, Y. (1995). On null subjects and null objects in generative grammar. *Linguistics, 33*, 1081-1123.

Huang, Y. (2000). *Anaphora. A Cross-linguistic Approach*. Oxford: Oxford University Press.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development, 1*(2), 151-195.

Jaeggli, O. (1982). *Topics in Romance Syntax*. Dordrecht: Foris.

Jaeggli, O., & Safir, K. J. (1989). The Null Subject parameter and parametric theory. In O. Jaeggli & K. J. Safir (Eds.), *The Null Subject Parameter* (pp. 1-44). Dordrecht: Kluwer.

Jara M., C. V. (1996). Sistema de tiempo-modo-aspecto en criollos de base española. *Filología y Lingüística, 22*(2), 105-132.

Jurafsky, D. (2001). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay & S. Jannedy (Eds.), *Probabilistic Linguistics*. Cambridge, MA: MIT Press.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. L. Bybee & P. J. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 229-254). Amsterdam: John Benjamins.

Kaschak, M. P., & Saffran, J. R. (2006). Idiomatic syntactic constructions and language learning. *Cognitive Science, 30*, 43-63.

Keenan, E. L., & Comrie, B. (1977). Noun Phrase Accessibility and Universal Grammar. *Linguistic Inquiry, 8*(1), 63-99.

Keller, F. (2000). Evaluating competition-based models of word order. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 747-752). Mahwah, NJ: Lawrence Erlbaum.

Kempe, V., & MacWhinney, B. (1999). Processing of morphological and semantic cues in Russian and German. *Language and Cognitive Processes, 14*(2), 129-171.

Kihm, A. (2000). Are creole languages "perfect" languages? In J. H. McWhorter (Ed.), *Language Change and Language Contact in Pidgins and Creoles* (pp. 163-199). Amsterdam: John Benjamins.

Kim, M., McGregor, K. K., & Thompson, C. K. (2000). Early lexical development in English- and Korean-speaking children: Language-general and language-specific patterns. *Journal of Child Language, 27*, 225-254.

Kim, Y.-J. (2000). Subject/Object drop in the acquisition of Korean: A cross-linguistic comparison. *Journal of East Asian Linguistics, 9*, 325-351.

Kirby, S. (1997). Competing motivations and emergence: explaining implicational hierarchies. *Linguistic Typology, 1*, 5-32.

Kouwenberg, S. (1990). Complementizer *pa*, the finiteness of its complements, and some remarks on empty categories in Papiamento. *Journal of Pidgin and Creole Languages, 5*(1), 39-51.

Kroch, A. (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change, 1*, 199-244.

Kuhn, J. (2003). *Optimality-Theoretic Syntax: A Declarative Approach*. Stanford, CA: CSLI Publications.

Kural, M. (1997). Postverbal Constituents in Turkish and the Linear Correspondence Axiom. *Linguistic Inquiry, 28*(3), 498-519.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar. Volume I: Theoretical Prerequisites*. Stanford, CA: Stanford University Press.

Langacker, R. W. (1987). Nouns and verbs. *Language, 63*, 53-94.

LaPolla, R. (1990). *Grammatical relations in Chinese: Synchronic and diachronic considerations*. Unpublished Doctoral dissertation, UC Berkeley.

Lehmann, C. (1984). *Der Relativsatz: Typologie seiner Strukturen, Theorie seiner Funktionen, Kompendium seiner Grammatik*. Tübingen: G. Narr.

Levy, E., & Nelson, K. (1994). Words in discourse: A dialectial approach to the acquisition of meaning and use. *Journal of Child Language, 21*, 367-389.

Levy, R. (2006). *Expectation-based syntactic comprehension*.Unpublished manuscript.

Levy, R., & Manning, C. D. (2003). *Is it harder to parse Chinese, or the Chinese Treebank*. Paper presented at the Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan.

Li, C. N. (1997). On zero anaphora. In J. L. Bybee, J. Haiman & S. A. Thompson (Eds.), *Essays on Language Function and Language Type: Dedicated to T. Givón* (pp. 275-300). Amsterdam: John Benjamins.

