Getting real about systematicity

Stefan L. Frank
s.frank@ucl.ac.uk

Department of Cognitive, Perceptual and Brain Sciences
University College London

This is a pre-final version of a book chapter to appear in:

P. Calvo & J. Symons (Eds.), *Systematicity and cognitive architecture: conceptual and empirical issues 25 years after Fodor & Pylyshyn's challenge to connectionism.* Cambridge, MA: MIT Press.

1. Introduction

1.1. Systematicity and reality

In the 25 years since its inception, the systematicity debate has suffered from remarkably weak empirical grounding. For a large part, the debate has relied on purely theoretical arguments, mostly from the Classists side (e.g., Aizawa, 1997a; Fodor & Pylyshyn, 1988; Phillips, 2000) but occasionally from the Connectionist camp as well (Bechtel, 1993; Van Gelder, 1990). And although there have been many attempts to empirically demonstrate (lack of) systematicity in connectionist models, it remains doubtful how these demonstrations bear upon reality considering that they are always restricted to hand-crafted, miniature domains. This is the case irrespective of whether they are presented by supporters of connectionist systematicity[1] (Bodén, 2004; Brakel & Frank, 2009; Chang, 2002; Christiansen & Chater, 1994; Elman, 1991; Farkaš & Crocker, 2008; Fitz & Chang, 2009; Frank, 2006a, 2006b; Frank & Čerňanský, 2008; Frank, Haselager, & van Rooij, 2009; Hadley, Rotaru-Varga, Arnold, & Cardei, 2001; Jansen & Watter, 2012; McClelland, St.John, & Taraban, 1989; Miikkulainen, 1996; Monner & Reggia, 2011; Niklasson & Van Gelder, 1994; Voegtlin & Dominey, 2005; Wong & Wang, 2007) or by those who are more skeptical (Marcus, 2001; Phillips, 1998; Van der Velde, Van der Voort van der Kleij, & De Kamps, 2004).

My goal in this chapter is to approach the systematicity problem in a fully empirical manner, by directly comparing a connectionist and a symbolic sentence-processing[2] model in a (more or less) realistic setting. As far as this chapter is concerned, getting real about systematicity means three things. Firstly, Connectionists can no longer get away with presenting models that only function within some unrealistic toy domain. To the extent that the systematicity issue is relevant to real-life cognitive systems, Connectionists should be able to demonstrate that (alleged) instances of systematicity do not crucially depend on the artificial nature of the simulation.

Secondly, I also aim to raise the bar for Classists, who need to empirically back up their claim that symbol systems are necessarily systematic. Aizawa (1997b) argues that compositionality is not a sufficient condition for systematicity and, indeed, to the best of my knowledge it has never been empirically demonstrated that symbol systems are any more systematic than neural networks. Nevertheless, even many Connectionists accept the premise that symbol systems explain systematicity.

---

[1] As a possible exception, the model by Borensztajn, Zuidema, & Bod (2009a), does learn from real-world data (2,000 utterances produced by a single child). However, it can be argued that the model forms a neural implementation of a symbolic parser, in which case it does not constitute a demonstration of eliminative connectionism (in the sense of Pinker & Prince, 1988).

[2] This chapter is only about language processing, arguably the main battleground for the systematicity debate, but I believe most of the same conclusions to hold for human cognition in general.

Thirdly, rather than defining particular levels of systematic behavior based on the specifics of training input and novel examples (as in, e.g., Hadley, 1994a, 1994b), the question of how systematic cognition really is will be avoided altogether. People learn language from what is "out there" and, subsequently, comprehend and produce more language "out there". Hence, the generalization abilities of the models presented here are investigated by training and testing both models on a large sample of sentences from natural sources. There is no invented, miniature language and no assumptions are made about which specific syntactic construction in the training data should result in which specific systematic generalizations.

1.2. Statistical modeling of language

While the systematicity debate in Philosophy and Cognitive Science revolved around theoretical arguments and unrealistic examples, actual progress was being made in the field of Computational Linguistics. The development of statistical methods for learning and processing natural language resulted in many successful algorithms for tasks such as sentence parsing, translation, and information retrieval. Recently, there has been a growing interest in applying such models to explain psychological phenomena in human language comprehension (e.g., Boston, Hale, Patil, Kliegl, & Vasishth, 2008; Brouwer, Fitz, & Hoeks, 2010; Levy, 2008), production (e.g., Levy & Jaeger, 2007) and acquisition (e.g., Bod & Smets, 2012; Borensztajn, Zuidema, & Bod, 2009b). The systematicity controversy tends not to arise here; for one because Computational Linguists are often more concerned with practical than theoretical issues. Also, and perhaps more importantly, these models are typically symbolic and thereby dodge the systematicity critique.

