In: Kokinov, B., Karmiloff-Smith, A., Nersessian, N. J. (eds.) ***European Perspectives on Cognitive Science***.
 ISBN 978-954-535-660-5

Modelling Language Acquisition in Children using Network Theory

Emil Gegov (emil.gegov@brunel.ac.uk)
School of Engineering and Design, Brunel University, Uxbridge, Middlesex UB8 3PH, UK

Fernand Gobet (fernand.gobet@brunel.ac.uk)
School of Social Sciences, Brunel University, Uxbridge, Middlesex UB8 3PH, UK

Mark Atherton (mark.atherton@brunel.ac.uk)
School of Engineering and Design, Brunel University, Uxbridge, Middlesex UB8 3PH, UK

Daniel Freudenthal (d.freudenthal@liv.ac.uk)
School of Psychology, University of Liverpool

Julian Pine (julian.pine@liv.ac.uk)
School of Psychology, University of Liverpool

Abstract

Research in children's language acquisition has recently benefited from the application of network theory to large sets of empirical data, which has illuminated interesting patterns and trends. Network theory is an extremely powerful modelling and analysis tool, and its full potential in terms of extracting useful information from raw data has yet to be exploited. In the present paper, we argue that well-established network analysis techniques can, and should be applied to the study of language acquisition, in order to reveal otherwise invisible patterns. We show that a key network parameter – the ranked frequency distribution of the links – provides useful information about the data, even though it had been previously neglected in this domain.

Keywords: language acquisition; network theory; word co-occurrence network.

Introduction

The ability of humans to communicate effectively through the use of a common language is fascinating. Even more impressive is the fact that young children, when learning their first language, are able to pick it up so well, even in the presence of noisy input. When a child produces an incorrect utterance, there is rarely any corrective feedback, so it is up to the child to filter those out over time.

Two opposing approaches try to explain this remarkable capability of children, and specifically, their syntax acquisition. Universal Grammar (UG) (Chomsky, 1957) argues that certain rules and parameters for language are hard-wired in each and every one of us, and the process of acquisition is simply a tuning of those variables. A key argument for UG is *the poverty of the stimulus argument*: the input received by children contains numerous errors, false starts, unfinished sentences, and thus is not sufficient for inferring the rules of language. UG has generally been accepted as the de facto theory for syntax acquisition, but an alternative approach developed in the 1990s is gaining increasingly more support in recent years. There is increasing evidence that the environment provides much more information than had been assumed by Chomsky, and a number of simulation models have shown that much grammatical knowledge can be learnt from child-directed speech (Freudenthal, Pine, Aguado-Orea, & Gobet, 2007; Redington, Chater, & Finch, 1998).

If one does not subscribe to UG's assumptions of innate abstract knowledge, one has to identify the specific mechanisms employed when young children acquire their first language. This can be achieved by using models, such as MOSAIC (Freudenthal et al., 2007), which are trained with maternal utterances, and then produce utterances that can be directly compared to children's utterances. By examining the quality of the obtained results, we can get a good idea of which mechanisms account for the empirical data. However, data sets of utterances are typically very large, noisy, and difficult to compare directly. Therefore, a worthwhile approach is to extract key characteristics of corpora first. This is what network modelling offers. It holds the key to embed all the information contained in a data set in a network, which can be analysed and compared with similar networks. It is important to identify which network parameters represent a useful statistic of the raw data, so they can be extracted from the network for analysis and cross-comparison.

Researchers modelling language using network theory have experimented with a wide range of parameters, but there seems to be no consensus on which parameters are essential for understanding language acquisition, and which are merely providing some additional network information. Table 1 summarises the parameters used by recent research involving the network modelling of language, to give an idea of the network properties that we are interested in. The publications in the Table were collected by identifying seven key papers (Numbers 1-6 and 11 in Table 1) on the modelling of language using network theory, and papers (Numbers 7-10 in Table 1) on the syntax of language that cite any one of the initial seven. The shaded entries correspond to the papers that investigate children's language

