The ontogeny of scale-free syntax networks through language acquisition

By Bernat Corominas-Murtra, Sergi Valverde and Ricard V. Solé

[1] *ICREA-Complex Systems Lab, Universitat Pompeu Fabra, Barcelona Biomedical Research Park, Dr. Aiguader 80, 08003 Barcelona, Spain*
[2] *Santa Fe Institute, 1399 Hyde Park Road, New Mexico 87501, USA*

The evolution of human language allowed the efficient propagation of nongenetic information, thus creating a new form of evolutionary change. Language development in children offers the opportunity of exploring the emergence of such complex communication and is considered as a window to the transition from protolanguage to language. Here we analyze available information from the CHILDES database and we study the emergence of syntax in terms of complex networks where words are connected through syntactic links, providing a global view of the organization of syntactic relations. A previously unreported, sharp transition is shown to occur at ≈ 2 years from a tree-like structure to a scale-free, small world syntax network. The nature of such transition supports the presence of an innate component pervading the emergence of full syntax.

Keywords: Language evolution, language acquisition, syntax, complex networks, small worlds

1. Introduction

Although human language stands as the greatest transitions in evolution (Maynard-Smith and Szathmàry, 1997) its exact origins remain a source of debate. Since language does not leave fossils, our windows to its evolution are limited and require extrapolation from different sources of indirect information (Bickerton, 1990)). Among the relevant questions to be answered is the leading mechanism driving language change: Is language the result of natural selection? The use of population models under noisy environments is consistent with such selection-driven scenario (Hurford, 1989; Nowak and Krakauer, 1999; Komarova and Niyogi, 2004). Other approaches have suggested the importance of communicative constraints canalizing the possible paths followed by language emergence (Ferrer-i-Cancho and Solé, 2003). Supporting such communication system there has to be a symbolic system which it has been for some authors the core question (Deacon, 1997). Finally, a rather different approach focuses on the evolution of the *machine* that generates human language. The most remarkable trait of such *machine*, is the possibility of generating infinite structures (Humboldt, 1999; Chomsky, 1957; Hauser et al., 2002) in a recursive fashion. The evolution of such ability alone, beyond its potential functionality is considered by some authors the main problem in language evolution (Hauser et al., 2002).

An important component of this debate is related to the tempo and mode of language acquisition in children. Actually, it has been pointed out that child language (together with ape and pidgin languages) may help understanding language origins

(Maynard-Smith and Szathmàry, 1997; Bickerton, 1990). Children are able to construct complex sentences by properly using phonological, syntactic and semantic rules in spite that no one teaches them. Specifically, they can generate a virtually infinite set of grammatically correct sentences in spite that they have been exposed to a rather limited number of input examples. And although the lexicon shows a monotonous growth as new words are learned, the pattern of change in syntactic organization is strongly nonlinear, with a well-defined transition from babbling, to single words, to the rude two-words grammar to a fully, complex adult grammar (Radford, 1990). How can children acquire such huge set of rules? Are there some specific, basic rules predefined as a part of the biological endowment of humans? If so, some biological program of rules (the Universal Grammar) should guide the acquisition process. In this way, models assuming a constrained set of accessible grammars have shown that final states (i.e., an evolutionary stable complex grammar) can be reached under a limited exposure to the right inputs (Komarova et al., 2001), (Niyogi, 2006). However, we cannot deny the fact that important features language acquisition process can be obtained by appealing only general purpose mechanisms of learning (Macwhinney, 2005; Newport, 1990; Elman, 1993).

The experimental analysis of language acquisition data is an important source of validation of different hypotheses about language origins and organization, as far as any reasonable theory of language should be able to explain how it is acquired. Here we analyze this problem by using a novel approximation to acquisition based on a global, network picture of syntax. We present evidence for the existence of predefined combinatorial features which are triggered at some point of the acquisition process, thus supporting the presence of some innate component underlying the combinatorial power of human grammar.