What recent models from Computational Linguistics share with connectionist ones is their statistical nature: They are concerned with the problem of extracting useful statistics from training data in order to yield optimal performance on novel input. As Hadley (1994a) pointed out, this issue of correct generalization to previously unseen examples is exactly what systematicity is all about. Hence, one might expect statistical Computational Linguistics to be vulnerable to the same systematicity critique as Connectionism. On the other hand, these statistical models for natural language processing are symbol systems in the sense of Fodor & Pylyshyn (1988) so, according to some, their systematic abilities are beyond any doubt.

1.3. Overview

In what follows, the systematic abilities of two sentence-processing models will be compared directly. The two models are of fundamentally different types: One is a thoroughly connectionist recurrent neural network (RNN); the other a truly symbolic probabilistic phrase-structure grammar (PSG). Both are trained on a large number of

naturally occurring English sentences, after which their performance is evaluated on a (much smaller) set of novel sentences. In addition, it is investigated how they handle ungrammatical word strings. The ideally systematic model would have no problem at all with the correct sentences but immediately "collapse" when ungrammaticality is encountered. As it turns out, neither model is perfectly systematic in this sense. In fact, the two models behave quite similarly, although the symbolic model displays slightly stronger systematicity. As will be discussed, these results suggest that, when dealing with real-world data, generalization or systematicity may not be a relevant property when arguing for or against a model's cognitive adequacy.

2. Simulations

2.1. Model training data

As discussed in the Introduction, connectionist models are typically trained on an artificial miniature language whereas models from Computational Linguistics are broad-coverage, being able to deal with sentences from natural sources. Although the latter approach was used here for both types of models, the task was made more manageable for the neural network by reducing the size of the language: The vocabulary was restricted to 7,754 word types (including the comma and the sentence-final period, which are treated as regular words) that occur with high frequency in the written-text part of the British National Corpus (BNC). The training data consisted of all 702,412 sentences from the BNC (comprising 7.6 million word tokens) that contain only words from the vocabulary. This is the same data set as used by Fernandez Monsalve, Frank, & Vigliocco (2012; Frank, 2013; Frank & Thompson, 2012).

The connectionist model was trained on just these sentences. In contrast, the symbolic model, being a probabilistic grammar, needs to be induced from a so-called treebank: a collection of sentences with syntactic tree structures assigned. To obtain these, the selected BNC sentences were parsed by the Stanford parser (Klein & Manning, 2003). The resulting treebank served as the training data for the grammar.

2.2. Models

2.2.1. Recurrent neural network

RNNs have formed the standard connectionist model of sentence processing ever since the seminal paper by Elman (1990). However, such models are difficult to scale up and were therefore always limited to unrealistic, miniature languages. In order to train the current model on the 7.6-million-word data set, the training process was separated into three distinct stages, as illustrated in Figure 1.

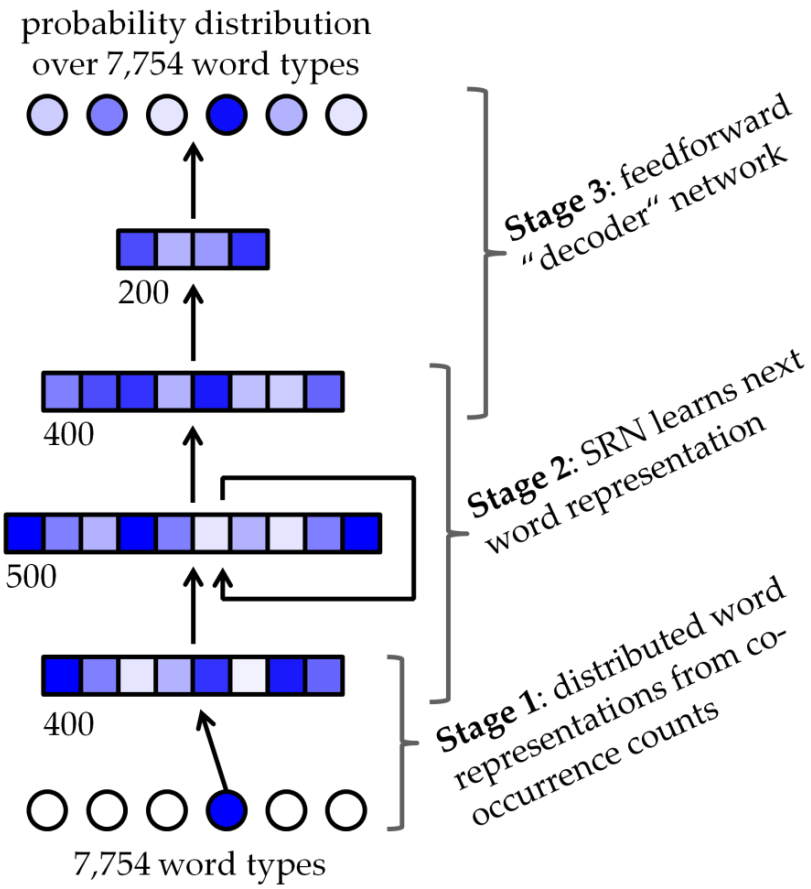

Figure 1. Architecture and training stages of the RNN model (reproduced from Fernandez Monsalve et al., 2012).