Li, C. N., & Thompson, S. A. (1976). Subject and topic: A new typology of language. In C. N. Li (Ed.), *Subject and Topic* (pp. 458-489). New York, NY: Academic Press.

Li, C. N., & Thompson, S. A. (1979). Third-person pronouns and zero-anaphora in Chinese discourse. In T. Givón (Ed.), *Syntax and Semantics 12. Discourse and Syntax* (pp. 311-335). New York: Academic Press.

Li, C. N., & Thompson, S. A. (1981). *Mandarin Chinese. A Functional Reference Grammar*. Berkeley, CA: University of California Press.

Li, P. (1998). Crosslinguistic variation and sentence processing: The case of Chinese. *Syntax and Semantics, 31*, 33-53.

Li, P. (2002). Emergent semantic structure and language acquisition: A dynamic perspective. In H. S. R. Kao, C. K. Leong & D. G. Gao (Eds.), *Cognitive neuroscience studies of the Chinese language* (pp. 79-98). Hong Kong: Hong Kong University Press.

Li, P. (2006). Modeling language acquisition and representation: Connectionist networks. In P. Li, H. Tan, E. Bates & O. J. L. Tzeng (Eds.), *Handbook of East Asian Psycholinguistics* (Vol. Volume 1. Chinese, pp. 320-329). Cambridge: Cambridge University Press.

Li, P., Bates, E., Liu, H., & MacWhinney, B. (1992). Cues as Functional Constraints on Sentence Processing in Chinese. In H.-C. Chen & O. J. L. Tzeng (Eds.), *Language Processing in Chinese* (pp. 207-234). Amsterdam: North-Holland.

Li, P., Bates, E., & MacWhinney, B. (1993). Processing a language without inflections: A reaction time study of sentence interpretation in Chinese. *Journal of Memory and Language, 32*, 169-192.

Li, P., & Bowerman, M. (1998). The acquisition of lexical and grammatical aspect in Chinese. *First Language, 18*, 311-350.

Li, P., Burgess, C., & Lund, K. (2000). The acquisition of word meaning through global lexical co-occurrences. In E. V. Clark (Ed.), *Proceedings of the Thirtieth Stanford Child Language Research Forum* (pp. 167-178). Stanford, CA: Center for the Study of Language and Information.

Li, P., Jin, Z., & Tan, L. H. (2004). Neural representation of nouns and verbs in Chinese: An fMRI study. *NeuroImage, 21*, 1533-1541.

Li, P., Shu, H., Yip, M., Zhang, Y., & Tang, Y. (2002). Lexical ambiguity in sentence processing: Evidence from Chinese. In M. Nakayama (Ed.), *Sentence Processing in East Asian Languages* (pp. 111-129). Stanford, CA: Center for the Study of Language and Information.

Lidz, J., Waxman, S. R., & Freedman, J. (2003). What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structures at 18 months. *Cognition, 89*, B65-B73.

Lieven, E. V. M. (1997). Variation in a crosslinguistic context. In D. I. Slobin (Ed.), *The Crosslinguistic Study of Language Acquisition. Volume 5: Expanding the Contexts* (pp. 199-263). Hillsdale, NJ: Lawrence Erlbaum.

Lieven, E. V. M., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: A usage-based approach. *Journal of Child Language, 30*, 333-370.

Lieven, E. V. M., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early development. *Journal of Child Language, 24*, 187-219.

Lightfoot, D. (1991). *How to set parameters: Arguments from language change*. Cambridge, MA: MIT Press.

Lipski, J. M. (2002). *Null subjects in (Romance-derived) creoles: Routes of evolution.*Unpublished manuscript.

MacWhinney, B. (1978). The Acquisition of Morphophonology. *Monographs of the Sociery for Research in Child Development, 43*(1-2), 1-123.

MacWhinney, B. (1987). The Competition Model. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition* (pp. 249-308). Hillsdale, NJ: Lawrence Erlbaum.

