Postmodern linguistics and the prospects of neural syntax: Some polemical remarks[1]

HANS-MARTIN GÄRTNER and BRYAN JURISH

Let us assume for the sake of argument that Martin Stokhof and Michiel van Lambalgen (S&vL) have correctly identified methodological flaws that obstruct progress in modern linguistics[2] and let us therefore comment on one of the alternatives S&vL point out as potentially promising. For reasons of curiosity we choose "Neural Syntax" (NS) (Fitz 2009), which instantiates one of the "approaches in which neuronal models of language acquisition and language use are studied." It is important for us to stress right at the outset that we cannot do justice to the intricacies of NS. In fact, we will be rather selective in focusing on issues that seem to us to cast doubt on the NS approach.

To begin with, let us consider the kinds of motivation that drive connectionist theories like NS:

> Specifying the model's learning environment, input/output encoding, and learning procedures does not involve theory-laden assumptions about the syntax underlying the target language. Neural network models 'find' syntactic representations autonomously in the process of generating a solution to a computational problem and these representations are not preconceived by the experimenter nor do they necessarily map onto the syntactic categories postulated by descriptive linguistics. Because of domain-general learning and the autonomy of representations, neural network models *prima facie* are ideally suited for modelling syntactic development and sentence processing. (Fitz 2009: 4)

Stated in a more provocative way, the ultimate aim of approaches like NS is to show that the results of "descriptive linguistics" concerning syntax are at best

[1] H.M.G. was supported by BMBF (Grant Nr. 01UG0711) and B.J. by a DFG-grant to the project "Deutsches Textarchiv."

[2] It seems to us that a fair portrayal of foundational assumptions made in generative linguistics would have profited from taking into account Chomsky (1975 [1955]: Chapters I–VI).

Theoretical Linguistics 37–1/2 (2011), 37–44 0301–4428/11/0037–0037
DOI 10.1515/THLI.2011.003
© Walter de Gruyter

epiphenomena of neural learning and processing. On this count, one has to study the degree to which NS has been able to show this on the basis of mechanisms that themselves are free of specifically syntactic assumptions.

A central controversy − familiar, of course, from earlier discussions of connectionist models (e.g. Fodor & Pylyshyn 1988) − concerns the status of syntactic constituents:

> The preceding analysis indicates that the Dual-path model did not develop robust internal representations of phrasal categories in the sense of traditional theory of syntax. I found some coherence among verb phrase elements, volatile clustering within prepositional phrases, and perfect separation within noun phrases. Yet, the model acquired a structurally complex target language with relative clauses and substantially generated beyond its immediate linguistic experience. What this suggests, if anything, is that the representation of phrasal categories is not required to successfully transduce between meaning representations and grammatical sentence forms. (Fitz 2009: 142)[3]

A noun phrase (NP) in NS is a simple determiner+head noun (D+N) pair as well as such a pair immediately followed by a relative clause. Networks in NS are trained to put NPs into the argument slots of intransitive, transitive and ditransitive constructions which can undergo "transformations" like the active/passive alternation. How is this possible if representationally there is "perfect separation within noun phrases"? The answer is that NS networks are trained on input that represents the core meaning of a clause in terms of thematic (macro-)role features "bound" to concept features. Thus, *the cat chased the dog* is the output of a network that binds *X(Agent)* to concepts *THE* and *CAT*, *A(Action)* to *CHASE*, and *Y(Theme)* to *THE* and *DOG* (Fitz 2009: 63).[4] Passivization to *the dog is chased by the cat* is the result of lowering activation of the *X*-role wrt to the *Y*-role on a separate layer. The gist of this method can be stated as follows:

> The results I presented from this approach suggest that the *structure of meaning* may obviate the need for [. . .] syntax-specific knowledge [. . .]. (Fitz 2009: 251)

This, of course, is suggestive of a hidden Montagovian agenda, on which a semantic and a syntactic algebra are connected by homomorphism (Partee &

---

3 "The Dual-path model is a model of sentence production and syntactic development. [. . .] it consists of an SRN [simple recurrent network; HMG&BJ] extended with a system to represent sentence meaning. Both components − the SRN and the meaning system − are arranged to form separate information channels or *pathways*" (Fitz 2009: 52).