2. Syntactic Networks: Data sets and Graph construction

Regarding English syntax, we find several well-known acquisition stages (Radford, 1990). The first stage is the so-called *Babbling*, where only single phonemes or short combinations of them are present. This stage is followed by the *Lexical spurt*, a sudden lexical explosion where the child begins to produce a large amount of isolated words. Such stage is rapidly overcome by the *two words stage*, where short sentences of two words are produced. In this period, we do not observe the presence of functional items nor inflectional morphology. Later, close to the two-years age, we can observe the *syntactic spurt*, where more-than-two word sentences are produced.

In this paper we analyse raw data obtained from children utterances, from which we extract a global map of the pattern of the use syntactic relations among words. In using this view, we look for the dynamics of large-scale organization of the use of syntax. This can be achieved by means of complex networks techniques, by aggregating all syntactic relationships within a graph. Recent studies have shown that networks reveal many interesting features of language organization (Hudson, 2006; Ferrer-i-Cancho and Solé, 2001; Melçuck, 1989; Ke, 2007; Sigman and Cecchi, 2002; Ferrer-i-Cancho et al., 2004) at different levels. These studies uncovered new regularities in language organization but so far none of them analyzed the emergence of syntax through language acquisition. Here we study in detail a set of quantitative, experimental data involving child utterances at different times of their development.

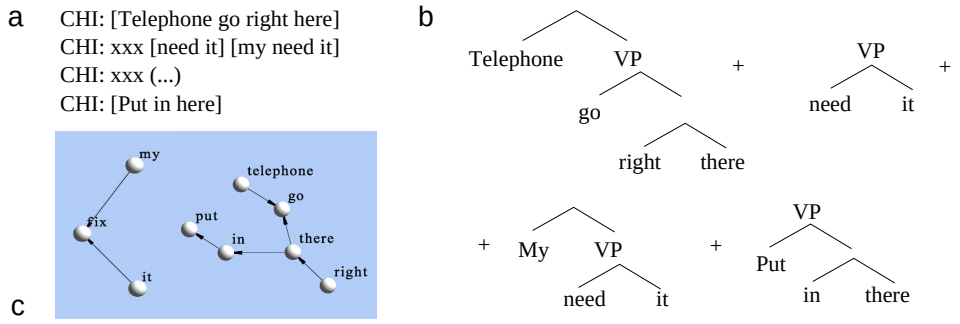

Figure 1. Building the networks of Syntax Acquisition. First (a) we identify the structures in child's productions using the lexico-thematic nature of early grammars (Radford, 1990), see (Corominas-Murtra, 2007). Then (b) a basic constituency analysis is performed, assuming that the semantically most relevant item is the head of the phrase and that the verb in finite form (if any) is the head of the sentence. Finally (c) a projection of the constituent structure in a dependency graph is obtained.

Formally, we define the *syntax network* $\mathcal{G} = \mathcal{G}(\mathcal{W}, E)$ as follows (1). Using the lexicon at any given acquisition stage, we obtain the collection of words $W_i (i = 1, ..., N_w)$, every word is a node $w_i \in \mathcal{G}$. There is a connection between two given words provided that they are syntactically linked†. The set of links E describes all the syntactic relationships in the corpus. For every acquisition stage, we obtain a syntactic network involving all the words and their syntactic relationships. The structure of syntax networks will be described by means of the *adjacency matrix* $A = [a_{ij}]$ with $a_{ij} = 1$ when there is a link between words w_i and w_j and $a_{ij} = 0$ otherwise.

Our corpora are extracted from a recorded session where a child speaks with adults spontaneously. We have collected them from the *CHILDES Database* (Macwhinney, 2000)†. Specifically, we choose Peter's corpora (Bloom et al., 1974, 1975). Time intervals are regular and the corpora spans a time window that can be considered large enough to catch global properties. Although this is a given sample, it seems to be a fairly good representation of common average patterns.