MacWhinney, B. (1995). *The CHILDES Project: Tools for Analyzing Talk*. Hillsdale, NJ: Lawrence Erlbaum.

MacWhinney, B. (2001). Emergentist approaches to language. In J. L. Bybee & P. J. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 449-470). Amsterdam: John Benjamins.

MacWhinney, B. (2004). A multiple process solution to the logical problem of language acquisition. *Journal of Child Language, 31*, 883-914.

MacWhinney, B., & Bates, E. (Eds.). (1989). *The crosslinguistic study of sentence processing*. Cambridge: Cambridge University Press.

Maling, J., & Zaenen, A. (1978). The nonuniversality of a surface filter. *Linguistic Inquiry, 9*, 475-497.

Mandel, D. R., Kemler Nelson, D. G., & Jusczyk, P. W. (1996). Infants remember the order of words in a spoken sentence. *Cognitive Development, 44*, 181-196.

Mandler, J. M. (2004). A synopsis if *The foundations of mind: Origins of conceptual thought* (2004). New York: Oxford University Press. *Developmental Science, 7*(5), 499-505.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Maratsos, M. (1998). Commentary: Relations of lexical specificity to general categories. *Linguistics, 36*(4), 831-846.

Maratsos, M., & Chalkley, M. A. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K. Nelson (Ed.), *Children's Language* (pp. 127-214). New York, NY: Gardner Press.

Marchman, V. (1997). Children's productivity in the English past tense: The role of frequency, phonology, and neighborhood structure. *Cognitive Science, 21*(3), 283-304.

Marchman, V., & Bates, E. (1994). Continuity in lexical and morphological development: A test of the critical mass hypothesis. *Journal of Child Language, 21*, 339-366.

Marchman, V., Martínez-Sussman, C., & Dale, P. S. (2004). The language-specific nature of grammatical development: Evidence from bilingual language learners. *Developmental Science, 7*(2), 212-224.

Marchman, V., Plunkett, K., & Goodman, J. (1995). Overregularization in English plural and past tense inflectional morphology:. *Journal of Child Language, 24*, 767-779.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by 7-month-old infants. *Science, 283*, 77-80.

Marshall, J. (2003). Noun-verb dissociations: Evidence from acquisition and developmental and acquired impairments. *Journal of Neurolinguistics, 16*, 67-84.

Matthews, D., Lieven, E. V. M., Theakston, A. L., & Tomasello, M. (2005). The role of frequency in the acquisition of English word order. *Cognitive Development, 20*, 121-136.

Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition, 78*, 91-121.

May Vihman, M., & Vija, M. (2006). The acquisition of verbal inflection in Estonian: Two case studies. In N. Gagarina & I. Gülzow (Eds.), *The Acquisition of Verbs and their Grammar. The Effect of Particular Languages* (pp. 263-296). Dordrecht: Springer Verlag.

Mazzocco, M. M. (1997). Children's interpretation of homonyms: A developmental study. *Journal of Child Language, 24*, 441-467.

Mazzocco, M. M., Myers, G. F., Thompson, L. A., & Desai, S. S. (2003). Possible explanations for children's literal interpretation of homonyms. *Journal of Child Language, 30*, 879-904.

McWhorter, J. H. (1997). *Towards a New Model of Creole Genesis*. New York: Peter Lang.

McWhorter, J. H. (1998). Identifying the creole prototype. Vindicating a typological class. *Language, 74*(4), 788-818.

McWhorter, J. H. (2001). The world's simplest grammars are creole grammars. *Linguistic Typology, 5*(2-3), 125-166.

Meyerhoff, M. (2000). The emergence of creole subject-verb agreement and the licensing of null subjects. *Language Variation and Change, 12*, 203-230.

Miao, X., & Zhu, M. (1992). Language development in Chinese children. In H.-C. Chen & O. J. L. Tzeng (Eds.), *Language Processing in Chinese* (pp. 237-276). Amsterdam: North-Holland.

Milostan, J. (1995). Connectionist modeling of the Fast Mapping phenomenon. *CRL Newsletter, 9*(3), 1-13.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90*, 91-117.

