Title
Non-Adjacent Transitional Probabilities and the Induction of Grammatical Regularities

Permalink
https://escholarship.org/uc/item/1wn8n7kn

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 27(27)

ISSN
1069-7977

Author
Garzon, Francisco Calvo

Publication Date
2005

Peer reviewed

Non-adjacent Transitional Probabilities and the Induction
of Grammatical Regularities

Francisco Calvo Garzón (fjcalvo@um.es)
Departamento de Filosofía, Campus de Espinardo
Murcia, 30100 SPAIN

Abstract

According to Peña et al. (2002), statistical computations based on nonadjacent transitional probabilities of the sort that are exploited in speech segmentation cannot be used in order to induce existing grammatical regularities in the speech stream. In their view, statistics are insufficient to support the discovery of the underlying grammatical regularities. In this note I argue that a single statistical mechanism can account for the data Peña et al. report.

Keywords: statistical computations; rule-based learning; speech segmentation; neural networks.

Introduction

Peña et al. (2002) consider whether statistical computations based on nonadjacent transitional probabilities of the sort that are exploited in speech segmentation (Saffran et al., 1996) can be used in order to induce existing grammatical regularities in the speech stream. To answer this question, they designed an experimental paradigm and performed five different experiments. The first three are briefly reviewed here. Under the light of these experiments Peña et al. conclude that statistics are insufficient to support the discovery of the underlying grammatical regularities, and that their results imply knowledge of rules. *Contra* Peña et al., I shall argue that a *single* statistical mechanism can account for the data they report.

Peña et al.'s experimental paradigm

Peña et al.'s experimental paradigm involves asking adult subjects to listen to a sequence of trisyllabic artificial words for 10 minutes. Words have the form A_iXC_i and are identified on the basis of their nonadjacent transitional probabilities, such that the transitional probability between any A_i and the following C_i is 1.0; between A_i and the intermediate X, and between X and the final C_i is 0.33; and between C_i and the next word's first syllable is 0.5. After familiarization, subjects are confronted with two linguistic stimuli and asked to offer a judgement as to which stimulus is more similar to chunks of the familiarization stream.

Experiment A

In experiment A, subjects were familiarized to a continuous speech stream as described above. In the test phase, they were exposed to a word (A_iXC_i) and a part-word (XC_iA_j)[1], both of them contained in the stream being heard during familiarization. The results reported show that subjects favour words over part-words (P < 0.0005; see Peña et al., 2002, for details). This behaviour backs up the hypothesis that statistics alone suffice to segment a linguistic stream, since what counts as a word is defined as a function of the higher transitional probabilities between specific non-adjacent items (in this case, between A_i and C_i).

Experiment B

In experiment B, Peña et al. wanted to know whether subjects are simply segmenting the stream by exploiting differential transitional probabilities between words and part-words, or whether they are also tuning to some more abstract underlying grammatical regularity. In order to answer this question, after having been familiarized to the same speech stream of experiment A, subjects were asked to choose between a part-word and what Peña et al. dubbed a rule-word. A rule-word is a sequence of three syllables of the form $A_iX^*C_i$, where X^* stands for a familiar syllable that occurs in familiarization, although never between A_i and C_i. In this way, although a rule-word represents a novelty, it is congruent with the structure of actual words by means of the preservation of a non-adjacent transitional probability of 1.0 between A_i and C_i. This time, interestingly, subjects prefer part-words over rule words.

Under the light of experiments A and B, Peña et al. conclude that a "computational mechanism sufficiently powerful to support segmentation on the basis of nonadjacent transitional probabilities [experiment A] is insufficient to support the discovery of the underlying grammatical-like regularity embedded in a continuous speech stream [experiment B]" (p. 605).

Experiment C

Peña et al. then introduced a 25-ms subliminal segmentation gap between words during familiarization. This time subjects manifested preference for rule-words over part-words, identifying them with above-chance accuracy (P < 0.0005; see Peña et al., 2002, for details).

In their view, these results imply knowledge of rules insofar as the very notion of an abstract rule-word underlies the successful discrimination of rule-words and part-words. Thus, they claim: "This seems to be due to the fact that the selected items are compatible with a generalization of the kind "If there is a [pu] now, then there will be a [ki] after an intervening X" (p. 606).