The data-set studied here includes the first eleven stages of Peter's corpora. The time period covers all the early, key changes in language acquisition, from non-grammatical to grammatical stages. Each corpus contains several conversations among adult investigators and the child. However, the raw corpus must be manipulated for our needs. In (Corominas-Murtra, 2007) we present a detailed description of the criteria and rules followed to pre-process the raw data. The main features of the procedure are in (fig.(1) and can be summarized as:

1. Select only child's productions rejecting imitations, onomatopoeia's and undefined lexical items.

2. Identify the *structures*, i.e., the minimal syntactic constructs.

† Recall that the net is defined as the projection of the constituency hierarchy. Thus, the *link* has not an ontological status under our view of syntax

† http://talkbank.org

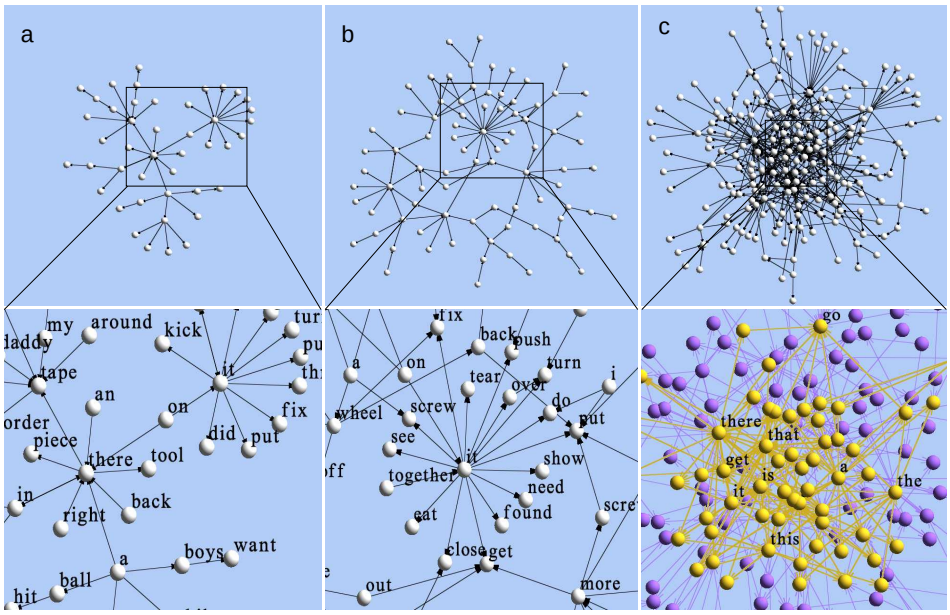

Figure 2. Transitions from tree-like graphs to scale-free syntax graphs through the acquisition process. Here three snapshots of the process are shown, at (a) 25 months, (b) 26 moths and (c) 28 months. Although a tree-like structure is shown to be present through the pre-transition (a-b) a scale-free, much more connected web suddenly appears afterward (c), just two months later. The lower pictures indicate how the hubs are organized and their nature. There is a critical change at the two-years age marked by a widespread re-organization of the network. Prior to the transition, semantically degenerated elements (such as *it* act as hubs. Key words essential to adult syntax are missing in these early stages. After the transition, the hubs change from degenerated to functional items (i.e., *a* or *the*. In (f) we highlight the core of this network using yellow nodes and links.

3. Among the selected structures, we perform a basic analysis of constituent structure, identifying the verb in finite form (if any) in different phrases.

4. Project the constituent structures into lexical dependencies. This projection is close to the one proposed by (Hudson, 2006) within the framework of the network-based *Word Grammar*. †:

5. Finally, we build the graph by following the dependency relations in the projection of the syntactic structures found above. Dependency relations allow us to construct a syntax graph.

With this procedure, we will obtain a graph for every corpus. The resulting graphs will be our object of study in the following section.

3. Evolving syntax Networks

Here we analyze the topological patterns displayed by syntax networks at different stages of language acquisition. In fig. (2) we show three examples of these networks.