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26*, 393-424.

Mithun, M. (1987). Is basic word order universal? In R. S. Tomlin (Ed.), *Coherence and Grounding in Discourse* (pp. 281-328). Amsterdam: John Benjamins.

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition, 96*, 143-182.

Monaghan, P., Gonitzke, M., & Chater, N. (2003). Two wrongs make a right: Learnability and word order consistency. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

Morimoto, Y. (2002). *From synchrony to diachrony: Topic salience and cross-linguistic patterns of agreement.*Unpublished manuscript.

Morris, W. C., Cottrell, G. W., & Elman, J. L. (2000). A connectionist simulation of the empirical acquisition of grammatical relations. In S. Wermter & R. Sun (Eds.), *Hybrid Neural Symbolic Integration* (pp. 175-193). Berling: Springer.

Müller, G. (2001). *Free word order, morphological case, and sympathy theory.*Unpublished manuscript.

Mufwene, S. S. (1988). The Small *pro* and Inflectional Morphology. *Linguistic Analysis, 18*(3-4), 235-242.

Munakata, Y., & McClelland, J. L. (2003). Connectionist models of development. *Developmental Science, 6*(4), 413-429.

Muysken, P. (1981). Creole tense/mood/aspect systems: The unmarked case. In P. Muysken (Ed.), *Generative studies on Creole languages* (pp. 181-199). Dordrecht: Cinnaminson, NJ.

Muysken, P., & Law, P. (2001). Creole studies. A theoretical linguist's field guide. *Glot International, 5*(2), 47-57.

Naigles, L. R. (2002). Form is easy, meaning is hard: Resolving a paradox in early child language. *Cognition, 86*, 157-199.

Naigles, L. R. (2003). Paradox lost? No, paradox found! Reply to Tomasello and Akhtar (2003). *Cognition, 88*, 325-329.

Naigles, L. R., Bavin, E. L., & Smith, M. A. (2005). Toddlers recognize verbs in novel situations and sentences. *Developmental Science, 8*(5), 424-431.

Naigles, L. R., & Kako, E. T. (1993). First contact in verb acquisition: Defining a role for syntax. *Child Development, 64*(6), 1665-1687.

Narayanan, S., & Jurafsky, D. (2002). A Bayesian model predicts human parse preference and reading times in sentence processing. In T. G. Dietterich, S. Becker & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14* (pp. 59-65). Cambridge, MA: MIT Press.

Newmeyer, F. J. (2003). Grammar is grammar and usage is usage. *Language, 79*(4), 682-707.

Newmeyer, F. J. (2005). *Possible and Probable Languages: A Generative Perspective on Linguistic Typology*. New York, NY: Oxford University Press.

Newmeyer, F. J. (2005). A reply to the critiques of 'Grammar is grammar and usage is usage'. *Language, 81*(1), 228-236.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science, 14*, 11-28.

Nichols, J. (1986). Head-marking and dependent-marking grammar. *Language, 62*(1), 56-119.

Nichols, J. (1992). *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

Nichols, J., Peterson, D. A., & Barnes, J. (2004). Transitivizing and detransitivizing languages. *Linguistic Typology, 8*, 149-211.

Nicolis, M. (2005). *On pro-drop*. Unpublished Ph.D. Dissertation, Universita degli Studi di Siena.

Nicolis, M. (2007). *The null subject parameter and correlating properties: The case of Creole languages*.Unpublished manuscript.

Olguin, R., & Tomasello, M. (1993). Twenty-five-month-old children do not have a grammatical category of verb. *Cognitive Development, 8*, 245-272.

Ozkaragoz, I. Z. (1987). *The Relational Structure of Turkish Syntax*. Unpublished Doctoral dissertation, UC San Diego.

Packard, J. L. (2006). The manifestation of aphasia syndromes in Chinese. In P. Li, H. Tan, E. Bates & O. J. L. Tzeng (Eds.), *Handbook of East Asian Psycholinguistics* (Vol. Volume 1. Chinese, pp. 330-345). Cambridge: Cambridge University Press.