Firstly, each word type was represented by a high-dimensional vector, based on the frequencies with which the words occur adjacently in the training data. More specifically, word co-occurrence frequencies were collected in a matrix with 7,754 rows (one per word type) and 2×7,754 columns (corresponding to the directly preceding and following word types). These frequencies were transformed into pointwise mutual information values, after which the 400 columns with the highest variance were selected, yielding a 400-element vector per word type. These representations encode some of the paradigmatic, distributional relations between the words. For example, words from the same syntactical category tend to be represented by more similar vectors than words from different categories (cf. Frank, 2013).

Secondly, the selected BNC sentences (in the form of sequences of word vector representations) served as training input and target output for the recurrent part of the network. As is common in simulations using Simple Recurrent Networks (e.g., Elman, 1990, 1991; Frank, 2006a, among many others), it was trained to predict, at each point in each sentence, what the next input will be. More specifically, for each sentence-so-far $w_1,\ldots,w_t$, the input sequence consisted of the words' vector representations and the target

output was the vector representing the sentence's next word w_{t+1}. The training data set was presented to this part of the network five times and standard backpropagation was used to update the connection weights.

After this second training stage, the outputs of the recurrent part of the network form a mix of distributed, 400-element word vectors, somehow representing the possible continuations w_{t+1} (and their probabilities) of the input sequence so far. In order to interpret this mix of word vectors, a two-layer feedforward "decoder" network was applied in the third stage of model training. After each sequence $w_1,\ldots,w_t$, the decoder network took the outputs of the trained recurrent part of the network as its input, and learned to activate only one of 7,754 units, corresponding to the upcoming word w_{t+1}. In practice, of course, the network will never actually activate just a single output unit because many different continuations are possible at each point in a sentence.

The training data were presented two times to this decoder network and standard backpropagation was used for training. An important difference with Stage-2 training is that the decoder network's output units have softmax activation functions, ensuring that output activations are always non-negative and sum to one. That is, they form a probability distribution. To be precise, if $w_1,\ldots,w_t$ represents the first t words of a sentence, then the network's output activation pattern forms the model's estimated probability distribution $P(w_{t+1} \mid w_1,\ldots,w_t)$ over all word types. So, for each word type, the model estimates a probability that it will be the next word given the sentence so far. These estimates are accurate to the extent that the model was trained successfully.

As should have become clear from the exposition above, the RNN model uses only non-symbolic, distributed, numerical representations and processes. Crucially, the model is non-compositional in the sense that the representation of a word sequence does not contain representations of the individual words. There are no symbolic components at all. Thus, according to Fodor & Pylyshyn (1988), the model should be unable to display any systematicity.

2.2.2. Phrase-Structure Grammar

The PSG, needless to say, operates very differently from the RNN. To begin with, it is based on linguistic assumptions about hierarchical constituent structure: a sentence consists of phrases, which consists of smaller phrases, etcetera, until we get down to individual words.

The grammar operationalizes this idea by means of context-free production rules. These rules are induced from a treebank, which in this case is formed by the sentences selected from the BNC together with their syntactic tree structures. Figure 2 shows one of the 702,412 training items. From this single example, we can observe that a noun phrase (NP) can consist of either a singular noun (NN) or of a determiner (DT) followed by a

singular noun. Hence, the two production rules "NP → NN" and "NP → DT NN" appear in the grammar.

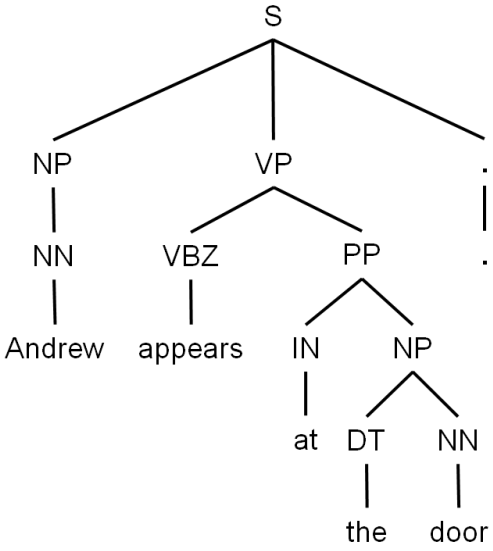

Figure 2. The syntactic tree structure assigned by the Stanford parser to the BNC sentence *Andrew appears at the door.* Syntactical category labels: NN (singular noun), VBZ (third person singular present-tense verb), IN (proposition), DT (determiner). Phrasal labels: S (sentence), NP (noun phrase), VP (verb phrase), PP (propositional phrase).

