Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Physics of Life Reviews ••• (••••) •••–•••

PHYSICS of LIFE reviews

www.elsevier.com/locate/plrev

Comment

Syntactic networks, do they contribute valid information on syntactic development in children? Comment on “Approaching human language with complex networks” by J. Cong and H. Liu

Anat Ninio

Received 7 May 2014; accepted 16 May 2014

Communicated by L. Perlovsky

In the target article [1] Cong and Liu provide a clear and informative introduction to the use of complex networks in research studying language. I would like to add the perspective of a researcher of language acquisition who has been hopeful that network theory illuminates processes of development [2,3], but feels a certain difficulty with studies applying network analysis to the development of syntax.

In syntactic networks, nodes in the network represent each different word used in a certain collection of texts, while edges are said to represent the syntactic relations existing between pairs of words. However, the conclusions arrived at regarding syntactic development from analysis of such networks raise fears that something is not quite successful in the modeling employed.

First, we know children’s syntactic advance is first and foremost making their syntactically connected sentences longer [4,5]. At the start of multiword speech, young children typically produce only short, two-word and three-word long sentences for a considerable time, only several months later increasing the mean length of the sentences they are producing. Although this is a crucial element of syntactic development in normal children, the modeling technique employed in these projects is unable in any way to reflect the developmental advance from short sentences to longer sentences. In fact, syntactic networks do not represent the length of the sentence from which the linked word-couples are taken in any manner. Instead, the network-graphs represent syntactic structure as if it were only the relation between two words that form a pair, namely, as a collection of single dependency pairs, but does not represent the structure of the sentences beyond the word-couple. It is as if the researchers cut all texts into two word pieces and mapped them into the network. The result is that the network is the same if the speaker produced only two-word long sentences and if he or she produced five or ten-word long ones. However, the true syntactic structure of a ten-word-long sentence is not identical to the structure of nine two-word sentences into which it is virtually “cut into” in the present technique, and the developmental achievement is not identical if a child only produces two-word utterances than if she produces ten-word long ones. This crucial aspect of normal development is completely un-represented in the projects modeling syntactic development in children by network analysis.

For instance, let us take the two sentences *need it* and *my need it* said by the same child and used as an illustration of the method of building the network ([6], Fig. 1). The syntactic relations of the two sentences are represented in the network as three nodes, each for the words *need*, *it* and *my*, connected by two links (one connecting *need* and *it*, the

DOI of original article: http://dx.doi.org/10.1016/j.plrev.2014.04.004.

http://dx.doi.org/10.1016/j.plrev.2014.05.004

other connecting *my* and *need*). There is no representation of the fact that *my need it* is a three-word sentence with two different dependency relations forming a tree-structure, whereas *need it* is a two-word sentence, with a single dependency relation in it.

As another example, let us take the second sentence used in the same Fig. 1: *Telephone go right there*. The representation of the syntax of this sentence in the syntactic network built by the authors is identical to the representation a series of two-word utterances would get: *telephone go*, *go there*, and *right there*. (These are the three pairs of words actually connected by dependency relations in the original four-word sentence.)

Let us assume that the child produced the three short sentences at Age One, and the long sentence, at an older Age Two. The change in the syntactic complexity of the child's sentences would not get any representation in the syntactic networks build by this method for the two ages – they would be absolutely identical.

The problem is even more acute in the case of children with language impairment. For instance, if a child with a language impairment continues to produce very short sentences for years and never advances to longer ones, this technique would be unable to show the nature of her syntactic disability. For example, the child called in the literature Genie never advanced past very short sentences, although she did learn a vocabulary of considerable size. If we mapped her sentences to a network of the kind used by the researchers in these projects, we would never be able to realize that Genie couldn't produce any but very short sentences, nor represent her disability by a formal quantitative measure. For instance, let us assume that Genie produced the three short sentences in the examples above and a child with normal language development, the long one. Using the present technique of building syntactic networks, there is no way to observe the difference, and hence characterize her language impairment.

I see a second problem for developmental significance in the modeling of syntactic relations with the specific words used. During the period when children first learn to combine words into syntactically connected sentences, they also learn new words, and their vocabulary increases significantly within a short time [7]. It appears, however, that using network-based measures demonstrates a negative correlation between a child's vocabulary size and the amount of connectivity in his syntax [8,9]. Thus, children with a smaller lexicon display a network with higher connectivity and children with a larger lexicon, a lower connectivity. This finding is very surprising and against the grain of what we know about the **positive** correlation of vocabulary size and syntactic development [10]. Close examination of the relevant projects reveals that this is not a true developmental finding but an artefact of the method of constructing networks.