Table 1: Parameters of interest in linguistic networks

Ref.	Network Type	Language	Length	MLU	N	E	GCC	NCC	<k>	AC	DC	NN	D	L	C	P(k)	P(f)	P(b)
1	co-occurrence	English		x	x	x			x									
2	co-occurrence	English			x	x			x					x	x	x		
3	semantic	English			x	x			x					x	x	x		
4	co-occurrence syntactic semantic	English			x	x			x					x	x	x		
5	syntactic	English		x	x	x			x					x	x	x	x	
6	co-occurrence	English		x	x	x			x		x							
7	dependency syntactic	English			x	x			x				x	x	x	x		
8	char. structure	Chinese			x	x			x	x				x	x	x		
9	co-occurrence	Chinese			x	x			x	x		x		x	x	x	x	x
10	co-occurrence	Chinese	x		x	x	x	x	x				x	x	x	x		
11	co-occurrence	Chinese English	x		x	x	x	x	x				x	x	x	x		

References: 1. (Ke & Yao, 2008); 2. (Cancho & Solé, 2001); 3. (Motter, De Moura, Lai, & Dasgupta, 2002); 4. (Solé, Corominas-Murtra, Valverde, & Steels, 2010); 5. (Corominas-Murtra, Valverde, & Solé, 2010); 6. (Adamo & Boylan, 2008); 7. (Haitao & Fengguo, 2008); 8. (Li & Zhou, 2007); 9. (Zhou, Hu, Zhang, & Guan, 2008); 10. (Shi, Liang, Liu, & Tse, 2008); 11. (Liang et al., 2009).

acquisition, as opposed to language in general. The research summarised in Table 1 typically focuses on three types of networks: co-occurrence (see Figure 2 for a simple example), syntactic, and semantic; and two languages: English and Chinese. A total of sixteen unique statistical parameters were investigated in these publications, and we will briefly describe them here.

The **Length** of a linguistic data set may refer to the total number of characters, words, or utterances within the sample. The Mean Length of Utterance (**MLU**) is the average number of words in an utterance. To be precise, these two are in fact data set parameters, but they are nevertheless treated like the network parameters here. The total number of network nodes (**N**) and the total number of network links (**E**) are the two most basic network measures. The Giant Connected Component (**GCC**) represents the largest connected part of the entire network, i.e., the cluster with the highest number of nodes. In a connected network (where there are no isolated nodes or clusters of nodes), the GCC is identical to the original network. The Number of Connected Components (**NCC**) is simply the number of network components that are disconnected from one another. The average degree ***<k>*** is the average number of links adjacent to a network node. This key parameter reflects the overall connectivity of the network. The assortativity coefficient (**AC**) measures how assortative the network is, i.e., to what extent high-degree nodes are connected to other high-degree nodes. The degree centralization (**DC**) measures to what extent the links are centralized on a small number of high-degree nodes. The average nearest-neighbour degree (**NN**) of a node is the average degree of the nodes that are connected to the given node. The network diameter ***D*** is the length of the longest path between a pair of nodes, when the shortest possible paths are considered, i.e., containing the fewest links. The average geodesic length ***L*** is the average length of the shortest paths between all pairs of nodes. The clustering coefficient ***C***, averaged among all nodes, measures the likelihood of the neighbours of a node being connected themselves. The node degree distribution ***P(k)*** is the probability distribution of a randomly chosen node with degree *k*. Similarly, the node frequency distribution ***P(f)*** is the probability distribution of a node with frequency *f*. Finally, the node betweenness distribution ***P(b)*** is the probability distribution of a node with betweenness centrality *b*.

In the present paper, we investigate eight of these parameters: **MLU**, **N, E**, **GCC**, ***<k>***, ***L***, ***C*** and ***P(k)***, because they are key parameters in network theory, and they are also widely used in language modelling. In addition, we

highlight another well-known network metric, called the *ranked link frequency distribution*, which has been neglected by the language research community, to the best of our knowledge.

Ranked Link Frequency Distribution

The ranked link frequency distribution shows how the magnitude of link frequencies decreases when the frequencies are sorted in descending order. To compute this function, all the links of the given network are ranked in order of frequency. This frequency is then normalised for consistency to a value between 0 and 1 by dividing all frequencies by the highest frequency in the network. The distribution is defined as the normalised frequency as a function of the rank. This distribution is particularly interesting for studying language because previous research (Corominas-Murtra et al., 2010) has shown that ***P(f)*** – the probability distribution of a node with frequency f – in children's syntactic networks follows a power-law. Figure 1 shows an example of a ranked link frequency distribution from our maternal results, which also seems to follow a power law.