The leafs of the tree are formed by words. Each word token belongs to one of a number of syntactical categories (also known as parts-of-speech) which forms its direct "parent" node in the tree. So, for example, the tree of Figure 2 provides evidence that the word *at* can be a preposition (i.e., there is a production rule "IN → *at*").

In a probabilistic grammar, like the one used here, each rule is assigned a probability conditioned upon the rule's left-hand side. Based on just the single example from Figure 2, each of the two NP-rules receives a probability of 0.5, indicating that, when faced with an NP, it becomes either NN or DT NN, with equal probability. Needless to say, the 702,412 tree structures in the training data set give rise to much more fine-grained and accurate probabilities.

In practice, more advanced statistical techniques are applied to induce the PSG used here. Firstly, the probability of a rule is conditional not only on its left-hand side but also on larger parts of the tree (see Frank & Bod, 2011, for details). This is to say that the probability of producing "DT NN" depends not only on the identity of their parent node (NP) but also on their (great)grandparents (PP and VP in the example above). This makes the grammar more sensitive to structure and greatly improves its performance (Fernandez Monsalve et al., 2012). Secondly, probabilities are not purely based on frequencies in the treebank: In order to improve the probability estimates of very low-frequency events, additional smoothing of probabilities is applied.

The grammar was induced by Roark's (2001) algorithm, and implementation thereof, using his default settings. Next, Roark's incremental parser was used to estimate conditional word probabilities $P(w_{t+1} | w_1,\ldots,w_t)$, just like the RNN did. These follow from the probabilities of sentence-initial word sequences because:

$$P(w_{t+1}|w_1, \ldots, w_t) = \frac{P(w_1, \ldots, w_t)}{P(w_1, \ldots, w_{t+1})}$$

Here, $P(w_1,\ldots,w_t)$ is the probability of a sentence-so-far, which equals the sum of probabilities of all complete sentences that begin with $w_1,\ldots,w_t$. The probability of a complete sentence equals the sum of probabilities of all its possible syntactic tree structures. The probability of a tree structure, in turn, is the product of probabilities of the rules used in its construction. In this manner, a PSG can be used to estimate the probability of a word given the sentence so far. As will be discussed in the next section, these conditional probability estimates are central to model evaluation.

In contrast to the RNN, the PSG is very much a symbol system. It applies symbolic operations over discrete units (words, syntactical category labels, and phrasal labels) and has compositional representations: A sentence's tree structure consists of representations of subtrees, of syntactical and phrasal categories, and of the sentence's words. Because of the model's probabilistic nature, numerical processing is also involved to compute the probabilities, but this does not diminish the system's symbolic character. Hence, Fodor & Pylyshyn (1988) would claim that the PSG necessarily displays systematicity.

2.3. Model evaluation

2.3.1. Evaluation sentences

In order to evaluate the models' ability to generalize to realistic input, 361 sentences were selected from three novels (for details, see Frank, Fernandez Monsalve, Thompson, & Vigliocco, in press). These novel sentences contained a total of 5,405 word tokens (including commas and periods), all of which are present in the vocabulary of 7,754 high-frequency words.

A properly systematic model should not only be able to deal with new sentences, but also reject ungrammatical input. To test how the models handle ungrammatical word strings, scrambled versions of the original 361 sentences were created. This was done by choosing, from each sentence, two words at random and swapping their positions. This is done n times, for $n = 0$ (i.e., the grammatical sentence) up to $n = 9$, creating ten different sets of test data. Table 1 shows an example of one sentence and its nine scrambled version. Note that every sentence ends with a period, which remains in its sentence-final location.

Table 1: Example of an evaluation sentence and its nine, increasingly scrambled versions.

Scrambling level	**Sentence**
0	andrew closed the office door on the way out .
1	andrew closed the office door on the out way .
2	andrew on the office door closed the out way .
3	andrew on the office door the closed out way .
4	closed on the office door the andrew out way .
5	closed on the out door the andrew office way .
6	closed the the out door on andrew office way .
7	closed the andrew out door on the office way .
8	the the andrew out door on closed office way .
9	the the andrew way door on closed office out .

The probability of the word string being grammatical decreases with larger n, and most often even $n = 1$ already yields an ungrammatical string.[3] A very rigid language model would estimate zero probability for an ungrammatical string, but such an estimate would necessarily be incorrect. After all, impossible things by definition do not happen, so any input that actually occurs must have had a nonzero probability.