To see the artefact, let us compare two children who master just one single syntactic relation: that of a transitive verb getting a direct object. Child A is a child with a large vocabulary; she is a so-called "nominal" child who has learned many different names of objects [11]; Child B is a "pronominal" child who uses many pronouns but few nouns. Let's now assume that both children generate verb-object sentences using the same verbs popular with young children, e.g., *want, open, take, close, see, fix*, and so forth. Child A will say things like *want banana, open door, take bottle, close window, see bird* and *fix TV*. Child B will say *want it, open it, take it, close it, see it* and *fix it*. When we compute the average degree – a measure of connectivity in the network – we find that the network of Child A with the large vocabulary has a really small connectivity, with no degree larger than one. By contrast, the network of Child B with the small vocabulary has a very high connectivity, with all his verbs linked through the shared word *it*. The problem is that in any other measure of syntactic knowledge the two children are identical; their mean length of utterance is 2.0; they produce the same number of syntactic relations (one) and the same number of distinct syntactic constructions (six). The much higher measure of connectivity, that is, the high average in-degree exhibited by the syntactic network of the child who uses a pronoun as a direct object, is a meaningless fact as far as the child's syntactic knowledge is concerned. This child knows no more syntax than Child A; the elegant connectivity measure we got from the network analysis teaches us nothing about the two children's acquisition of syntax. Unfortunately, the studies demonstrating the reputed negative correlation between vocabulary size and connectivity fall into this trap, so that, for instance, the child Ruth with the low vocabulary and high average degree of her syntactic network in Ke and Yao's study [8] is a typical noun-avoiding child while Joel of the same study with his large vocabulary and low average degree in his syntactic network is a typical referential noun-loving child, as the authors themselves report in the relevant article (on p. 94).

By the requirement that two words are connected when and only when they share a specific word as a syntactic associate, connectivity becomes a measure of the semantic homogeneity of a speaker's vocabulary, not of the speaker's mastery of syntax. In a small speech sample such as a young child's, when words do not share specific syntactic associates, they tend to generate isolated sub-graphs. Their presence reduces the measured degree of connectivity of the network. This means that if a young child has a large vocabulary, containing semantically diverse items, the syntactic network generated by this technique from his word-combinations will tend to present many isolated sub-graphs and hence a low connectivity. When the child uses general terms such as pronouns and demonstratives, this boosts measures of connectivity. Unfortunately, this technical artefact voids most of the conclusions reached in these papers regarding the developmental changes supposedly occurring in children's syntactic system.

The last problem concerns a methodological shortcut often taken in studies of syntactic networks and that is taking collocation (or co-occurrence) as an approximation to syntactic relations. In a collocation network sentences are not coded for grammatical relations among words. Instead, two words are deemed to be connected if they are direct neighbors in a sentence or, sometimes, if they appear one word apart from each other. The advantage of this shortcut is that it makes it possible to automatically analyze large written corpora and derive networks from them, whereas true syntactic dependency requires syntactic annotation, which is costly and time consuming. This methodology is also employed in studies describing syntactic development under the assumption that simply marking linear sequencing is a good substitute for grammatical analysis [8,9]. However, the methodology is problematic, as it creates two kinds of errors.

First, above 50% of syntactic dependency relations in different types of languages are between pairs of words which are not adjacent in sentences; and increasing the window of collocation to two words may only cover 60–90% of syntactic relations, dependent on the language [12].

Second, adjacent words are often not grammatically related; including all words within a two-word radius from a target results in 50% false identification of syntactic relations between unrelated words [13]. For example, the verb form *likes* in the sentence *John likes very cold milk* does not have a direct grammatical relation with the words *very* and *cold*, but in a collocation network they would be treated as syntactically linked. Thus, studies purporting to map syntactic development in children that employ this method of modeling, introduce a substantial level of error, with unknown consequences for their findings.

In conclusion, work on syntactic networks may need some additional methodological effort before it can provide valid information on children's developing syntactic knowledge. A similar caveat may probably apply to all of the studies pertaining to syntactic networks reviewed in this article and to their conclusions. I am certain that the field will provide the necessary adjustments and I am looking forward to them.

References

[1] Cong J, Liu H. Approaching human language with complex networks. Phys Life Rev 2014. http://dx.doi.org/10.1016/j.plrev.2014.04.004 [in this issue].
[2] Ninio A. Language and the learning curve: a new theory of syntactic development. Oxford: Oxford University Press; 2006.
[3] Ninio A. Syntactic development, its input and output. Oxford: Oxford University Press; 2011.
[4] Brown R. A first language: the early stages. Cambridge: Harvard University Press; 1973.
[5] Ingram D. First language acquisition. Cambridge: Cambridge University Press; 1989.
[6] Corominas-Murtra B, Valverde S, Solé RV. The ontogeny of scale-free syntax networks: phase transitions in early language acquisition. Adv Complex Syst 2009;12:371–92.
[7] Nelson K. Structure and strategy in learning to talk. Monogr Soc Res Child Dev 1973;38 [Serial No. 149].
[8] Ke J, Yao Y. Analysing language development from a network approach. J Quant Linguist 2008;15:70–99. http://dx.doi.org/10.1080/09296170701794286.
[9] Solé RV, Corominas Murtra B, Valverde S, Steels L. Language networks: their structure, function and evolution. Santa Fe Institute Working Paper 05-12-042; 2005.
[10] Horgan D. Nouns: love 'em or leave 'em. In: Teller V, White SI, editors. Studies in child language and multilingualism. New York: New York Academy of Sciences; 1980. p. 5–26.
[11] Bloom L, Lightbown P, Hood L. Structure and variation in child language. Monogr Soc Res Child Dev 1975;40(2) [Serial No. 160].
[12] Liu H. Dependency distance as a metric of language comprehension difficulty. Cogn Sci 2008;9(2):159–91.
[13] Ferrer-i-Cancho R. The Euclidean distance between syntactically linked words. Phys Rev E 2004;70:056135.