† note that the operation is reversible, since can rebuild the tree from the dependency relations

At early stages, (fig. 2a,b) most words are isolated (not shown here) indicating a dominant lack of word-word linkage. Isolated words are not shown in these plots. For each stage, we study only the largest subset of connected words or *giant component* (GC). The reason for considering the largest connected component is that, from the very beginning, the GC is much larger than any other secondary connected component and in fact the system shows an almost all-or-none separation between isolated words and those belonging to the GC. In other words, the giant component captures almost all word-word relations. By sampling corpora at different times, we obtain a time series of connected networks $\mathcal{G}(\mathcal{W}_T, E_T)$, where $\mathcal{W}_T$ and E_T are the set of words and links derived from the T-th corpus, $T = 1, ..., 11$.

The most salient qualitative feature of language acquisition is the existence of two clearly differentiated regimes, already visible in (fig(2a-c). These distinct regimes have an impact in the organization of syntactic networks at different stages. For instance, we find that networks before the two-year transition show a tree-like organization. However, pre-transition networks are suddenly replaced by much larger, heterogeneous networks after the two-year transition -see fig. (2c)- which are very similar to adult syntactic networks (Ferrer-i-Cancho et al., 2004). This abrupt change indicates a global pattern of language re-organization marked by a shift in grammar structure. There is actually a large change in the nature of hubs before and after the transition. Highly connected words in the pre-transition stage are semantically degenerated lexical items, such as *it*. After the transition, hubs emerge as functional items, such as *a* or *the*. These hubs were essentially nonexistent in previous stages -see fig.(3).

A first quantitative measure is the connectivity of every element. The number of links (or *degree* $k_i = k(w_i)$ of a given word $w_i \in \mathcal{W}$ gives a measure of the number of syntactic relations between such word and its neighbors. Figure (3) shows the time series evolution of k for several relevant words. All of them display a sharp change around two-years ($T = 5$). The advantage of using degree as a measure of the relevance of a given word is that this topological trait is largely independent on the the frequency of appearance of the word or why it appears in a given corpus.

Two important measures allow to characterize the overall structure of these graphs. These are the average path length (L_T) and clustering coefficient (C_T). The first measure is defined as $D_T = \langle D_{min}(i,j) \rangle$, where $D_{min}(i,j)$ indicates the length of the shortest path connecting nodes w_i and w_j. The average is performed over all pairs of words. Roughly speaking, short path lengths means that it is easy to reach any given word w_i starting from another arbitrary word w_j. Small path lengths in sparse networks are often an indication of efficient information exchange. The clustering coefficient C_T is defined as the probability that two words that are neighbors of a given word are also neighbors of each other (i. e. that a triangle is formed). In order to estimate C_T, we define for each word w_i a neighborhood Γ_i. Each word $w_j \in \Gamma_i$ is syntactically related (at least once) with w_i in a production. The words in Γ_i can also be linked to each other, and the clustering $C(\Gamma_i)$ is defined as

$$C(\Gamma_i) = \frac{1}{k_i(k_i - 1)} \sum_j \sum_{k \in \Gamma_i} a_{jk} \tag{3.1}$$

The average clustering of the G_T network is simply $C_T = \langle C(\Gamma_i) \rangle$ i.e, the average