2.3.2. Evaluation measure

The question remains how to quantify the models' ability to generalize. Earlier evaluation measures, such as those proposed by Frank (2006a) and Christiansen & Chater (1999), require knowledge of all grammatical next-word predictions, so these measures can only be used if the true grammar is known, as is the case when using artificial languages. For natural language, the true grammar is (arguably) unknowable[4] or possibly even non-existent. Therefore, the evaluation measure applied here uses only the model-estimated next-word probabilities, $P(w_{t+1} \mid w_1,\ldots,w_t)$, for the *actual* next word w_{t+1}. The larger these probabilities, the more accurate were the model's expectations and, therefore, the better the model captured the statistics of the language.

Rather than using the conditional probability itself, it is transformed by the (natural) logarithm, that is, we take $\log(P(w_{t+1} \mid w_1,\ldots,w_t))$, which ranges from negative infinity (when $P(w_{t+1} \mid w_1,\ldots,w_t) = 0$) to zero (when $P(w_{t+1} \mid w_1,\ldots,w_t) = 1$). The negative of this value is an information-theoretic measure known as *surprisal*, which expresses the amount

[3] Here, grammaticality is a subjective notion. There is no (known) "true" grammar to objectively decide whether a word string is grammatical.

[4] Although there is a sense in which speakers of a language know its grammar, this knowledge is mostly procedural in nature (cf. Bybee, 2003). That is, people know how to *use* their language but have little or no conscious access to its grammar.

of information conveyed by word w_{t+1}. A word's surprisal is also of cognitive interest because it is believed to be indicative of the amount of 'mental effort' required to understand the word in sentence context (Hale, 2001; Levy, 2008). Indeed, surprisal values have been shown to correlate positively with word-reading times (e.g., Boston et al., 2008; Demberg & Keller, 2008; Fernandez Monsalve et al., 2012; Frank & Bod, 2011; Frank & Thompson, 2012; Smith & Levy, 2013, among many others).

Note that surprisal (i.e., mental effort and amount of information) is infinitely large if a word appears that was estimated to have zero probability. A more appropriate way to put this is that a model is infinitely wrong if something happens that it considers impossible. In practice, however, both models always estimate strictly positive probability for each word type at any point in the sentence. In the PSG, this is due to the smoothing of probabilities. In the RNN, this is because the connection weights have finite values.

3. Results

The leftmost panel of Figure 3 shows how the next-word probabilities that are estimated by the models decrease as the RNN and PSG are made to process increasingly scrambled sentences.[5] For correct sentences (i.e., $n = 0$), the PSG performs slightly better than the RNN. As expected, both models make worse predictions as the input contains an increasing number of grammatical errors. This effect is stronger for the PSG than for the RNN. Although the difference between the two models is small for lower levels of scrambling, all differences were statistically significant (all $t_{5341} > 3.19$; $p < .002$, in paired t-tests) due to the large number of data points.

Figure 3 also presents the coefficient of correlation between the RNN's and PSG's estimates of $\log(P(w_{t+1} | w_1,\ldots,w_t))$, as a function of scrambling level. This correlation is clearly very strong, and although it weakens as sentences are scrambled more, it seems to stabilize at around $r = .9$.

[5] For a fair comparison between models, clitics were excluded because the two models treat these differently: A word like *isn't* is considered a single word by the RNN whereas the PSG parser splits it into *is* and *n't*. After removing such cases, there are 5,344 probability estimates for each model and level of scrambling.

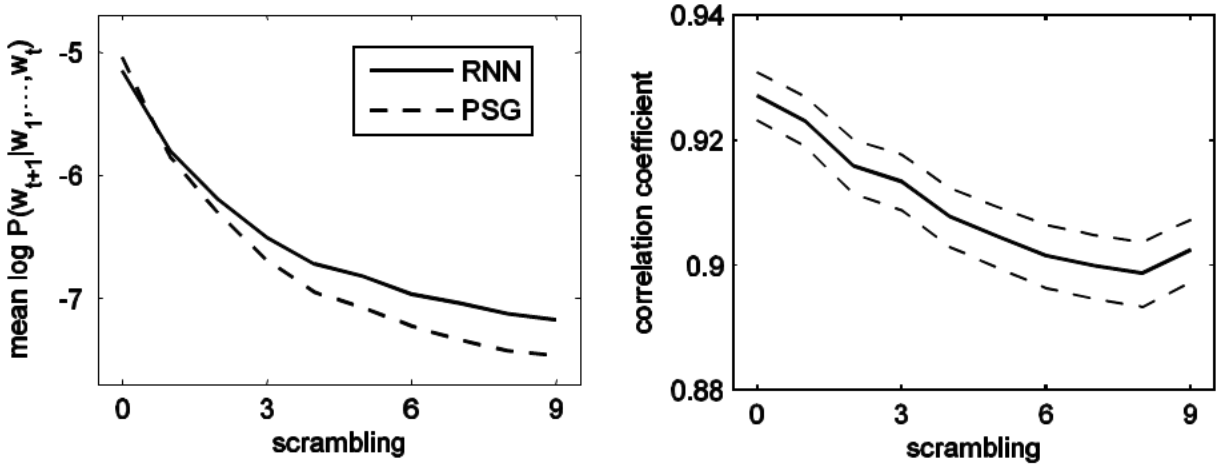

Figure 3. Left: average log-probability estimate as a function of scrambling level. Right: coefficient of correlation between RNN's and PSG's log-probability estimates, as a function of scrambling level (dashed lines indicate the 95% confidence interval).