---

[1] Part-words can also be of the form C_kA_iX (see below).

Summing up, Peña et al. contend that two different computational mechanisms must be responsible for the results of experiments A-C: Namely, a statistical mechanism for performing speech segmentation (experiment A), and a rule-governed mechanism responsible for the induction of grammatical structural regularities in the corpus (experiment C).

Subliminal segmentation gaps

In a footnote, however, although they consider a potential rejoinder according to which a single statistical mechanism may be responsible for the induction of the structural regularity in experiment C, they dismiss that alternative. As I shall try to show in what follows there's a misunderstanding in Peña et al.'s line of reasoning that invalidates the conclusion they reach. Let's first of all rescue their whole footnote into the main text for clarity's sake:

"Participants might have included the gaps as separate elements for computing transitional probabilities. As a result, they may have preferred rule words, not because they extracted the structure of the stream, but because they computed probabilities over syllables, pauses, and absence of pauses in the stream and the test items. Thus participants may have analyzed the rule words in the test as having the structure #A1X*C1# and the part words as having the structure #XC2@A3# (where # indicates a pause and @ the absence of a pause). In this case, the transitional probabilities between adjacent elements would favor rule words over part words and no sensitivity to the structure of the rule words would be needed to prefer rule words. This hypothesis makes a prediction that has not been confirmed in a control experiment. Though in experiment [C] the test items do not contain pauses, in this control experiment we tested participants (n = 14) with items including the pauses. Thus, participants compared rule words with structure #A1X*C1# to part words with structure #XC2#A3#. In this case, the presence of the pause in the part words makes the transitional probability of the part word higher than that of the rule word. Therefore, if pauses counted as separate events in the computation, participants should favor the part words over the rule words. Nevertheless, contrary to this prediction, participants still preferred rule words to part words" (fn. 27, p. 607).

In effect, there is no principled reason to exclude the possibility that subliminal segmentation gaps can be exploited statistically. Notice that the very fact that these gaps are subliminal doesn't prevent them from carrying potentially relevant information. It simply means that their presence is not available to conscious access. As a matter of fact, as Peña et al. well observe, they must carry *the* critical piece of information for the mastery of structural induction since the inclusion of the gaps is the one and only difference between experiment B, where the part-word is

preferred, and experiment C, where the rule-word is favoured.

Thus, Peña et al. consider the prediction that subjects would choose rule-words (#A1X*C1#) over part-words (#XC2@A3#), once we consider "probabilities over syllables, pauses, and absence of pauses in the stream and the test items", since "[t]ransitional probabilities *between adjacent elements* favours rule words over part words" (emphasis added).

Non-adjacent transitional probabilities

The problem with this comment is that no reason is offered as to why the only transitional probabilities to be computed must be those between adjacent elements in the speech stream. They arbitrarily assume that a statistical learning mechanism can only be sensitive to immediately adjacent patterns. However, there's no reason not to believe that such mechanisms can be sensitive to higher order (i.e., non-immediately adjacent) regularities. This is so, especially since Peña et al.'s experimental setting was precisely designed by constructing a lexicon mainly characterized in terms of *nonadjacent transitional* probabilities; probabilities which, as they themselves acknowledge, are the cornerstone of the segmentation task in experiment A: "[We] explore whether participants can segment a stream of speech by means of *nonadjacent* transition probabilities, and we also ask whether the same computations are used to promote the discovery of its underlying grammatical structure" (pp. 604-605; emphasis added).

One first hurdle before we can decide whether the experimental results of Peña et al. point towards a non-statistical route, is to specify exactly what it is that we are comparing in terms of probability. Since, they simply comment that adjacent transitional probabilities for rule-words are higher than for part-words, but do not especify which generalizations support that choice, it is not clear how neat the prediction of their working hypothesis is.

We may then ask which test items subjects "should" choose once both adjacent as well as nonadjacent transitional probabilities are considered. Table 1 shows (adjacent and nonadjacent) transitional probabilities between syllables in the familiarization stream, and the corresponding probability values. By taking into account these probabilities, we can see if Peña et al.'s comments in footnote 27 are justified.