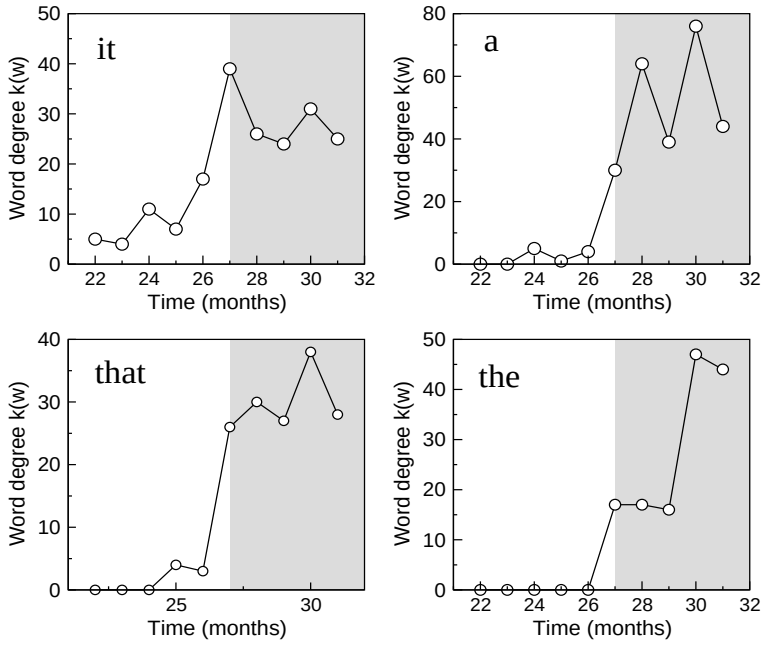

Figure 3. Time evolution of word degrees through language acquisition. Here four relevant words have been chosen: *it, a, that, the*. Their degree has been measured in each corpus and display a well-defined change close to the critical age of ≈ 24 months. Interestingly, *it* is rapidly replaced by *a* as the main hub as soon as purely functional words emerge.

over all $w_i \in W$. Most complex networks in nature and technology are known to be *small words*, meaning that they have short path lengths and high clustering (Watts and Strogatz, 1998) Although language networks have been shown to have small world structure (Ferrer-i-Cancho and Solé, 2001; Steyvers and Tenenbaum, 2005; Ferrer-i-Cancho et al., 2004; Sigman and Cecchi, 2002) little is known about how it emerges in developing systems.

Two regimes in language acquisition can be observed in the evolution of the average path length (figure 4a). It grows until reaches a peak at the transition. Interestingly, at $T = 5$ the network displays the highest number of words for the pre-transition stage. For $T > 5$, the average path length stabilizes $D_T \approx 3.5$ (see fig. (4 b)). The increasing trend of D_T in $T < 5$ may be an indication that combinatorial rules are not able to manage the increasing complexity of the lexicon. In figure 4b we plot the corresponding number of words N_T and links L_T of the GC as filled and open circles, respectively.

We can see that the number of connected words that belong to the GC increases in a monotonous fashion, displaying a weak jump at the age of two. However, the number of links (and thus the richness of syntactic relations) experiences a sharp change.

The rapid increase in link numbers indicates a qualitative change in network properties and pervades the reduction of average path length. A similar abrupt transition is observed for the clustering coefficient: In the pre-transition stage C_T is small (zero for $T = 1, 2, 3$). After the transition, it experiences a sudden jump.

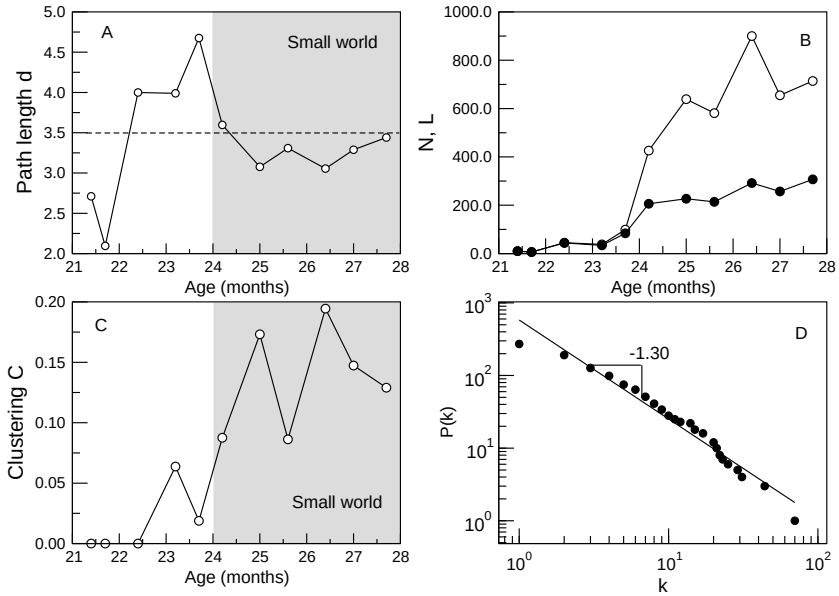

Figure 4. Changes in the structure of syntax networks in children are obtained by means of several quantitative measures. Here we display: (a) the average path length D_T, (b) The number of words (N_w) and links L (c) the clustering coefficient. As shown in (a) and (c), a small world pattern suddenly emerges after an age of ≈ 24 months. A rapid transition from a large L and low C takes place towards a small world network (with low D and high C). After the transition, well-defined scale-free graphs, with $P(k) \propto k^{-2.30}$, are observed (d).