4. Discussion

Compared to the connectionist RNN model, the symbolic PSG generalizes better to grammatical sentences and (correctly) estimates lower next-word probabilities for scrambled sentences. Both these results suggest that the symbolic model is indeed more systematic than the connectionist one. However, the difference between the two models' performance is one of degree rather than categorical. Both models show a fairly slow decrease (instead of a sudden drop) in prediction accuracy as scrambling level increases. Crucially, the correlation between the two models' next-word probability estimates is very strong, which means that the models behave similarly.

Why would models that are so different in their underlying assumptions nevertheless show such remarkably similar behavior? Note that they are faced with the same, complex computational problem: to extract linguistic patterns from 7.6 million word tokens and to apply the discovered statistics when processing novel sentences. In terms of Marr's (1982) famous levels of description, the models differ strongly at the algorithmic level but they are similar at the computational level, which specifies the task to be performed. As Marr argued, task requirements, rather than representations and algorithms, often form the most important factor in shaping behavior. Since the models need to perform the same task, they display similar observed behavior.

Interestingly, to the (small) extent that the models do differ, the RNN seems to be closer to cognitively reality: A comparison between the two models' surprisal estimates and human word-reading times (over the same test sentences) revealed that the RNN's surprisals explained significantly more variance in reading time, whether these were collected by self-paced reading (Fernandez Monsalve et al., 2012) or by eye-tracking (Frank & Thompson, 2012). This is in line with earlier findings using different models and sentences (Frank & Bod, 2011). A similar result emerged from a sentence-reading

experiment in which participants' brain activity was recorded by EEG. Although both RNN- and PSG-based surprisal values predicted the size of the N400 event-related potential component, the RNN model yielded more accurate predictions (Frank, Otten, Galli, & Vigliocco, 2013). In fact, the PSG did not account for any unique variance in N400 size over and above the RNN's predictions. This strongly suggests that, despite being slightly less systematic, the RNN is the more accurate cognitive model.

Compared to earlier demonstrations of models' (lack of) systematicity, the simulations presented here were based on a much more complex and realistic task. Nevertheless, it was still far from what people face in the real world. For example, the amount of training data was far smaller than what children learn language from, and the input stream was noise-free and pre-segmented into words. Also, the models "merely" needed to learn syntax so they could ignore phonology, prosody, co-speech gesture, pragmatic constraints, the discourse setting, and any extra-linguistic context. If noisy and ambiguous information from all these different sources must be integrated in real time, this puts further pressure on the system by narrowing the gap between how well it *needs to* perform and how well it *can* perform. It is only within this narrow gap that any difference in systematic abilities can occur.

So, to what extent do the models display systematicity? In absence of any gold standard of systematicity to compare the models' performance with, this question cannot be answered. All we know is that the RNN and PSG are similarly (un)systematic. If symbol systems are indeed necessarily systematic (Fodor & McLaughlin, 1990), then the PSG must be systematic and, therefore, so is the RNN (or, at least, it is 'almost' systematic, if some arbitrary cutoff-point is introduced just below the PSG's level of performance). Conversely, if systematicity is indeed beyond the reach of connectionist models (Fodor & Pylyshyn, 1988), then the RNN cannot be systematic and, therefore, neither is the PSG (or, at best, it is minimally systematic).

It may seem unlikely that a PSG could fail to be systematic. After all, provided that *Mary* and *John* belong to the same syntactical category (which is the case in the PSG used here), a grammar trained on *John loves Mary* necessarily generalizes to *Mary loves John*, thereby realizing Fodor and Pylyshyn's (1988) standard example of systematic behavior. Moreover, grammars as specified by production rules have traditionally been viewed as typical examples of symbol systems, with full-fledged systematic abilities against which generalization by connectionist models is evaluated (e.g., Christiansen & Chater, 1994; Frank, 2006a; Van der Velde et al., 2004). As argued by Aizawa (1997b), however, even classical symbols systems are only systematic when combined with a proper mechanism for manipulating the symbols. Perhaps our PSG lacks such a systematic algorithm, but this does raise the question why that would be so. Roark's (2001) algorithms, which were used

here for inducing the grammar and obtaining surprisal estimates, certainly do not intend to reduce the grammar's ability to generalize to new input (i.e., to display systematicity).