Let us first take the rule-word #PUBEKI# and the part-word #RAKI@BE#. According to Peña et al., a choice based on the computation of adjacent transitional probabilities should favour the rule-word. Their prediction is correct. Notice that whereas the rule-word #PUBEKI# is backed up by predictions 1 and 4 in table 1 (summed transitional probabilities = 1.5), the part-word #RAKI@BE# is only backed up by prediction 3 (which has a transitional probability of 0.33).

Table 1: Some adjacent and nonadjacent transitional probabilities between syllables/pauses extracted from Peña et al.'s (2002) familiarization stream, and their corresponding probability values.

Familiarization stream: .....#PURAKI#BELIGA#TAFODU#PUFOKI#TALIDU#BERAGA#TARADU#.....	
Predictions between adjacent items	**Transitional probability**
1. # predicts PU	0.5
2. PU predicts RA	0.33
3. RA predicts KI	0.33
4. KI predicts #	1.0
5. # predicts BE	0.5
6. BE predicts LI	0.33
Predictions between non adjacent items	**Transitional probability**
7. # _ predicts RA	0.33
8. # _ _ predicts KI	0.33
9. # _ _ _ predicts #	1.0
10. # _ _ _ _ predicts BE	0.5
11. PU _ predicts KI	1.0
12. PU _ _ predicts #	1.0
13. PU _ _ _ predicts BE	0.5
14. RA _ predicts #	1.0
15. RA _ _ predicts BE	0.5
16. KI _ predicts BE	0.5
17. # _ predicts LI	0.33
18. KI _ _ predicts LI	0.33

What happens then in Peña et al.'s control experiment when test items include segmentation gaps? They contend that the transitional probability of the part-word (#XC2#A3#) is higher than that of the rule-word (#A1X*C1#). Table 2 shows that their claim is correct *only if* adjacent transitional probabilities are computed

exclusively. Once nonadjacent transitional probabilities are taken into account, the transitional probability of the rule-word becomes higher than that of the part-word. This means that participants in the control experiment may be computing statistical information about segmentation gaps. The prediction would be that they should favour rule-words over part-words, which is exactly what happens in Peña et al.'s control experiment. Therefore, statistical computations can, not only perform speech segmentation, but also promote the discovery of the underlying structural regularities in the corpus.

Table 2: Adjacent and nonadjacent transitional probabilities for part-word #RAKI#BE# and for rule-word #PUBEKI#

Experiment C	**Predictions between adjacent items supporting test preferences**	
		Summed probabilities
Rule-word #PUBEKI#	1, 4	1.5
Part-word #RAKI#BE#	3, 4, 5	1.83
	Predictions between non-adjacent items supporting test preferences	
Rule-word #PUBEKI#	8, 9, 11, 12	3.33
Part-word #RAKI#BE#	14, 15, 16	2
	TOTAL	**adjacent + nonadjacent probability based predictions**
Rule-word #PUBEKI#		4.83
Part-word #RAKI#BE#		3.83

On the other hand, part-words may be of two different types. They can be constructed by taking the last two syllables of a word and the first one of the next word (XCiAj), or by joining the last syllable of a word and the first two syllables of the next word (CkAiX). Peña et al. only consider part-words of the first sort. But the aforementioned results can be consistently extended to the

second sort of part-words. As table 3 shows, were we to insert segmentation gaps in part-words of the second sort (#KI#BELI#), predictions would still favour rule-words over them in statistical terms.

Table 3: Adjacent and nonadjacent transitional probabilities for part-word (type 2) #KI#BELI#.

Part-word (type 2)	Predictions between adjacent items supporting test preferences	
		Summed probabilities
#KI#BELI#	4, 5, 6	1.83
	Predictions between non-adjacent items supporting test preferences	
#KI#BELI#	16, 17, 18	1.16
	TOTAL	adjacent + nonadjacent probability based predictions
#KI#BELI#		2.99

Grammatical induction in SRNs

In order to back up empirically these results, I run a series of connectionist simulations that illustrate the exploitation of statistically-driven information. Following Elman (1990), I trained a simple recurrent network (SRN) on a prediction task to test if it could generalize to novel rule-words in the line of Peña et al.'s experiments A, B and C, and their control experiment.

