

Case, Word Order, and Language Learnability: Insights from Connectionist Modeling

Gary Lupyan (il24@cornell.edu)
Morten H. Christiansen (mhc27@cornell.edu)

Department of Psychology
Cornell University
Ithaca, NY 14853 USA

Abstract

How does the existence of case systems, and strict word order patterns affect the learnability of a given language? We present a series of connectionist simulations, suggesting that both case and strict word order may facilitate syntactic acquisition by a sequential learning device. Our results are consistent with typological data concerning the frequencies with which different type of word order patterns occur across the languages of the world. Our model also accommodates patterns of syntactic development across several different languages. We conclude that non-linguistic constraints on general sequential-learning devices may help explain the relationship between case, word order, and learnability of individual languages.

Introduction

In language acquisition, children are faced with many formidable tasks, yet they normally acquire most of their native language within the first five years of life. One of the most difficult of these tasks involves mapping a sequence of words onto some sort of interpretation of what that sequence is supposed to mean. That is, in order for the child to understand a sentence, she needs to determine the grammatical roles of the individual words so that she can work out who did what to whom. Although the children appear to bring powerful statistical learning mechanisms to bear on the acquisition tasks (e.g., Saffran, Aslin, & Newport 1996), the existence of linguistic universals common across radically different languages (Greenberg 1963) points to the presence of innate constraints on such learning. Without such constraints, it becomes difficult to explain why there are few, if any, Object-Subject-Verb (OSV) languages (van Everbroeck, 1999) even though in principle such a language appears to be as good as any other. In this paper, we propose that these constraints may arise from non-linguistic limitations on the sequential learning of statistical structure, and examine how this perspective may shed light on how children learn to map the words in sentences onto their appropriate grammatical roles. There are two major ways in which languages signal syntactic relationships and grammatical roles—word order (WO),

and case markings. In a strict WO language like English, declarative sentences follow a Subject-Verb-Object (SVO) pattern. It is the occurrence of the subject in the first position, and the object in the second, that allows the hearer to comprehend who did what to whom. In contrast, languages such as Russian or Japanese allow multiple word orders and rely on case markings to disambiguate subjects from objects. For instance, *Masha lubit Petyoo* (SVO), *Petyoo lubit Masha* (OVS), and *Lubit Petyoo Masha* (VOS) are all grammatical in Russian and all mean *Mary loves Peter* (albeit with different emphases on the constituents), due to the nominative *-a*, and accusative *-u* case markers.

While long-standing theories describe acquisition of language through an innate language acquisition device (e.g., Pinker, 1995), an alternative approach that is gaining ground is the adaptation of linguistic structures to the human brain rather than vice versa (e.g., Christiansen, 1994; Kirby, 1998). On this account, language universals may reflect non-linguistic cognitive constraints on learning and processing of sequential structure, rather than constraints prescribed by an innate universal grammar. Previous work has shown that sequential-learning devices with no language-specific biases are better able to learn more universal aspects of language as compared to aspects encountered in rare languages (e.g., Ellefson & Christiansen, 2000; Christiansen & Devlin, 1997; Van Everbroeck, 1999, 2001).

Here, we examine the ways in which case markings and word order may function as cues for a sequential learning device acquiring syntactic structure. In simulation 1, we model different word orders, and hypothesize that typologically common languages should be easier to learn by a sequential-learning device than the more rare ones. We expand on this idea in simulation 2 by studying the performance of networks trained on languages of varying degrees of case markings and flexibility. Finally, in simulation 3, we establish that our trained networks are able to mimic syntactic performance of children learning English, Italian, Turkish, and Serbo-Croatian (Slobin and Bever, 1982).

Acquisition of Word Order

Generative linguists have long relied on parameter setting to explain how children acquire the distinct patterns of their native language. For instance, it has been assumed that the way a child knows to generate SVO and not SOV English sentences is through the setting of a VO/OV parameter (Neeleman, 1994). This account has been unsatisfactory because it does not account for many observed correlations; for instance, OV languages typically have flexible word orders (Koster, 1999). More generally, parameter theory has been largely unable to account for the asymmetries and patterns in the distribution of world languages. Why, for instance, are the most common word orders SOV, SVO, and VSO (Greenberg 1963: Universal 1)? Why do verb-final languages almost always have a case system (Greenberg 1963: Universal 41)? And even more fundamentally, why do case languages have flexible word orders to begin with? It is our position that these observations can be at least partially accounted for by examining the learnability of languages from the viewpoint of sequential learning.