Perlmutter, D. (1971). *Deep and Surface Constraints in Syntax*. New York: Holt, Rinehart & Winston.

Peters, A. M. (1997). Language typology, prosody, and the acquisition of grammatical morphemes. In D. I. Slobin (Ed.), *The Crosslinguistic Study of Language Acquisition. Volume 5: Expanding the Contexts* (pp. 135-197). Hillsdale, NJ: Lawrence Erlbaum.

Piatelli-Palmarini, M. (1989). Evolution, selection and cognition: From "learning" to parameter setting in biology and in the study of language. *Cognition, 31*, 1-44.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*, 56-115.

Plunkett, K., & Marchman, V. (1991). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. In D. S. Touretzky, J. L. Elman, T. J. Sejnowski & G. E. Hinton (Eds.), *Connectionist Models: Proceedings of th 1990 Summer School* (pp. 87-106). San Mateo, CA: Morgan Kaufman.

Plunkett, K., & Marchman, V. (1996). Learning from a connectionist model of the acquisition of the English past tense. *Cognition, 61*, 299-308.

Polinsky, M. S. (2005). Word class distinctions in an incomplete grammar. In D. Ravid & H. Bat-Zeev Shyldkrot (Eds.), *Perspectives on Language and Language Development* (pp. 419-436). Dordrecht: Kluwer.

Polinsky, M. S., & Van Everbroeck, E. (2003). Development of gender classifications: Modeling the historical change from Latin to French. *Language, 79*, 356-390.

Prince, A., & Smolensky, P. (1993). *Optimality Theory: Constraint interaction in generative grammar.*Unpublished manuscript.

Prince, E. F. (1998). Subject-Prodrop in Yiddish. In P. Bosch & R. A. van der Sandt (Eds.), *Focus: linguistic, cognitive, and computational perspectives* (pp. 82-104). Cambridge: Cambridge University Press.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*(4), 425-469.

Reyes, I. (2003). A study of sentence interpretation in Spanish monolingual children. *First Language, 23*(3), 285-309.

Rizzi, L. (1982). *Issues in Italian Syntax*. Dordrecht: Foris.

Roberts, I. (1999). Verb movement and markedness. In M. F. DeGraff (Ed.), *Language Creation and Language Change* (pp. 287-327). Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland & t. P. R. Group (Eds.), *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Volume 1: Foundations* (pp. 318-362). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart & t. P. R. Group (Eds.), *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models* (pp. 216-271). Cambridge, MA: MIT Press.

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language, 47*, 172-196.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926-1928.

Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multilevel statistical learning by 12-month-old infants. *Infancy, 4*(2), 273-284.

Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multilevel statistical learning by 12-month-old infants. *Infancy, 4*(2), 273-284.

Samek-Lodovici, V. (2001). Crosslinguistic typologies in Optimality Theory. In G. Legendre, J. Grimshaw & S. Vikner (Eds.), *Optimality-Theoretic Syntax* (pp. 315-354). Cambridge, MA: MIT Press.

Samek-Lodovici, V. (2003). *Agreement impoverishment under subject inversion: A crosslinguistic analysis.*Unpublished manuscript.

Sandhofer, C. M., Smith, L. B., & Luo, J. (2000). Counting nouns and verbs in the input: Differential frequencies, different kinds of learning? *Journal of Child Language, 27*, 561-585.

Sapir, E. (1921). *Language. An Introduction to the Study of Speech*. San Diego: Harcourt Brace.

Savage, C., Lieven, E. V. M., Theakston, A. L., & Tomasello, M. (2003). Testing the abstractness of children's linguistic representations: Lexical and structural priming of syntactic constructions in young children. *Developmental Science, 6*(5), 557-567.

Schafer, G., & Plunkett, K. (1996). Rapid word learning by 15-month-olds under tightly controlled conditions. *CRL Newsletter, 10*(5), 1-13.