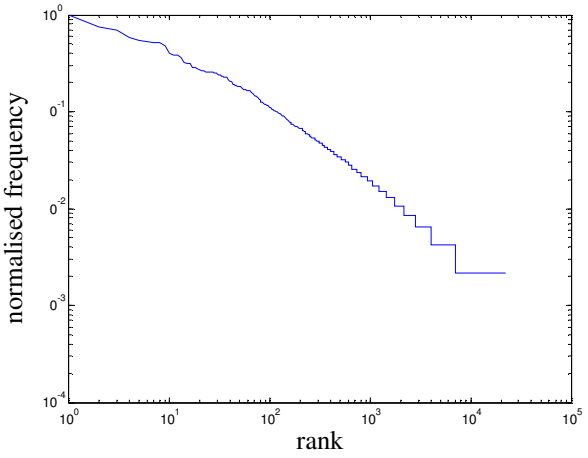

Figure 1: Ranked link frequency distribution

A power law is usually described by two parameters: a scaling factor **a** and an exponent **n** (thus, $f(x) = ax^n$). The scaling factor simply determines how much the function is shifted along the y-axis. The exponent controls the slope of the function; thus, a higher **n** (in absolute sense) results in a more skewed distribution. In a co-occurrence network, the presence of a power law means that language productivity is very biased towards some word co-occurrences, which are produced much more often than most other co-occurrences. Based on this, we expect a linear relationship between incremental stages of linguistic development and **n**.

This paper is structured as follows: in Method we describe the data collection and the construction of the networks; in Results we summarise our main findings in terms of parameter correlations and trends; in Discussion we evaluate our results and provide a conclusion.

Method

This section consists of four subsections: Data Sets, Filtering and Reduction, Construction of Networks, and Analysis of Networks. In Data Sets we describe how linguistic data sets are obtained from two sources: mothers and their children. In Filtering and Reduction we explain the various filtering techniques we used to ensure that the data are consistent for network modelling and analysis. In Construction of Networks we present the steps involved in the creation of word co-occurrence networks from the data. Finally, in Analysis of Networks we outline the analysis done on the networks.

Data Sets

We use data from the Manchester corpus of the CHILDES database (MacWhinney, 2000; Theakston, Lieven, Pine, & Rowland, 2001), which holds large files of logged conversations between mothers and their children, produced while they are interacting at home. Over a significant developmental time period, the children are regularly visited by an experimenter that records all the interactions for a fixed time period. The utterances are recorded on audio tapes that are transcribed to text files by keeping all clearly audible utterances and ignoring anything inaudible. The children's files are partitioned into three discrete, non-overlapping stages of development, and all the files for a given stage are combined to produce three data sets. We call these stage 1, stage 2 and stage 3. For each mother, the files at the three stages are combined to produce just one data set, as their language should remain fairly stable.

Filtering and Reduction

Since the raw data set consists of a long list of utterances, some of which have duplicates, those duplicates are removed in order to obtain consistent networks.

After they are filtered, all the data files' lengths – in terms of the number of utterances they contain – are recorded. To make the files as consistent as possible for later systematic analysis, all files are reduced to the length of the shortest file for a particular stage by randomly deleting utterances. The reduction is done using Matlab's random permutation function, which assigns each utterance a unique natural number between 1 and the total number of utterances. Then, to complete the reduction, all utterances that were assigned a number that is larger than the size of the required data set are discarded.

Construction of Networks

Word co-occurrence networks are built using the data in a process consisting of four stages. Figure 2 illustrates a simple version of such a network for the following two sentences: *The cat sleeps*. *The dog wakes the cat*.

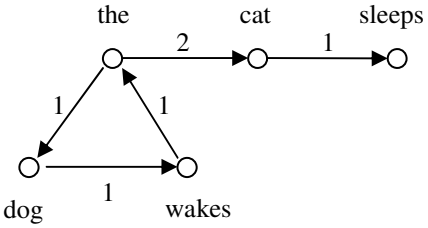

Figure 2: Simple word co-occurrence network

In the first stage, the utterances are split into overlapping pairs of adjacent words, such that each co-occurrence is represented by a single pair. For example, the first two pairs in this sentence are "For example" and "example the", when ignoring any punctuation marks. Also, single word utterances are kept unchanged because they become isolated nodes in the final network, if they do not appear in any pair elsewhere.

In stage two, the pairs are transformed into a corresponding network of nodes and links using a simple mapping: for every pair, insert a directed link from the first word to the second word. The network is represented as a list of nodes and links, where the nodes are trivially obtained by taking all words from the pairs, and removing duplicates.