Generative linguistics also leaves largely unexplained the process children use to actually set the parameters. With regard to word orders, an explanation espoused by Pinker (e.g., 1995) involves the so-called Subset Principle. According to the Subset Principle, children take the most conservative strategy and so, by default, assume a fixed order. Alternative word orders are only accepted if a child is exposed to these orders, at which time a free word order parameter gets switched on. Under this assumption, FWO languages are predicted to be more difficult to learn. Although the idea that all languages are initially approached as having strict word-order (SWO) was popular in the sixties and seventies (Slobin, 1966), Slobin and Bever (1982) conclude that the primacy assigned to word order was unduly influenced by languages such as English.

There is ample evidence that children learning a strict word-order language such as English never leap to the conclusion that it is a free-word order language (Pinker 1995). While Pinker has used this evidence for reinforcing parameter-setting—the reason children never leap to such conclusions is because a word-order parameter has been set—we suggest an alternative explanation. Simply, children learning English generally do not produce non-SVO sentences because non-SVO sentences are incomprehensible in English. In the absence of case markings, *Kicked John Bill* is ambiguous as to who did the kicking. Children learning English, use the statistical properties of the language to learn that word order is a reliable cue to syntactic relationships. Children learning a case-based language such as Russian, make a similar observation about case markings. This view obviates the need for a default strategy. What is important is that there exist some set of cues to in-

dicating syntactic relationships—there is nothing inherently special about word order or case markings. In short, we posit that a major reason for the observable asymmetries among the world’s languages is that certain patterns make a language more easily learnable by a general sequential-learning device, ensuring the proliferation of such a language in the human population.

Simulation 1: Exploring the Learnability of Case and Word Order

In the view that the frequency of certain WOs is correlated with their learnability, we hypothesized that typologically rare languages will be more difficult to learn by a sequential-learning device than the more common languages. To test this prediction, we trained simple recurrent networks (SRNs: Elman, 1990) on a total of 14 artificial grammars, reflecting the 6 possible strict word orders (SWO) and a flexible word order (FWO), with or without the presence of case markings.

Method

Networks Ten SRNs were used in each condition. The networks were initialized with random weights in the interval $[-0.1, 0]$.¹ Each input to the networks consisted of a distributed representation of a word, spliced with a case marker. Words were represented by 20-unit randomly generated bit-vectors. Although some vectors were bound to be close in the representation space, random assignment to words assured that any such interaction would not bias the results. Having words represented by random vectors may seem odd considering the complex phonology that underlies human languages. However, for present purposes such a representation seems to work just as well as phonological (e.g., van Everbroeck 2001), while dramatically decreasing training time. Case markings (nominative, accusative, dative, and genitive) were represented by a four-bit vector appended to the word vector. This made for a total of 24 input units. There were seven output units, corresponding to the grammatical roles the network was supposed to predict: subject, direct object, indirect object, genitive noun, verb, or end-of-sentence. In all simulations, the learning rate was set to 0.1, and momentum to 0.01. Each SRN had 30 hidden units and 30 context units.

Materials The lexicon contained 300 nouns and 100 verbs. This noun-to-verb ratio is generally consistent with human languages (e.g., British National Corpus). The verbs were evenly divided into intransitive, transitive, and ditransitive categories. As illustrated in Table 1, each grammar included three

¹It was found that the slightly inhibitory starting weights provided for better performance across the board. A similar conclusion was reached by van Everbroeck (2001).