4 Concepts have been trained to be associated to words in a one-to-one fashion.

Hendriks 1997: Section 3.1). The more transparent the homomorphism is, the more "trivial" becomes the "elimination of syntax," i.e. the direct association of surface forms with semantic representations. That NS has so far been able to work with much simpler meaning representations is clearly related to its much narrower scope. Treating determiners as "concepts," for example, on a par with concepts like *CHAIR* and *TABLE*, is only possible because the combinatorics of determiners is also controlled by *n*-gram learning in the additional sequencing layer. This enforces realization of D+N strings once the underlying concepts are activated in the meaning layer.[5] Equally, adjacency of head noun and relative marker *that* is a consequence of bigram learning controlled by an extra feature that marks the head noun as an attachment site for a relative clause. Finally, integrity of the relative clause is ensured by co-indexing the meaning constituents of different clauses. Thus, *the cat chased the dog that stole the cake* is triggered by features *0X(Agent)* >> (= bound to) *THE*,*CAT; 0A(Action)* >> *CHASE; 0Y(Theme)* >> *THE*,*DOG; 1X(Agent)* >> *THE*,*DOG; 1A(Action)* >> *STEAL; 1Y(Theme)* >> *THE*,*CAKE;* and *0Y(Relativized)*.[6] What remains to be shown, however, is the architecture required for more involved NP configurations such as NPs with multiple modifiers (and spreading number, gender, and case agreement), relatives with multiple heads ("hydras") (Link 1998), as well as discontinuous NPs resulting, among other things, from relative clause extraposition and left-branch extractions.

"Coherence among verb phrase elements," on the other hand, solely concerns cooccurrence patterns involving main verbs, auxiliaries, and participles (Fitz 2009: 139). All of these cluster around the element controlled from the *A(Action)* node in meaning representation. This, of course, shows nothing about the ultimate dispensability of complex VP-structure as long as the system hasn't begun to model things like VP-fronting and VP-ellipsis:

---

5 Preventing determiners from following nouns, i.e. preventing N+D sequences, is actually more involved. The decision to start with *the cat chased* . . . instead of *cat the chased* . . . seems (in part) to be due to the fact that "all sentences in the input started with either a definite or indefinite article" (Fitz 2009: 64). It would be important to see what happens when bare nouns and inversions (*Broccoli the boss doesn't like*) enter the input. This would also raise worries about generating unwelcome preverbal subject-object-seqences (* *The boss broccoli doesn't like*) etc.

6 The feature responsible for attachment of relative clauses would actually be written *0YYT*, saying that the main clause theme is also "topic" (T) of the event described by the relative clause. An additional feature is responsible for suppression of *1X(Agent)*, i.e. for creating a gap inside the relative clause (Fitz 2009: 96).

(1) a. *Mary wanted to read Ulysses, and read Ulysses she did*
b. A: *Mary wanted to read Ulysses*.
B: *She did!*

Let us next turn to some slightly more abstract matters and look at the inner workings of NS networks. The following passage describes node activation patterns for the relative clause in (2):

(2) *A woman that the boy played with is hitting the father*

> In the verb position, the embedded action role *1A* wins the competition. There is residual activation of the *1Y* role and rather strong activation of the *1X* role. The *1X* role is not linked to conceptual content in oblique constructions so there are no causal consequences to activating this role. Most likely, the model is preparing to sequence this role in the post-verbal slot because in the passive transitive construction the agent role *1X* succeeds the *1Y* role. If this is the correct explanation, this behavior indicates that the Dual-path model is sensitive to statistical regularities in two-role chunks within clauses [. . .]. (Fitz 2009: 134f.)

Crucially, intransitives like *play* are analyzed with theme subjects, i.e., they are agent-less. Thus, the relative clause in (2) would be triggered by *1Y(Theme)* >> THE,BOY; *1A(Action)* >> PLAY; and *1Z(Oblique)* >> A,WOMAN. Additional activation of *1X(Agent)* may indeed be an interesting by-product of statistical learning. However, it may also have detrimental "causal consequences" when combined with an insightful account of reflexivization and control by "implicit (agent) arguments," which is dealt with in terms of "unassigned theta roles" (e.g. *The ship was sunk to collect the insurance*) (Lasnik 1988; Williams 1987). Spurious agents could lead to all kinds of serious overgeneration here. Of course, one will have to await an NS analysis of such phenomena to see whether our hunch is right.