Schwegler, A. (1993). Subject pronouns and person/number in Palenquero. In F. Byrne & J. Holm (Eds.), *Atlantic meets Pacific. A global view of pidginization and creolization* (pp. 145-161). Amsterdam: John Benjamins.

Seidenberg, M. S., & MacDonald, M. C. (1999). A probabilistic constraints approach to language acquisition and processing. *Cognitive Science, 23*(4), 569-588.

Seidenberg, M. S., MacDonald, M. C., & Saffran, J. R. (2002). Does grammar start where statistics stop? *Science, 298*, 553-554.

Shi, D. (2000). Topic and topic-comment constructions in Mandarin Chinese. *Language, 76*(2), 383-408.

Shi, R. (2006). Basic syntactic categories in early language development. In P. Li, H. Tan, E. Bates & O. J. L. Tzeng (Eds.), *Handbook of East Asian Psycholinguistics* (Vol. Volume 1. Chinese, pp. 90-102). Cambridge: Cambridge University Press.

Shultz, T. R. (2007). The Bayesian revolution approaches psychological development. *Developmental Science, 10*(3), 357-364.

Siewierska, A. (1988). *Word Order Rules*. London: Croom Helm.

Siewierska, A. (1996). Word order type and alignment type. *Zeitschrift für Sprachtypologie und Universalienforschung, 49*(2), 149-176.

Siewierska, A. (Ed.). (1998). *Constituent Order in the Languages of Europe*. Berlin: Mouton de Gruyter.

Siewierska, A. (1998). Variation in major constituent order: a global and a European perspective. In A. Siewierska (Ed.), *Constituent Order in the Languages of Europe* (pp. 475-551). Berlin: Mouton de Gruyter.

Slobin, D. I. (1973). Cognitive prerequisites for the development of grammar. In C. A. Ferguson & D. I. Slobin (Eds.), *Studies of Child Language Development* (pp. 175-208). New York, NY: Holt, Rinehart and Winston.

Slobin, D. I. (1981). Introduction: Why study acquisition crosslinguistically? In D. I. Slobin (Ed.), *The Crosslinguistic Study of Language Acquisition. Volume 1: The Data* (pp. 3-24). Hillsdale, NJ: Lawrence Erlbaum.

Slobin, D. I. (1985). Crosslinguistic evidence for the language-making capacity. In D. I. Slobin (Ed.), *The Crosslinguistics Study of Language Acquisition* (Vol. Volume 2: Theoretical Issues, pp. 1157-1249). Hillsdale, NJ: Lawrence Erlbaum.

Slobin, D. I., & Bever, T. G. (1982). Children use canonical sentence schemas: A crosslinguistic study of word order and inflections. *Cognition, 12*, 229-265.

Song, J. J. (2001). *Linguistic Typology: Morphology and Syntax*. Harlow: Longman.

Speas, M. (1996). Null objects in functional projections. In J. Rooryck & L. Zaring (Eds.), *Phrase structure and the lexicon* (pp. 187-211). Dordrecht: Kluwer.

Speas, M. (2001). Constraints on null pronouns. In G. Legendre, J. Grimshaw & S. Vikner (Eds.), *Optimality-Theoretic Syntax* (pp. 393-426). Cambridge, MA: MIT Press.

Steele, S. (1978). Word Order Variation: A Typological Study. In J. H. Greenberg (Ed.), *Universals of Human Language* (Vol. Volume 4, pp. 585-624). Stanford, CA: Stanford University Press.

Su, I.-R. (2001). Context effects on sentence processing: A study based on the Competition Model. *Applied Psycholinguistics, 22*, 167-189.

Sun, C. (1996). *Word-Order Change and Grammaticalization in the History of Chinese*. Stanford, CA: Stanford University Press.

Sun, C., & Givón, T. (1985). On the so-called SOV word order in Mandarin Chinese: A quantified text study and its implications. *Language, 61*(2), 329-351.

Syea, A. (1993). Null subject in Mauritian Creole and the Pro-drop parameter. In F. Byrne & J. Holm (Eds.), *Atlantic meets Pacific. A global view of pidginization and creolization* (pp. 91-102). Amsterdam: John Benjamins.