Table 1: A Sample SOV Grammar Used to Generate Training Corpora

| |
|---|
| S → Intransitive [.35] Transitive [.35] Ditransitive [.3] |
| Intransitive → NP-nom V-intrans |
| Transitive → NP-nom NP-acc V-trans |
| Ditransitive → NP-nom NP-acc NP-dat V-ditrans |
| NP → N N N-gen [.25] |

types of sentences: intransitive, transitive, and ditransitive. A sentence consisted of noun phrases (NP) and one of three verb classes. Twenty-five percent of noun phrases contained a noun in the genitive form (e.g., *John’s brother*). The simplest sentence generated by such a grammar was a simple intransitive: e.g., *John walks*. The most complex sentence contained 7 words: *Mary’s friend gave Peter’s key [to] John’s brother*. A fully flexible grammar was identical to the strict WOs except the order within each element was randomly varied from sentence to sentence. In an effort to model the languages as naturalistically as possible, we modeled genitives based on Greenberg’s (1963) universal 2: in typically prepositional languages (SVO and VSO) genitives generally follow the governing noun, while in postpositional languages (SOV), the reverse is true. We modeled the remaining three word orders with genitives following the noun. We also added a genitive case-marking to SWO-no case languages. Without this, it was impossible for the networks to discern governing nouns from genitives. This addition is motivated by the observation that even normally case-less languages have some form of genitive case markings (e.g., in English: *Mary’s house*) (van Everbroeck, 2001).

Procedure We used a crossover design of 7 word orders (6 strict, and one flexible), by two case conditions (with or without case) resulting in 14 training corpora. For each condition, we generated 3,000 random sentences of the appropriate order. Such a corpus occupies a very small part of the possible sentences that can be generated by the corpus. For instance, 9 million different sentences are possible for a transitive SOV configuration (300 x 300 x 100).

The networks were trained for 100,000 sweeps (input/output pairs), corresponding to about 7 passes through the corpus. During each training sweep, the network was presented with a word, and depending on the condition, a case marking. A corpus of 200 novel sentences was created for testing. In the testing corpus, 50% of words were completely new—ones to which the network has never been exposed. Performance was measured by assessing the network’s ability to map a given word to its correct grammatical role. During testing, the network’s highest-activated output unit was compared to the expected output. If the units matched, the word was marked

Table 2: Network performance and Language Distributions

| Word Order | Words Correct – No Case Condition (%) | Attested Frequency (%) |
|------------|---------------------------------------|------------------------|
| SOV | 90 | 51 (most w/cases) |
| VOS | 85 | 8 |
| OVS | 80 | 0.75 |
| OSV | 74 | 0.25 |
| Flexible | 65 | 0 (all w/cases) |

Note. Attested language frequencies taken from Van Everbroeck (1999).

as being correctly mapped.

It may seem that providing the networks with direct mapping from word to grammatical category is not ecologically valid. After all, it has long been recognized that kids are not given sufficient ostensive cues to syntactic relationships and word meanings. No one explains to the child after each encountered sentence that word *A* referred to the “do-er” and word *B* to the “do-ee”. However, Tomasello and colleagues have shown that children are able to use pragmatic cues such as eye gaze to help figure out which object is being referred to (Tomasello & Akhtar, 1995). Twenty-four month olds show understanding of adult intentions in inferring meanings of novel verbs (Tomasello & Barton, 1994), and 18-month old children are able to learn new words in non-ostensive contexts (Tomasello, Strosberg, & Akhtar, 1996). Such use of pragmatic cues enables children to map words onto meanings and correctly infer who did what to whom. Considering that our networks live in a purely linguistic world, the method of direct mapping seems reasonable.

Results

All networks trained in the case condition were able to map 100% of the words to the correct categories. When case was not available, the network performance roughly corresponded to attested language frequencies (Table 2). Only two caseless WOs obtained nearly perfect performance: SVO and VSO (99%). There, however, appears to be a discrepancy. While SOV is the most common WO, it is outperformed by both SVO and VSO. According to Greenberg’s Universal 41, however, the great majority of SOV languages have case, and most caseless languages turn out to be either SVO or VSO (e.g., English, Welsh). This finding supports our learnability hypothesis: verb-final languages presumably have a case system because reliance on WO results in suboptimal learnability.

The likely reason SOV-no case grammars did not achieve perfect accuracy is because they contained two unmarked nouns prior to the verb. Since the networks learn to map different types of verbs to different argument constructions, verb-final grammars are at a disadvantage—in these grammars the

grammatical role that provides the most information about what is to come, is received last (van Everbroeck, 2001). The poor performance of VOS is due to the intervening indirect subject in ditransitive sentences. We should also note that even though FWO-no case languages perform poorly, their performance is consistently above chance. This can be explained by the networks’ learning to map familiar verbs to intransitive, transitive or ditransitive word schemas.