Both D_T and C_T are very similar to the measured values obtained from syntactic graphs from written corpus (Ferrer-i-Cancho et al., 2004).

The small world behavior observed at te second phase comes from the presence of a heterogeneous distribution of links in the syntax graph. Specifically, we measure the degree distribution $P(k)$, defined as the probability that a node has k links. Global patterns of syntactic networks revealed a scale-free degree distributions $P(k) \propto k^{-\gamma}$, with $\gamma \approx 2.3 - 2.5$. An example is shown in (fig. (4 d)) where we plot the cumulative degree distribution, i.e:

$$P_{>}(k) = \int_k^{\infty} P(k)dk \sim k^{-\gamma+1} \tag{3.2}$$

which gives a $\gamma \approx 2.3$, also in agreement with adult studied corpora. Scale-free webs are characterized by the presence of a few elements (the hubs) having a very large number of connections. They are responsible for the very short path lengths and thus for the efficient information transfer in complex networks. The relationships between hubs are also interesting: the syntax graph is *dissassortative* (Newman, 2002), meaning that hubs tend to avoid to be connected among them (Ferrer-i-Cancho et al., 2004). In our networks, this tendency also experiences a sharp change close to the transition domain (not shown) thus indicating that strong constraints emerge strongly limiting the syntactic linking between functional words.

4. Discussion

Our study reveals two clearly differentiated behaviors in the early stages of language acquisition. Rules governing both grammatical and global behavior seem to be qualitatively and quantitatively different. Could we explain the transition in terms of self-organizing or purely external-driven mechanism? Clearly not, given the special features exhibited by our evolving webs, not shared by *any* current model of evolving networks (Dorogovtsev and Mendes, 2001, 2003).

Beyond the transition, some features diverge dramatically from the pre-transition graph. Particularly interesting is the changing role of the hubs: as soon as the purely functional words emerge they automatically replace semantically-degenerated words as the main hubs. Such features cannot be explained from external factors (such as communication constraints among individuals). Instead, it seems tied to changes in the internal grammar machinery. The sharp transition from small tree-like graphs to much larger scale-free nets, and the sudden change of the nature of hubs are the footprints of the emergence of new, powerful rules of exploration of the combinatorial space, i.e., the emergence of full adult syntax. What we see is an abrupt change on the underlying expressive power of the grammar, jumping into a recursive, unbounded one beyond the transition. This seems to support the hypotheses suggested by Hauser et al. (Hauser et al., 2002); see also (Nowak and Krakauer, 1999).

A further line of research should extend the analysis to other (typologically different) languages and clarify the nature of the innovation, and how general principles of communication and cognition predate such an innovation to generate grammar as we know. Preliminary work using three different european languages supports our previous results. Moreover, modeling the transitions from finite grammars to unbounded ones by means of connectionist approximations (Szathmáry et al., 2007) could shed light on the neuronal prerequisites that guide the acquisition process to a fully developed grammar as described and measured by our network approach.

The authors thank the members of the CSL for useful discussions. We also acknowledge Liliana Tolchinsky and Joana Rosselló for helpful comments on theory of syntax acquisition. Finally, we acknowledge Maria Farriols i Valldaura for her support during the whole process of this work. This work has been supported by grants FIS2004-0542, IST-FET ECAGENTS, project of the European Community founded under EU R&D contract 01194, the McDonnell foundation (RVS) and by the Santa Fe Institute.

References

Bickerton D. 1990. *Language and Species.* University of Chicago Press. Chicago.

Bloom L, Hood L, Lightbown P. 1974. Imitation in language development if when and why. *Cognitive Psychology.* (6):380–420.

Bloom L, Lightbown P, Hood L. 1975. Structure and variation in child language. *Monographs of the society for Research in Child Development. Serial 160.* (40).