However, this network is a multidigraph, i.e., there exist multiple links from a source to a target node, which is to be avoided if possible since it is difficult to process. Therefore, in stage three, the network is converted to a weighted digraph – with no parallel links – where the weight (frequency) of a link denotes the original number of links from the source to the target node in the multidigraph. This digraph is used in all of the analysis except for the calculation of two network parameters – the average geodesic length and the clustering coefficient – which require a simple graph, i.e., an unweighted, undirected graph with no self-loops.

Hence, for the final stage four, the weighted digraph is converted to a simple graph by erasing all link weights, converting all links from directed to undirected (and removing duplicates), and removing all self-loops.

Analysis of Networks

First, we compute the network parameters and then we follow an analysis process – from networks to conclusions – which has two branches.

The first branch is concerned with the average for a given source and stage, thereby ignoring the specifics of the individual children. For each parameter, the average value across the six children is calculated and a summary plot is produced.

The second branch of the analysis deals with correlations between mothers and children, for a given source and stage, thereby focusing on the details of the individual children. The process begins with a decision. If a single parameter is being analysed, the value of this parameter is used and the fitting step is by-passed. By contrast, if a function (i.e. the frequency distribution or the degree distribution) is being analysed, it is necessary to fit a best-fit curve to the data so that the parameters of the function can be used for further analysis. For both functions we use a power-law fit of the form $f(x) = ax^n$ where **a** is the coefficient and **n** is the exponent. In the frequency distributions, **x** represents the rank and **f(x)** represents the normalised frequency for that rank. In the degree distributions, **x** represents the degree and **f(x)** represents the probability of a randomly chosen node with degree **x**.

Finally, the correlations between the parameters of the six children and those of their mothers are calculated.

Results

The results are presented in two parts. First we summarise the results obtained for the ranked link frequency distributions using summary plots of averages, and we briefly describe the other parameters that were obtained. Then, all the parameter correlations between the two sources are reported. At this point we would like to note that the quality of all best-fit functions in this research is checked by computing the correlation coefficient between the real data and the fitted function, in order to ensure the reliability of the results. For the degree distribution fits, all correlations are above 0.99, indicating an almost perfect power-law relationship in the data. For the frequency distribution fits, all correlations are above 0.92, which is still an excellent fit between data and fitted function.

After we calculated the node in-degree and out-degree distributions, we found that the first data point (representing in/out-degree 0) is far from a power-law fit due to its unique nature, and therefore regarded it as an additional parameter, which we called **p.** The reason for this behaviour is that the in/out-degree 0 nodes decrease as the network becomes increasingly connected.

Summary Plots

The summary plots for the parameters of the ranked link frequency distributions demonstrate their effectiveness in capturing linguistic development, and also show a much unexpected trend in one of the two parameters. For the mothers and for each stage of the children's development, we calculated the mean and the Standard Error of the Mean (SEM) of the six subjects, and plotted the results on two graphs (one for each parameter). The points on the graphs represent the mean and the vertical bars represent the SEM. The mothers' results are shown next to the children's stage 3 results for comparison. In Figure 3 we have shown parameter **a**, and in Figure 4 – parameter **n**. Note that whereas the children are steadily approaching their mothers in Figure 3, they diverge in Figure 4, but the mothers are exactly where we would have expected their children to be by linear extrapolation.