These simulation results confirm the idea that FWO-case languages are no more difficult to learn than common SWO-no case languages such as SVO and VSO, counter to predictions of the subset principle. The difficulty associated with learning a FWO language without case markings is underscored by typological evidence, suggesting that such languages use case markings to signal grammatical relationships (Payne 1992).

Simulation 2: The Impact of Case on Word Order Flexibility

In natural languages, case markings are not wholly deterministic. For instance, Slavic languages such as Russian and Serbo-Croatian, contain a number of nouns which, perhaps for historical reasons, do not take case markings. Additionally, because these markings often take the form of suffixes, they change the phonology of words. This results in potential phonological ambiguity. For instance, in Russian *stali* is either the genitive form of *steel* or a conjugated verb meaning *we stopped*. By examining the effects of varying cases on different word orders, we hoped to show that (1) even probabilistic case markings improve performance for FWO languages, and (2) case markings do not improve performance for languages that already rely on WO.

Method

Networks Ten SRNs were used for each condition. The initial conditions and training details were the same as in Simulation 1.

Materials We generated five artificial grammars varying on the salience of case markings—from only genitive markings, to full case markings. A grammar with case marked 50% of the time corresponded to a language in which 50% of case markings are possibly phonologically ambiguous, or a language in which certain of nouns do not take on case markings. Five more grammars varying on strictness of word order—from a completely flexible order, to a completely strict one (SVO). The word orders approximated distributions found in natural languages (Italian, and Turkish: Slobin & Bever 1982); Polish: Jacennik & Dryer 1992). The two conditions were crossed, for a 5x5 matrix. As in simulation 1, 3000 sentences were generated for each condition.

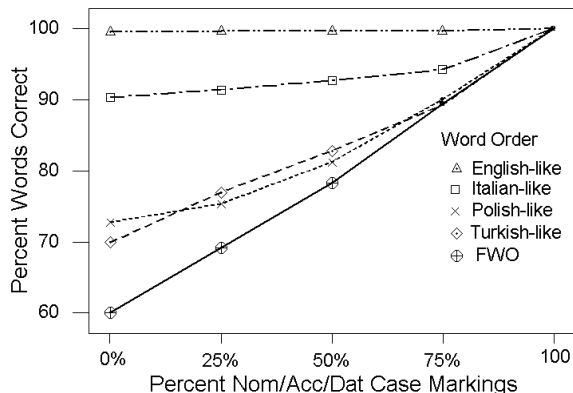


Figure 1: Network performance in simulation 2 for increasing degrees of case markings as a function of word order.

Procedure Each group of networks had case cues added to the sentences based on case condition. The testing proceeded as in Simulation 1.

Results

As expected, SWO languages such as English and Italian were little-benefited by case (Figure 1). In contrast, the probabilistic addition of case markings to FWO languages consistently improved performance. The slightly lower performance of Italian is due to it having a more flexible word order than English (see Table 3). To compensate for possible ambiguities, Italian relies heavily on prosodic and contextual information (Slobin & Bever, 1982) which was not available to our networks. In summary, the precise nature of the cue: case, or WO, does not seem to matter. Neither needs to be primary.

Simulation 3: Interactions between Case and Word Order Flexibility in Development

In this simulation, we demonstrate that networks trained on corpora similar to those used in simulation 2 are able to mimic syntactic performance of children learning English, Italian, Turkish, and Serbo-Croatian. Slobin and Bever (1982) tested 48 children divided into 8 age groups (24-52 months). Each child was tested on their ability to demonstrate familiar actions (e.g., *scratch*, *bump*, *pick up*) using familiar toy animals after hearing a transitive language in their native language. The authors hypothesized that Turkish, English, and Italian-speaking children would have the easiest time due to the consistent, unambiguous case markings available in the case of Turkish, and the consistent word-order information available in English and Italian. Children

Table 3: Word Order Distributions for Simulation 3

| Language | Words Order | Cases |
|----------------|---|---|
| English | 100% SVO | Genitive only |
| Italian | 82% SVO, 2% SOV, 11% VSO, 5% OVS | Genitive only |
| Serbo-Croatian | 55% SVO, 16% SOV, 16% VSO, 3% VOS, 2% OVS, 8% OSV | Full for non-SVO For SVO: 55% nom, 55% acc, 100% dat, 70% gen |
| Turkish | 48% SOV, 25% SVO, 13% OVS, 8% OSV, 6% VSO | Full |

acquiring Serbo-Croatian would have a more difficult time due to its more ambiguous case markings, requiring them to pay attention to word-order as well as cases.