Chomsky N. 1957. *Syntactic Structures.* The Hague: Mouton & Co. Paris.

Corominas-Murtra B. 2007. Network statistics on early english syntax: Structural criteria. *arXiv.org:0704.3708*.

Deacon TW. 1997. *The Symbolic Species: The Co-evolution of Language and the Brain*. New York: W.W. Norton.

Dorogovtsev SN, Mendes JFF. December 2001. Language as an evolving word web. *Proc. Royal Soc. London B*. 268.

Dorogovtsev SN, Mendes JFF. 2003. *Evolution of Networks*. Oxford University Press. New York.

Elman JL. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*. 48(1):71–99.

Ferrer-i-Cancho R, Solé RV. 2003. Least effort and the origins of scaling in human language. *Proc. Natl. Acad. Sci. USA*. 100:788–791.

Ferrer-i-Cancho R, Köhler R, Solé RV. 2004. Patterns in syntactic dependency networks. *Phys. Rev. E*. 69:051915.

Ferrer-i-Cancho R, Solé RV. November 2001. The small world of human language. *Proc. Royal Soc. London B*. 268.

Hauser MD, Chomsky N, Fitch TW. 11 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science*. 298:1569–1579.

Hudson R. 2006. *Language Networks: The New Word Grammar*. Oxford University Press. New York.

Humboldt WV. 1999. *On Language: the Diversity of Human Language Construction and its influence on the Mental Development of the Human Species, 2nd edn. Lomansky M., trans, Heath, P.L. (eds)*. Cambridge U. Press. Cambridge.

Hurford J. 1989. Biological evolution of the saussurean sign as a component of the language acquisition device. *Lingua*. 77(2):187–222.

Ke J. 2007. Complex networks and human language. *arXiv:cs/0701135*.

Komarova N, Niyogi P, Nowak M. 2001. The evolutionary dynamics of grammar acquisition. *J. Theor. Biol.* 209(1):43–59.

Komarova NL, Niyogi P. April 2004. Optimizing the mutual intelligibility of linguistic agents in a shared world. *Art. Int.* 154(1-2):1–42.

Macwhinney B. 2000. *The CHILDES project: Tools for analyzing talk. Third Edition.* Lawrence Erlbaum Associates. Mahwah, NJ.

Macwhinney B. 2005. The emergence of linguistic form in time. *Connection Science*. 17(3):191–211.

Maynard-Smith J, Szathmàry E. 1997. *The Major Transitions in Evolution*. University of New York Press. New York.

Melçuck I. 1989. *Dependency Grammar: Theory and Practice.* Oxford University Press. New York.

Newman MEJ. 2002. Assortative mixing in networks. *Phys. Rev. Lett.* 89(208701).

Newport EL. 1990. Maturational constraints on language learning. *Cogn. Sci.* 14 (1):11–28.

Niyogi P. 2006. *The Computational Nature of Language Learning and Evolution.* MIT Press. Cambridge, Mass.

Nowak MA, Krakauer D. 1999. The evolution of language. *Proc. Nat. Acad. Sci. USA*. 96(14):8028–8033.

Radford A. 1990. *Syntactic Theory and the Acquisition of English Syntax: the nature of early child grammars of English.* Oxford. Blackwell.

Sigman M, Cecchi G. 2002. Global organization of the wordnet lexicon. *Proc. Nat. Acad. Sci. USA*. 99(3):1742–1747.

Steyvers M, Tenenbaum JB. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cogn. Sci.* 29:41–78.

Szathmáry E, Szatmáry Z, Ittzés P, Orbán G, Zachár I, Huszár F, Fedor A, Varga M, Számadó S. 2007. In silico evolutionary developmental neurobiology and the origin of natural language. In: Lyon, C. and Nehaniv, C. L. and Cangelosi, A. *Emergence of Communication and Language.* Springer-Verlag. London pp. 151 – 187.

Watts DJ, Strogatz SH. 1998. Collective dynamics of 'small-world' networks. *Nature*. 393(6684):440–442.