Stimuli

Familiarization corpus
The familiarization corpus consisted of the same strings of syllables used by Peña et al. (table 1). The corpus thus consisted of CV syllables formed by concatenating all legal combinations of consonants and vowels. CV syllables were encoded in a localist way.

Network architecture and Task

The network had 10 input and output units, and 3 units in

both the hidden and context layers. In the familiarization phase the network was fed with 5,000 syllable tokens. The task was to predict the next item in the sequence.[2]
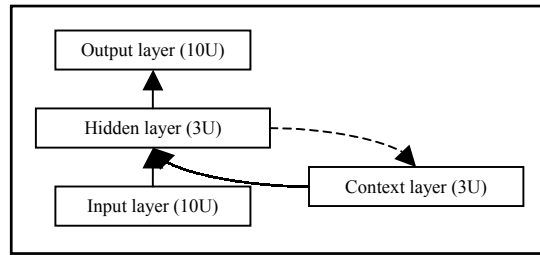

Figure 1: Architecture of SRN used to illustrate the exploitation of statistically-driven information. (the dashed line represents a copy connection).

Results

To confirm the robustness of the results, 5 extended test corpora were created to investigate the predictions of Peña et al: (i) words; (ii) part-words of type 1 (XCiAj); (iii) part-words of type 2 (CkAiX); (iv) rule-words; and (v) part-words that include segmentation gaps in between of the sort considered in footnote 27 of Peña et al., (2002). With the weights from the familiarization phase frozen, networks were tested on these five corpora. Figures 2-6 show differences in prediction root-mean-square (rms) error on test items for all five corpora.[3]
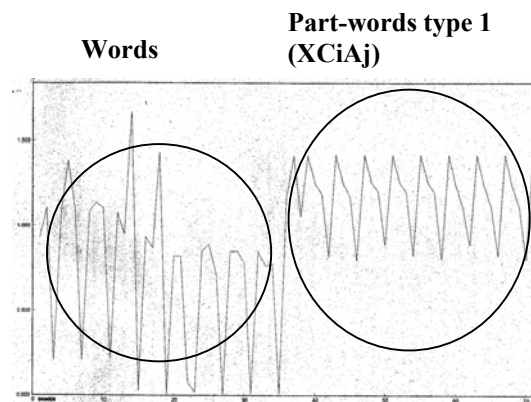

Figure 2: Network performance (RMS error) on words *versus* "type 1" part-words (XCiAj).

---

[2] SRNs were trained with a learning rate of 0,1 during habituation.
[3] Although calculating error measures of probability-based predictions against likelihood vectors would have been more informative, for current purposes sum_rms values suffice.
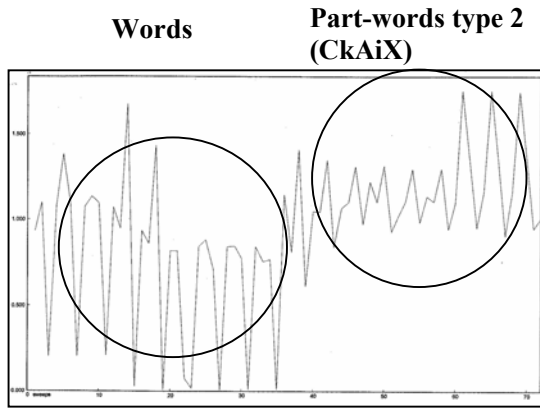
Words   **Part-words type 2 (CkAiX)**

Figure 3: Network performance (RMS error) on words *versus* "type 2" part-words (CkAiX).
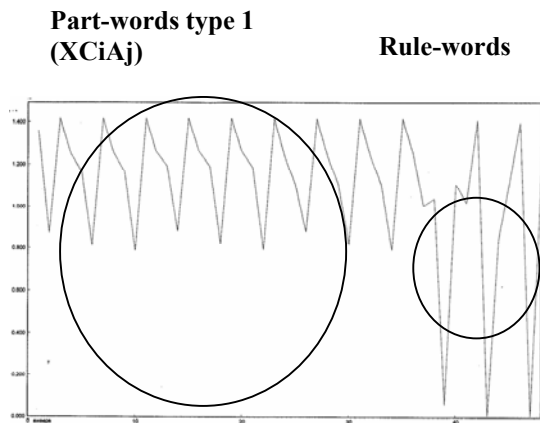
Part-words type 1 (XCiAj)   **Rule-words**

Figure 4: Network performance (RMS error) on "type 1" part-words (XCiAj) *versus* rule-words with structure #A1X*C1#.