Method

Networks The networks and training details were identical to simulation 2. We used 12 SRNs in each condition, mirroring the number of subjects used by Slobin and Bever (1982).

Materials We created 4 types of grammars motivated by the four languages used in the study. English was modeled as being 100% SVO, and having only genitive case markings. The word orders for the remaining languages were modeled based on the data provided by Slobin and Bever’s (1982) corpus of adult speech, reflecting the linguistic input available the children.

Although Turkish does not have an explicit nominative case, it was found that such a marker was necessary in this simulation. In the absence of semantic information and case markings, the networks must rely on the syntactic position of a word to correctly identify its category. However, in a relatively FWO language such as Turkish, this information is ambiguous. Without a nominative case, both verbs and subjects are unmarked, and the network naturally has trouble telling them apart. In contrast to these networks, children rely on semantic information, in addition to syntax, to tell apart verbs and nouns. In other words, a Turkish child knowing the meanings of “dog” and “sniff”, will not confuse the two even when “dog” is an unmarked agent in the sentence.

Serbo-Croatian has all four of the cases we were modeling, however, only masculine and feminine nouns take on accusative and nominative markings. Sentences containing one or more neuter nouns are typically ordered as SVO. We did not have data on the proportion of neuter nouns in Serbo-Croatian or the percentage of SVO sentences containing such nouns. It was estimated that about 55% of SVO sentences would contain one such noun, therefore case

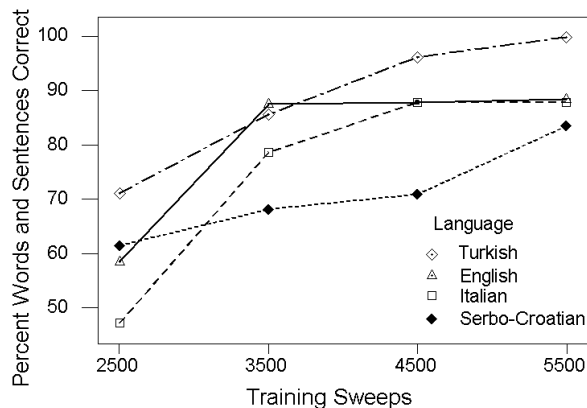


Figure 2: The pattern of performance across training for Turkish, English, Italian, and Serbo-Croatian in simulation 3

Table 4: Percentage correct performance for grammatical sentences in a given language in the Slobin and Bever (1982) study.

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

markings were deleted from 55% of nouns in SVO sentences. Serbo-Croatian neuter nouns do have dative case-markings, hence the datives are marked 100% of the time. However, plural neuter nouns are not declined in genitive constructions. If plural genitive nouns are used an estimated 30% of the time, then 70% of SVO sentences will have genitive case markers.

Procedure Training proceeded as in simulation 1. The extent of training was varied for networks corresponding to different age groups. Testing was done following the procedure employed by Slobin and Bever (1982). We used transitive sentences using only words which the networks have seen during training. Performance was quantified by measuring the percentage of subjects and objects the network identified correctly, and averaging the data with the overall percentage of words correctly identified.

Results

The networks’ performance (Figure 2) closely matched Slobin and Bever’s (1982) data (Table 4). As predicted, networks trained and tested on Turkish had the easiest time mapping words onto grammatical roles. Networks trained on Serbo-Croatian,

had the most difficult time, highlighting the higher processing-cost associated with having to pay attention to WO and case markings. This pattern of results runs counter to the subset principle since the latter predicts case-languages to be more difficult to acquire. Performance on Italian was slightly worse than on English, reflecting the more flexible WO of Italian. It is predicted that with the addition of prosodic and semantic cues, the performance of Italian would more closely parallel that of a fully SWO language such as English.

Conclusion

Our findings confirm that learnability of languages may be a major factor in the frequency of certain language types. In the view of language as an organism (Christiansen, 1994), languages which are easily learnable by the human sequential-learning device proliferate, while languages not easily learnable die out or never come into existence. Our simulations suggest that all that is needed to learn syntactic relations is a reliable cue: case, or word order—neither needs to be primary. As such, no parameter-setting or subset principle is needed to account for the data. These results also provide added support for a connectionist approach to studying acquisition and evolution of language.