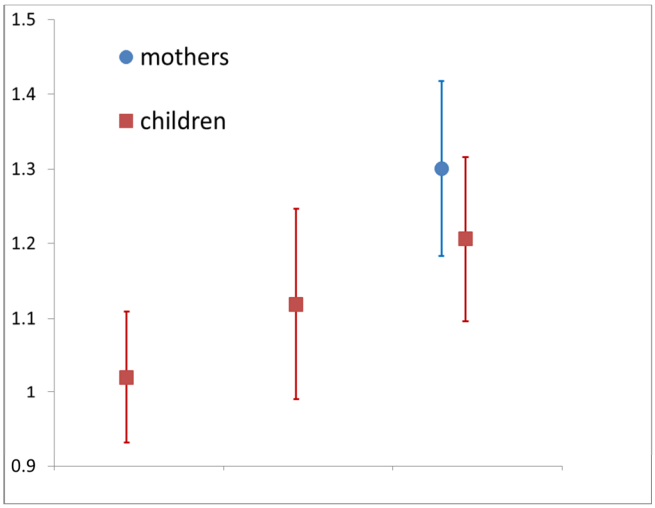

Figure 3: Summary plot for parameter **a** of the link frequency distributions

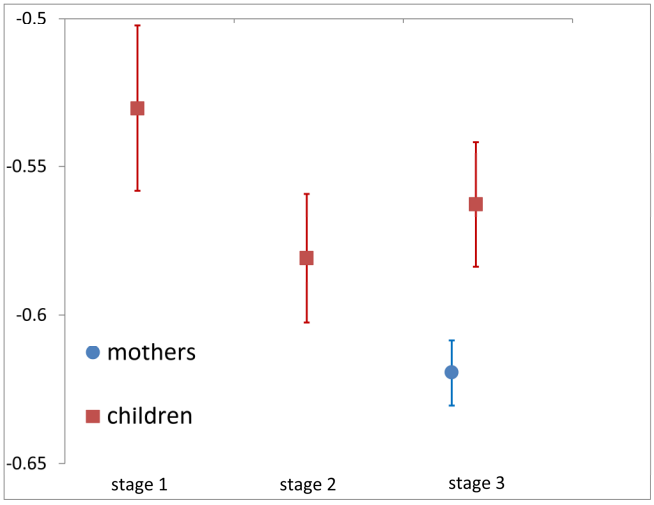

Figure 4: Summary plot for parameter **n** of the link frequency distributions

Inspection of each of the single network parameters on a summary plot (not shown), shows that **MLU**, **N**, **E**, **GCC**, **<*k*>** and ***L*** are behaving predictably with respect to the developmental stage of the children – the former five are increasing and ***L*** is decreasing because the networks are becoming more connected. On the other hand, ***C*** drops slightly in stage 3, suggesting that the children probably experimented with a wider variety of utterances, resulting in fewer word co-occurrence loops. This curvilinear function is interesting and is in line with what we have found for the **n** parameter of the ranked link frequency distribution.

By examining the in-degree and out-degree distribution parameters we observe a decrease in **p** and a small increase in **a**, but non-linear behaviour in **n**. However, the magnitude of these changes in parameter **n** is small. Also, the children appear to be close to the mothers with respect to all three degree distribution parameters. This leads us to believe that children in general produce utterances that are statistically similar to adults, in terms of the bias to particular word usage.[1]

Correlations

The correlation coefficients between the parameters of the mothers and their corresponding children at stage 3 were calculated, in order for us to identify which parameters, if any, in the children resemble their mothers'. Table 2 reports the single network parameter correlations. Table 3 presents the correlations of parameters of the best-fit functions of the node in-degree and out-degree distributions. The correlations of parameters **a** and **n** of the best-fit functions of the ranked link frequency distributions are -.56 and .06, respectively. Overall, some of these correlations are moderate ($0.30 \leq r < 0.50$) or even strong ($r \geq 0.50$), to use Cohen's (1988) criteria. However, given the small number of observations (6 child/mother pairs) and hence degree of freedom ($n = 4$), none of these correlations are statistically significant.

Table 2: Single network parameter correlations between mothers and children

MLU	**nodes**	**links**	**GCC**	**<*k*>**	***L***	***C***
.62	.39	-.02	.39	-.14	-.19	-.39

Table 3: Correlations of parameters of node degree distributions between mothers and children

In-degree			**Out-degree**		
p	a	n	p	a	n
.43	.41	.32	.54	.54	.41

Discussion

We begin this section by focusing on the ranked link frequency distributions in Figures 3 and 4. Clearly, coefficient **a** of the children is increasing in time in a linear fashion, converging to the mothers'. Here, **a** is the best-fit estimate of the top-ranked normalised frequency. Therefore, it should be close to 1, and the fact that it is increasing above 1 over the three stages implies that the best-fit is diverging from the data for the top-ranked frequency, in order to provide a better fit to the rest of the data. This

[1] The standard deviation of both mothers and children is significantly different from 0, for all 15 parameters under study. For 9 of the parameters, the SEM, and therefore, the variability of the children, is greater than that of the mothers. Specifically, for the more complex parameters (***L***, ***C***, **n** (in-degree), **n** (out-degree), and **n** (frequency)), the children have a significantly higher variability compared to the mothers, and vice versa, for the more simple parameters (**MLU**, **N**, **E**, **GCC**), the mothers have higher variability. For the moderately complex parameters (**<*k*>**, **p** (in-degree), **p** (out-degree), **a** (in-degree), **a** (out-degree), and **a** (frequency)), the children and the mothers have relatively similar variability.