So far we have pointed at tensions that threaten to arise between the NS models and adequacy conditions set by "competence grammars." There are, however, problematic issues for NS in the area of acquisition and processing of syntax as well. Thus, it must be considered disappointing that the model is unable to generalize from input structures with relative clause embedding of depth 1 to structures with greater depth of embedding, i.e. structures where relatives are embedded in relatives (Fitz 2009: 194). This is particularly interesting, given that such structures would be amenable to iterative instead of recursive generation, as the famous *This is the cat that ate the rat that stole the cheese* . . . indicates.

Equally disappointing is the fact that subject-auxiliary inversion around complex subjects in polar questions could not be learned by NS networks from being exposed to subject-auxiliary inversion in simple polar questions and declaratives with complex subjects. Instead, generating structures like (3a) could only be induced by putting subject-auxiliary inversion around complex subjects in constituent questions, (3b), into the input (Fitz 2009: 224).

(3) a. *Is the cat that ate the rat sitting on the mat?*
b. *What is the cat that ate the rat sitting on?*

This is worrisome because it hasn't been shown how (3b) itself is learned. The issue is of wider importance as it is supposed to contribute a building block to showing the famous argument for an innate structure-dependency assumption by Chomsky (cf. Crain & Nakayama 1987) to be flawed. Let us therefore briefly look at an attempt by Fitz (2010) to improve on the situation.

Fitz (2010) develops an "Adjaceny-Prominence" (AP) model (Chang, Lieven & Tomasello 2008) for generating complex polar questions such as (3a) from a bag of words based on traditional bigram ("adjacency") and precedence ("prominence") statistics gathered from an artificial training set. While the experiments were explicitly designed to preclude objections on the basis of an impoverished training set (e.g. Kam, et al. 2008 on Reali & Christiansen 2005), no provision at all is made for constiuent questions such as (3b).

Auxiliary fronting as such is easily accounted for in the AP model (as in Reali & Christiansen 2005) by the empirical necessity in the training set of an inflected verb immediately following the sentence-initial "seed marker" *[quest]* indicating a question. Clearly, such a necessity would no longer obtain if constituent questions were included in the training data.

That the embedded auxiliary is (mostly) sequenced correctly can be attributed to the use of (cumulative) precedence statistics: in the training data, every predicate is preceded by at least one auxiliary, so the AP model strongly favors sequencing all auxiliaries before all predicates as in (4a) rather than the ungrammatical (4b), in which the main auxiliary follows the embedded predicate.

(4) a. *Is the cat that is loud __ hungry?*
b. **Is the cat that __ loud is hungry?*

It remains an open question whether an AP model can successfully learn to sequence complex polar questions for languages such as German and Dutch with auxiliary-final relatives:

(5) a. **Ist die Katze die ist laut hungrig?*
b. *Ist die Katze die laut ist hungrig?*

The tendency of the AP model to sequence auxiliaries before predicates is also observable in the (comparatively few) errors it makes on the artificial test set, in particular the generation of grammatical subject-relatives, (6b), in place of object-relatives, (6a), in embedded transitive constructions:

(6) a. *Is the rat [that the cat is chasing] tasty?*
b. ☹ *Is the rat [that is chasing the cat] tasty?*

The Trigram model often converted [object-relativized questions] into grammatical [subject-relatived] questions. The AP-learner also made such conversion errors, but less frequently. (Fitz, 2010: 2696)[7]

Fitz (2010) offers no suggestions regarding how this bias might be overcome by an AP model; and indeed it is difficult to imagine how bigram and precedence relations alone could be made to account for the differences between (6a) and (6b) without reference to some representation of the constituent structure of the target utterances, whether explicitly as in traditional formal grammars or implicitly as in the the semantic role encoding used in NS.[8,9]

A final observation concerns the role accorded to frequency in NS. Certain passages are suggestive of a full-scale endorsement of "real" frequencies from realistic corpora (e.g. Fitz 2009: Section 8.5). Other passages make it quite clear that the NS-model has to be shielded from "rich" and varied input in order to produce any (reasonable) results, as, for example, the following statement indicates (cf. also Footnote 5):

For statistical learning systems like the recursive Dual-path model, low frequency can be detrimental. (Fitz 2009: 267)

Given the growing influence of belief in frequencies one would have liked a more comprehensive reflection on its proper place in the world of NS.

7 The trigram model is a purely *n*-gram based statistical learning model whose behavior was contrasted with the behavior of the AP-model for heuristic reasons.