The simulations described here have several notable limitations. The sentences used for training were admittedly simple. Although simple intransitive, transitive, and ditransitive sentences are very frequent in speech, natural languages are rife with more complex structures such as relative clauses and embedding. Offsetting the simple grammars, however, were the limited cues available to the networks, which relied solely on distributional information of grammatical categories. In contrast, children routinely use semantics and prosodic cues, and even more subtle cues such as differential word length of nouns and verbs (Cassidy & Kelly, 1991—see Christiansen & Dale, 2001, for a review). It is therefore quite remarkable that relying only on word order or case, the performance of the networks was near-perfect for common language types.

References

- British National Corpus*. Located at: www.hcu.ox.ac.uk/BNC
- Cassidy, K.W. & Kelly, M.H. (1991). Phonological information for grammatical category assignments. *Journal of Memory and Language*, 30: 348-369.
- Christiansen, M.H. & Dale, R.A.C. (2001). Integrating distributional, prosodic and phonological information in a connectionist model of language acquisition. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 220-225). Mahwah, NJ: Lawrence Erlbaum.
- Christiansen, M.H. & Devlin, J.T. (1997). Recursive inconsistencies are hard to learn: A connectionist perspective on universal word order correlations. In *The Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. (pp. 113-118). Mahwah, NJ: Erlbaum
- Christiansen, M. H. (1994). Infinite languages, finite minds: Connectionism, learning and linguistic structure. Unpublished doctoral dissertation, Centre for Cognitive Science, University of Edinburgh, U. K.
- Ellefson, M.R. & Christiansen, M.H. (2000). Subjacency constraints without universal grammar: Evidence from artificial language learning and connectionist modeling. In *The Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. (pp. 645-650). Mahwah, NJ: Erlbaum
- Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, 14, 179-211.
- Greenberg, J. H. (1963) Some universals of grammar with particular reference to the order of meaningful elements. In: *Universals of language*. Greenberg, J.H. (ed.). Cambridge, MA.: MIT Press.
- Jacennik, B. & Dryer, M.S. (1992). Verb-subject order in Polish. In *Pragmatics of word order flexibility*, ed. Payne, D. L. Amsterdam: John Benjamins.
- Kirby, S. (1998). Language evolution without natural selection: From vocabulary to syntax in a population of learners. *Edinburgh Occasional Paper in Linguistics*, EOPL-98-1.
- Koster, J. (1999) The word orders of English and Dutch: Collective vs. individual checking. In: *Groninger Arbeiten zur germanistischen Linguistik*, Abraham, W. (ed.), University of Groningen, Groningen. 1-42.
- Neeleman, A. (1994) *Complex predicates*. PhD Dissertation Utrecht University, Utrecht.
- Payne, D. L.(1992). Verb-Subject Order in Polish. In *Pragmatics of Word Order Flexibility*, Amsterdam: John Benjamins.
- Pinker, S. (1995). Language acquisition In *An invitation to cognitive science. Vol 1: Language* Gleitman, L.R. & Liberman, M. (Eds.) Cambridge, MA.: MIT Press.
- Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926-1928.
- Slobin, D. I. (1966). The acquisition of Russian as a native language. in F. Smith and G. A. Miller (eds.), *The genesis of language: A psycholinguistic approach*. Cambridge, MA: MIT Press
- Slobin, D. I., & Bever, T. G. (1982). Children use canonical sentence schemas: A crosslinguistic study of word order and inflections. *Cognition* 12: 229-265
- Tomasello, M. & Akhtar, N. (1995). Two-year-olds use pragmatic cues to differentiate reference to objects and actions. *Cognitive Development*, 10, 201-224.
- Tomasello, M. & Barton, M. (1994). Learning words in non-ostensive contexts. *Developmental Psychology*, 30, 639-650.
- Tomasello, M., Strosberg, R., & Akhtar, N. (1996). Eighteen-month-old children learn words in non-ostensive contexts. *Journal of Child Language*, 23, 157-176.
- 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: Erlbaum
- Van Everbroeck, E. (2001). *Language type frequency and learnability from a connectionist perspective*. Submitted Manuscript.