suggests that over the stages, the top frequency is falling below the expected power-law frequency. In other words, the most common pair of words occurs a little less frequently than a power-law would predict. However, the exponent **n** is in fact the key parameter in a power-law, as it controls how skewed the function is. Therefore, an increasing **n** in absolute terms (going down in Figure 4 due to negative sign), implies more bias towards certain pairs of words in language productivity. We note that the mothers are well-aligned with the children of stages 1 and 2, but the stage 3 children have in fact diverged completely from the expected linear trend. This is a clear indicator that in terms of word combinations, children are still developing their linguistic skills at this stage 3. We suspect that the children probably experimented with a variety of new, or relatively new, word combinations when producing utterances, thereby reducing the frequency of the more common word pairs, which are already well-known. Even though the other network parameters (except ***C***) and the previous two stages for this parameter suggest otherwise, we have seen that the exponent of the ranked link frequency distribution has uncovered something very surprising. This result is also supported by the trend in ***C***, but more importantly, the frequency distribution is a function parameter of the link weights, so it is telling us a lot more about the dynamics on the network.

Now we briefly concentrate on the correlation coefficients that we obtained between children and mothers. These correlations are difficult to interpret without strong *a priori* theoretical hypotheses, which we do not have at the moment. In addition, the small number of degrees of freedom means that it is very difficult to reach the significance level, which none of our correlations did. Thus, research with larger samples must be awaited before stronger conclusions can be made about the meaning of these correlations.

In summary, we have reached two main conclusions. Firstly, we have demonstrated that the ranked link frequency distribution of a word co-occurrence network is a powerful tool that can provide deeper insights into the language acquisition of young children. Therefore, we encourage its use in future research in this field and hope that the language acquisition community will exploit its potential. Secondly, we have found an interesting pattern of correlations between children and mothers for the parameters under study, but have suggested that clear-cut theoretical hypotheses are necessary to make sense of them. In general, then, network analysis provides powerful constrains for theories of language acquisition. Future research in this area will focus on the testing of syntax acquisition models, the search for more useful network parameters, and the robustness of linguistic networks to perturbation.

Acknowledgements

This research is funded by the Research Support and Development Office at Brunel University.

References

Adamo, M., & Boylan, S. (2008). *A network approach to lexical growth and syntactic evolution in child language acquistion.* Unpublished manuscript.

Cancho, R. F. I., & Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society B: Biological Sciences, 268*(1482), 2261-2265.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton and Co.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Corominas-Murtra, B., Valverde, S., & Solé, R. V. (2010). Emergence of scale-free syntax networks. *Evolution of Communication and Language in Embodied Agents*, 83.

Freudenthal, D., Pine, J. M., Aguado-Orea, J., & Gobet, F. (2007). Modeling the developmental patterning of finiteness marking in English, Dutch, German, and Spanish using MOSAIC. *Cognitive Science, 31*(2), 311-341.

Haitao, L., & Fengguo, H. (2008). What role does syntax play in a language network? *Europhysics Letters, 83*(1) 18002.

Ke, J., & Yao, Y. (2008). Analysing language development from a network approach. *Journal of Quantitative Linguistics, 15*(1), 70.

Li, J., & Zhou, J. (2007). Chinese character structure analysis based on complex networks. *Physica A: Statistical Mechanics and its Applications, 380*(1-2), 629-638.

Liang, W., Shi, Y., Tse, C. K., Liu, J., Wang, Y., & Cui, X. (2009). Comparison of co-occurrence networks of the Chinese and English languages. *Physica A: Statistical Mechanics and its Applications, 388*(23), 4901-4909.

MacWhinney, B. (2000). *The CHILDES project: Tools for analysing talk* (3rd ed.). Mahwah, NJ: Erlbaum.

Motter, A. E., De Moura, A. P. S., Lai, Y., & Dasgupta, P. (2002). Topology of the conceptual network of language. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics, 65*(6), 065102/1-065102/4.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*(4), 425-469.

Shi, Y., Liang, W., Liu, J., & Tse, C. K. (2008). Structural equivalence between co-occurrences of characters and words in the Chinese language. In: 2008 International Symposium on Nonlinear Theory and its Applications.

Solé, R. V., Corominas-Murtra, B., Valverde, S., & Steels, L. (2010). Language networks: Their structure, function, and evolution. *Complexity, 15*(6), 20-26.

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*(1), 127-152.

Zhou, S., Hu, G., Zhang, Z., & Guan, J. (2008). An empirical study of Chinese language networks. *Physica A: Statistical Mechanics and its Applications, 387*(12), 3039-3047.