8 The notion of meaning (constraints) underlying the AP model − fixing a multiset of words and collecting linear precedence statistics − strikes us as rather far removed from standard perspectives on the matter (e.g. Hovy 1988).

9 It would be interesting to contrast the AP-model with important related work such as done by de Marcken (1995; 1996) on induction of linguistic structure from string-based statistics and by Heinz (2010) and Heinz & Rogers (2010) on strictly piecewise distributions.

Let us finish by suggesting a re-contextualization of S&vL's endorsement of NS. Criticizing the (generative) foundations of "modern linguistics" has clearly become a trend, which recently culminated in the debate provoked by Evans & Levinson (2009) and a call for a new age ("sea change") (Levinson & Evans 2010). What at least some of their opponents seem to have overlooked is that the original paper was not addressed to linguists but to cognitive scientists. NS can also be read as, in the main, instructions for how formal modeling can make more precise and enlighten debates on the psychology of language. This fact allows the conjecture that − like in other fields (Habermas 1985; 1988) − *modern* linguistics is envisaged by S&vL to coexist with *postmodern* linguistics, the latter a field in which people interested in antecedent contained deletion, remnant scrambling, and LF-pied-piping do not dwell.

Correspondence address: Zentrum für Allgemeine Sprachwissenschaft, Berlin
gaertner@zas.gwz-berlin.de
Berlin-Brandenburgische Akademie der Wissenschaften
jurish@bbaw.de

References

Chang, Franklin, Elena Lieven & Michael Tomasello. 2008. Automatic evaluation of syntactic learners in typologically-different languages. *Cognitive Systems Research* 9. 198–213.
Chomsky, Noam. 1975 [1955]. *The logical structure of linguistic theory*. New York: Plenum.
Crain, Stephen & Mineharu Nakayama. 1987. Structure dependence in grammar formation. *Language* 63. 522–543.
de Marcken, Carl. 1996. Linguistic structure as composition and perturbation. *ACL* 34. 335–341.
de Marcken, Carl. 1995. Lexical heads, phrase structure and the induction of grammar." *SIGDAT*. 14–26.
Evans, Nicholas & Stephen Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32. 429–492.
Fitz, Hartmut. 2009. *Neural syntax*. Amsterdam: ILLC dissertation.
Fitz, Hartmut. 2010. Statistical learning of complex questions. In Stellan Ohlsson & Richard Catrambone (eds.), *Proceedings of the 32nd annual conference of the cognitive science society*, 2692–2698. Austin TX: Cognitive Science Society.
Fodor, Jerry & Zenon Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition* 28. 3–71.
Habermas, Jürgen. 1985. *Der philosophische Diskurs der Moderne*. Franfurt/M.: Suhrkamp.
Habermas, Jürgen. 1988. *Nachmetaphysisches Denken*. Frankfurt/M.: Suhrkamp.
Heinz, Jeffrey. 2010. String extension learning. *ACL* 48. 897–906.
Heinz, Jeffrey & James Rogers. 2010. Estimating strictly piecewise distributions. *ACL* 48. 886–896.

Hovy, Eduard. 1988. *Generating natural language under pragmatic constraints*. Hillsdale NJ: Lawrence Erlbaum.

Kam, Xuân-Nga Cao, Iglika Stoyneshka, Lidiya Tornyova, Janet Fodor & William Sakas. 2008. Bigrams and the richness of the stimulus. *Cognitive Science* 32. 771–787.

Lasnik, Howard. 1988. Subjects and the *Θ*-criterion. *Natural Language and Linguistic Theory* 6. 1–17.

Levinson, Stephen & Nicholas Evans. 2010. Time for a sea-change in linguistics: Response to comments on 'The myth of language universals'." *Lingua* 120. 2733–2758.

Link, Godehard. 1998. Hydras. On the logic of relative clause constructions with multiple heads. In *Algebraic semantics in language and philosophy*, 77–88. Stanford CA: CSLI-Publications.

Partee, Barbara & Herman Hendriks. 1997. Montague grammar. In Johan van Benthem & Alice ter Meulen (eds.), *Handbook of logic and language*, 5–91. Amsterdam: Elsevier.

Reali, Florencia & Morten Christiansen. 2005. Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science* 29. 1007–1028.

Williams, Edwin. 1987. Implicit arguments, the binding theory, and control. *Natural Language and Linguistic Theory* 5. 151–180.