Possibly, then, the PSG's statistical nature somehow reduces its systematic abilities to that of a connectionist model. Although this may indeed be the case, it is likely that there is no viable alternative. Arguably, only a statistical system can learn to generalize appropriately from very complex natural data, without failing when faced with a highly unexpected event, in a noisy, ambiguous, widely varying real-world environment that provides a massive number of potentially inconsistent cues. As a case in point, it is only since the advent of statistical methods that Computational Linguistics has made significant progress in automatic natural language processing (cf. Manning & Schütze, 1999). And indeed, the probabilistic nature of human cognition is increasingly recognized in Cognitive Science (cf. Oaksford & Chater, 2007). So, if we do need to choose between a statistical model and a systematic model, the safest bet may well be against systematicity.

5. Conclusion

As far as systematicity is concerned, there may not be any important difference between connectionist and statistical symbolic models, as long as the models are powerful enough to perform real-world tasks. This is not to say that the difference between model types is not of interest. On the contrary, it is quite relevant to Cognitive Science which representations and algorithms best describe the language processing system. There will certainly be significant differences among models regarding, for example, their learning and processing efficiency, their ability to connect to non-linguistic cognitive modalities, their performance under adverse conditions, and the ability to explain (psycho)linguistic phenomena and neuropsychological disorders. It is those kind of issues, rather than systematicity, on which the discussion about the value of Connectionism should focus. When facing reality, systematicity is not something worth worrying about.

Acknowledgments

I am thankful to Gideon Borensztajn for his comments on an earlier version of this chapter. The research presented here was funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant number 253803. The UCL Legion High Performance Computing Facility, and associated support services, were instrumental in the completion of this work.

References

Aizawa, K. (1997a). Exhibiting versus explaining systematicity: a reply to Hadley and Hayward. *Minds and Machines*, *7*, 39–55.

Aizawa, K. (1997b). Explaining systematicity. *Mind & Language*, *12*(2), 115–136.

Bechtel, W. (1993). The case for connectionism. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *71*, 119–154.

Bod, R., & Smets, M. (2012). Empiricist solutions to nativist puzzles by means of unsupervised TSG. In *Proceedings of the Workshop on Computational Models of Language Acquisition and Loss* (pp. 10–18). Avignon, France: Association for Computational Linguistics.

Bodén, M. (2004). Generalization by symbolic abstraction in cascaded recurrent networks. *Neurocomputing*, *57*, 87–104. doi:10.1016/j.neucom.2004.01.006

Borensztajn, G., Zuidema, W., & Bod, R. (2009a). The hierarchical prediction network: towards a neural theory of grammar acquisition. In N. A. Taatgen & H. Van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2974–2979). Austin, TX: Cognitive Science Society.

Borensztajn, G., Zuidema, W., & Bod, R. (2009b). Children's grammars grow more abstract with age: evidence from an automatic procedure for identifying the productive units of language. *Topics in Cognitive Science*, *1*(1), 175–188. doi:10.1111/j.1756-8765.2008.01009.x

Boston, M. F., Hale, J., Patil, U., Kliegl, R., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*, 1–12.

Brakel, P., & Frank, S. L. (2009). Strong systematicity in sentence processing by simple recurrent networks. In N. A. Taatgen & H. Van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1599–1604).

Brouwer, H., Fitz, H., & Hoeks, J. (2010). Modeling the noun phrase versus sentence coordination ambiguity in Dutch: evidence from surprisal theory. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics* (pp. 72–80). Uppsala, Sweden: Association for Computational Linguistics.

Bybee, J. (2003). Cognitive processes in grammaticalization. In M. Tomasello (Ed.), *The New Psychology of Language, Volume II* (pp. 145–167). New Jersey: Lawrence Erlbaum.

Chang, F. (2002). Symbolically speaking: a connectionist model of sentence production. *Cognitive Science*, *26*, 609–651.

Christiansen, M. H., & Chater, N. (1994). Generalization and connectionist language learning. *Mind & Language*, *9*, 273–287.

Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*, 157–205. doi:10.1016/S0364-0213(99)00003-8

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210. doi:10.1016/j.cognition.2008.07.008

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195–225.

Farkaš, I., & Crocker, M. W. (2008). Syntactic systematicity in sentence processing with a recurrent self-organizing network. *Neurocomputing*, *71*, 1172–1179.

Fernandez Monsalve, I., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In W. Daelemans (Ed.), *Proceedings of the 13th conference of the European chapter of the Association for Computational Linguistics* (pp. 398–408). Avignon, France: Association for Computational Linguistics.