Tai, J. H.-Y. (1994). Chinese classifier system and human categorization. In M. Y. Chen & O. J. L. Tzeng (Eds.), *In Honor of Professor William S-Y. Wang: Interdisciplinary Studies on Language and Language Change* (pp. 479-494). Taipei: Pyramid Press.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science, 12*, 49-100.

Tan, F. (1991). *Notion of subject in Chinese*. Unpublished Doctoral dissertation, Stanford.

Tao, H. (1996). *Units in Mandarin conversation. Prosody, discourse, and grammar*. Amsterdam: John Benjamins.

Taraldsen, K. T. (1980). *On the nominative island constraint. Vacuous application, and the That-trace filter*. Bloomington, IN.

Tardif, T. (1996). Nouns are not always learned before verbs: Evidence from Mandarin speakers' early vocabularies. *Developmental Psychology, 32*(3), 492-504.

Tardif, T. (2006). The importance of verbs in Chinese. In P. Li, H. Tan, E. Bates & O. J. L. Tzeng (Eds.), *Handbook of East Asian Psycholinguistics* (Vol. Volume 1. Chinese, pp. 124-135). Cambridge: Cambridge University Press.

Tardif, T., Gelman, S. A., & Xu, F. (1999). Putting the "noun bias" in context: A comparison of English and Mandarin. *Child Development, 70*(3), 620-635.

Tardif, T., Shatz, M., & Naigles, L. R. (1997). Caregiver speech and children's use of nouns versus verbs: A comparison of English, Italian, and Mandarin. *Journal of Child Language, 24*, 535-565.

Tepper, J. A., Powell, H. M., & Palmer-Brown, D. (2002). A corpus-based connectionist architecture for large-scale natural language parsing. *Connection Science, 14*(2), 93-114.

Tesar, B. (2004). Using inconsistency detection to overcome structural ambiguity. *Linguistic Inquiry, 35*(2), 219-253.

Thal, D., & Flores, M. (2001). Development of sentence interpretation strategies by typically developing and late-talking toddlers. *Journal of Child Language, 28*, 173-193.

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2004). Semantic generality, input frequency and the acquisition of syntax. *Journal of Child Language, 31*, 61-99.

Thomason, S. G., & Kaufman, T. (1988). *Language Contact, Creolization, and Genetic Linguistics*. Berkeley: University of California Press.

Tomasello, M. (1992). *First Verbs. A Case Study of Early Grammatical Development*. Cambridge: Cambridge University Press.

Tomasello, M. (1998). The return of constructions. *Journal of Child Language, 25*, 431-442.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition, 74*, 209-253.

Tomasello, M. (2000). First steps towards a usage-based theory of language acquisition. *Cognitive Linguistics, 11*(1-2), 61-82.

Tomasello, M. (2004). What kind of evidence could refute the UG hypothesis? *Studies in Language, 28*(3), 642-645.

Tomasello, M., & Abbot-Smith, K. (2002). A tale of two theories: response to Fisher. *Cognition, 83*, 207-214.

Tomasello, M., & Akhtar, N. (2003). What paradox? A response to Naigles (2002). *Cognition, 88*, 317-323.

Tomasello, M., Akhtar, N., Dodson, K., & Rekau, L. (1997). Differential productivity in young children's use of nouns and verbs. *Journal of Child Language, 24*, 373-387.

Tomlin, R. S. (1986). *Basic Word Order. Functional Principles*. London: Croom Helm.

Traxler, M. J. (2002). Plausibility and subcategorization preference in children's processing of temporarily ambiguous sentences: Evidence from self-paced reading. *The Quarterly Journal of Experimental Psychology, 55A*(1), 75-96.

Trueswell, J. C., Sekerina, I., Hill, N. M., & Logrip, M. L. (1999). The kindergarten-path effect: studying on-line sentence processing in young children. *Cognition, 73*, 89-134.

Tsunoda, T. (1981). Split case-marking patterns in verb-types and tense/aspect/mood. *Linguistics, 19*, 389-438.