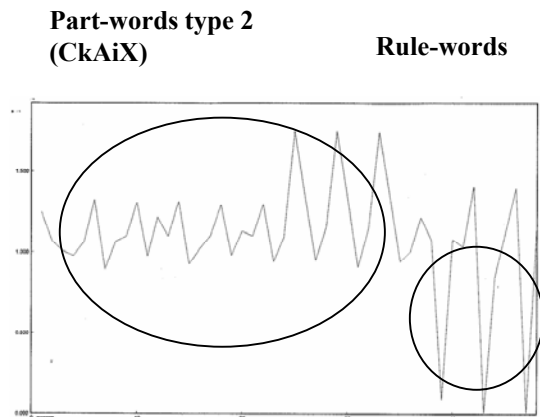**Part-words type 2 (CkAiX)**   **Rule-words**

Figure 5: Network performance (RMS error) on "type 2" part-words (CkAiX) *versus* rule-words with structure #A1X*C1#.

If the network has abstracted the structural regularities that underlie the familiarization corpus, prediction errors in

congruent patterns should be smaller. This prediction is confirmed (figs. 2-6). The results of this simulation show that simple recurrent networks can generalize the abstract patterns embodied in their training set and gain an advantage in processing subsequent patterns of the same grammatical type (i.e., rule-words).

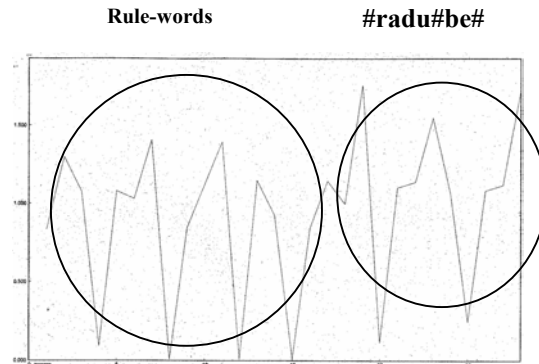**Rule-words**   **#radu#be#**

Figure 6: Network performance (RMS error) on rule-words with structure #A1X*C1# and part-words with structure #XC2#A3# (see fn. 27 from Peña et al., 2002, above).

Conclusion

The results reported here show that frequency and distributional properties in the corpus not only serve to segment statistically the data into its constitutive legal words, but also to explain the choices made by subjects that apparently involve manipulation of non-statistical information.

In this way, statistical computations based on nonadjacent transitional probabilities of the sort that are exploited in speech segmentation (Saffran et al., 1996) may be used in order to induce existing grammatical regularities in the speech stream.[4]

The arguments offered here don't attempt to show that this indeed is the case, but rather to illustrate that this is an option that cannot be discarded beforehand.

Acknowledgements

[4] There is in fact a fairly clear literature demonstrating that recurrent networks can induce grammars from examples of context-free and context-sensitive languages; grammars that are precisely of a form in which there are long-distance dependencies. See for example Boden & Wiles (2000), and Chalup & Blair (2003). Many thanks to Jeff Elman for bringing this to my attention.

References

Boden, M., & Wiles, J. (2000). Context-free and context-sensitive dynamics in recurrent neural networks. *Connection Science*, *12*, 197-210.

Chalup, S. K., & Blair, A. D. (2003). Incremental training of first order recurrent neural networks to predict a context-sensitive language", *Neural Networks*, *16*, 955-972.

Elman, J. (1990). Finding Structure in Time. *Cognitive Science*, *14*, 179-211.

Peña, M., Bonatti, L., Nespor, M., & Mehler, J. (2002). Signal-Driven computations in Speech Processing. *Science*, *298*, 604-607.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926-1928.