Fitz, H., & Chang, F. (2009). Syntactic generalization in a connectionist model of sentence production. In J. Mayor, N. Ruh, & K. Plunkett (Eds.), *Connectionist models of behaviour and cognition II: Proceedings of the 11th Neural Computation and Psychology Workshop* (pp. 289–300). River Edge, NJ: World Scientific Publishing.

Fodor, J. A., & McLaughlin, B. P. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, *35*(2), 183–204.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition*, *28*, 3–71.

Frank, S. L. (2006a). Learn more by training less: systematicity in sentence processing by recurrent networks. *Connection Science*, *18*(3), 287–302. doi:10.1080/09540090600768336

Frank, S. L. (2006b). Strong systematicity in sentence processing by an Echo State Network. In S. D. Kollias, A. Stafylopatis, W. Duch, & E. Oja (Eds.), *Proceedings of the 16th international conference on Artificial Neural Networks, Part I* (Vol. 4131, pp. 505–514). Berlin: Springer.

Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive effort in sentence comprehension. *Topics in Cognitive Science*, *5*, 475–494.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*(6), 829–834. doi:10.1177/0956797611409589

Frank, S. L., & Čerňanský, M. (2008). Generalization and systematicity in echo state networks. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 733–738). Austin, TX: Cognitive Science Society.

Frank, S. L., Fernandez Monsalve, I., Thompson, R. L., & Vigliocco, G. (in press). Reading-time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*.

Frank, S. L., Haselager, W. F. G., & van Rooij, I. (2009). Connectionist semantic systematicity. *Cognition*, *110*(3), 358–379. doi:10.1016/j.cognition.2008.11.013

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2013). Word surprisal predicts N400 amplitude during reading. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria: Association for Computational Linguistics.

Frank, S. L., & Thompson, R. L. (2012). Early effects of word surprisal on pupil size during reading. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1554–1559). Austin, TX: Cognitive Science Society.

Hadley, R. F. (1994a). Systematicity in connectionist language learning. *Mind & Language*, *9*(3), 247–272.

Hadley, R. F. (1994b). Systematicity revisited: reply to Christiansen and Chater and Niklasson and van Gelder. *Mind & Language*, *9*, 431–444.

Hadley, R. F., Rotaru-Varga, A., Arnold, D. V., & Cardei, V. C. (2001). Syntactic systematicity arising from semantic predictions in a Hebbian-competitive network. *Connection Science*, *13*(1), 73–94.

Hale, J. T. (2001). A probabilistic Early parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics* (Vol. 2, pp. 159–166). Pittsburgh, PA: Association for Computational Linguistics.

Jansen, P. A., & Watter, S. (2012). Strong systematicity through sensorimotor conceptual grounding: an unsupervised, developmental approach to connectionist sentence processing. *Connection Science*, *24*, 25–55.

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics* (pp. 423–430).

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schlökopf, J. Platt, & T. Hoffman (Eds.), *Advances in Neural Information Processing Systems* (Vol. 19, pp. 849–856). Cambridge, MA: MIT Press.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Marcus, G. F. (2001). *The algebraic mind: integrating connectionism and cognitive science*. Cambridge, MA: MIT Press.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman and Company.

McClelland, J. L., St.John, M. F., & Taraban, R. (1989). Sentence comprehension: a parallel distributed processing approach. *Language and Cognitive Processes*, *4*, 287–335.

Miikkulainen, R. (1996). Subsymbolic case-role analysis of sentences with embedded clauses. *Cognitive Science*, *20*, 47–73.

Monner, D. D., & Reggia, J. A. (2011). Systematically grounding language through vision in a deep, recurrent neural network. In *4th International Conference on Artificial General Intelligence* (Vol. 6830, pp. 112–121).

Niklasson, L. F., & Van Gelder, T. (1994). On being systematically connectionist. *Mind & Language*, *9*, 288–302.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford, UK: Oxford University Press.

Phillips, S. (1998). Are feedforward and recurrent networks systematic? Analysis and implications for a connectionist cognitive architecture. *Connection Science*, *10*, 137–160.

Phillips, S. (2000). Constituent similarity and systematicity: The limits of first-order connectionism. *Connection Science*, *12*, 45–63.

Pinker, S., & Prince, A. (1988). On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*, 73–193.

Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Computational Linguistics*, *27*, 249–276.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*, 302–319.

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*, 21–46.

Van Gelder, T. (1990). Compositionality: a connectionist variation on a classical theme. *Cognitive Science*, *14*, 355–384.

Voegtlin, T., & Dominey, P. F. (2005). Linear recursive distributed representations. *Neural Networks*, *18*, 878–895.

Wong, F. C. K., & Wang, W. S.-Y. (2007). Generalisation towards combinatorial productivity in language acquisition by simple recurrent networks. In *International Conference on Integration of Knowledge Intensive Multi-Agent Systems* (pp. 139 –144).