Tyler, L. K., Russell, R., Fadidi, J., & Moss, H. E. (2001). The neural representation of nouns and verbs: PET studies. *Brain, 124*(8), 1619-1634.

Ueno, M., & Polinsky, M. S. (2002). *Maximizing processing in an SOV language: A corpus study of Japanese and English.*Unpublished manuscript.

Uziel-Karl, S. (2006). Acquisition of verb argument structure from a developmental perspective: Evidence from Child Hebrew. In N. Gagarina & I. Gülzow (Eds.), *The Acquisition of Verbs and their Grammar. The Effect of Particular Languages* (pp. 15-40). Dordrecht: Springer Verlag.

van der Velde, F., van der Voort van der Kleij, G. T., & de Kamps, M. (2004). Lack of combinatorial productivity in language processing with simple recurrent networks. *Connection Science, 16*(1), 21-46.

Van Everbroeck, E. (2003). Language type frequency and learnability from a connectionist perspective. *Linguistic Typology, 7*(1), 1-50.

Van Everbroeck, E. (1999). Language type frequency and learnability: A connectionist approach. In *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 755-760). Mahwah, NJ: Lawrence Erlbaum.

Van Everbroeck, E. (in preparation). *Modeling cross-linguistic aphasia.*Unpublished manuscript.

Vennemann, T. (1975). An Explanation of Drift. In C. N. Li (Ed.), *Word Order and Word Order Change* (pp. 269-305). Austin, TX: University of Texas Press.

Waskan, J. A. (2001). A critique of connectionist semantics. *Connection Science, 13*(3), 277-292.

Weekes, B. S., Chan, A., Kwok, J. S. W., Hai Tan, L., & Jin, Z. (2004). AoA effects on Chinese language processing: An fMRI study. *Brain and Language, 91*, 33-34.

Weiss, D. J., & Newport, E. L. (2006). Mechanisms underlying language acquisition: Benefits from a comparative approach. *Infancy, 9*(2), 241-257.

Wilson, S. (2003). Lexically specific constructions in the acquisition of inflection in English. *Journal of Child Language, 30*, 75-115.

Xu, F., & Tenenbaum, J. B. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental Science, 10*(3), 288-297.

Yang, C. (2006). Grammar acquisition via parameter setting. In P. Li, H. Tan, E. Bates & O. J. L. Tzeng (Eds.), *Handbook of East Asian Psycholinguistics* (Vol. Volume 1. Chinese, pp. 136-147). Cambridge: Cambridge University Press.

Yang, C. L., Gordon, P. C., & Hendrick, R. (2006). The comprehension of coreference in Chinese discourse. In P. Li, H. Tan, E. Bates & O. J. L. Tzeng (Eds.), *Handbook of East Asian Psycholinguistics* (Vol. Volume 1. Chinese, pp. 257-267). Cambridge: Cambridge University Press.

Year, J. (2003). Sentence processing within the Competition Model. *Working Papers in TESOL and Applied Linguistics, 3*(1).

Zhang, Y., Wu, N., & Yip, M. (2006). Lexical ambiguity resolution in Chinese sentence processing. In P. Li, H. Tan, E. Bates & O. J. L. Tzeng (Eds.), *Handbook of East Asian Psycholinguistics* (Vol. Volume 1. Chinese, pp. 268-278). Cambridge: Cambridge University Press.

Zhiming, B. (2001). The origins of empty categories in Singapore English. *Journal of Pidgin and Creole Languages, 16*(2), 275-319.

Zhiming, B. (2005). Diglossia and register variation in Singapore English. *World Englishes, 25*, 105-114.

Zhou, X., Ostrin, R. K., & Tyler, L. K. (1993). The noun-verb problem and Chinese aphasia: Comments on Bates et al. (1991). *Brain and Language, 45*, 86-93.

Zhu, L., & Gigerenzer, G. (2006). Children can solve Bayesian problems: The role of representation in mental computation. *Cognition, 98*, 287-308.