Mind the gap between hierarchy and time

Cristiano Chesi
IUSS Scuola Universitaria Superiore Pavia
cristiano.chesi@iusspavia.it

Andrea Moro
IUSS Scuola Universitaria Superiore Pavia
andrea.moro@iusspavia.it

1. Introduction to hierarchy and “time” in linguistics

Consider time as a total order among discrete events.

If we look at human languages, a total order is always established among discrete events, which are the distinct pronunciations of the words (and morphemes they are formed by) in a sentence. Notice that this is not a necessary condition: if we think of the total order as a restriction imposed by the spoken modality, we could imagine a sign language in which, for instance, two arguments (e.g. “Eva” and “the apple”) of a certain predicate (e.g. “eats”) could be co-articulated exactly at the same time in space, one with one hand and another with the other hand in a temporally and spatially symmetrical situation. To our knowledge this possibility is limited by several restrictions in any sign language around the world. [1]

[1] In American Sign Language, for instance, (Napoli & Sutton-Spence 2010), the arguments of a sentence like "A man looking for the meowing cat in Paul Scott's ‘Tree’" are attested to be signed simultaneously (the man is represented by a raised finger of the right hand, while the ground where the tree is located is signed with the flat left hand) but eye gaze direction (“man

Here we assume that “time” is an essential ingredient of any human language and not an articulatory-perceptual restriction (also known as a “Phonetic Form effect”, Fox & Pesetsky 2005), hence it must be included in a theory of language that aims at being cognitively adequate. More precisely, our hypothesis is that “time” induces an asymmetric (total) relation among spelled-out linguistic units and this is a necessary, though not sufficient, condition for the correct structural analysis of the sentence, hence for its correct interpretation. The other crucial property is hierarchy: implicit groups of words (phrase structures) that constraint the sequence of expectations (Chesi 2015) and create complex (recursive) meaningful units.

Our intent here is to argue, by analyzing simple recurrent networks expectations and a very simple linguistic fact like anaphoric binding, that linear order is necessary but not sufficient to account for the hierarchical restrictions limiting pronominal interpretation.

2. On the interaction between hierarchy and time

2.1 Creating expectations

Elman (1993) succeeded in representing time in linguistic processing using Simple Recurrent Networks (SRN). SRNs are artificial neural networks that use a copy of the hidden layer activation status at time *t* and re-submit such activation to the same hidden units at time *t*+1, summing it up with the activation of the afferent input layer at that time. SRNs of the Elman’s kind are evaluated on its ability to predict next input token, namely next word: in a sentence like “Eva eats the apple”, we expect a network that has learned to perform a correct “prediction” for a third person singular verb after the subject “Eva” and expect a common noun after the article “the”. This approach is rather coherent with our idea of “time” as sequence of discrete events since the sentence is chunked and the network gets fed word by word. Elman shows that SRNs succeed in learning many grammatical constraints, like subject-verb agreement both locally (“*Eva eats*…”) and non-locally (“*Eva*, who Adam knows very well, *eats*…”).[2]

looking”) and mouth movement (“cat meowing”) pose an asymmetrical order among the distinct arguments, unambiguously indicating who does what and where.

[2] We disregard here Elman’s “starting small” idea and its criticism (Rohde & Plaut 2001).

2.2 Expectations on recursive structure

SRNs seem to be able to model even more subtle properties of the human performance in allegedly recursive structures: Christiansen & Charter (1999) show that a recursive network with a decent number of nodes in the hidden layer (more than 10) can easily learn up to three non-local dependencies both of the nested kind (e.g. *a b c c b a*) and cross-serial kind (*a b c a b c*), performing slightly better with the second kind than with the first one. Although this is coherent with the psycholinguistic evidence, this is surprising according to Chomsky's generative power hierarchy for phrase structure grammars that ranks the cross-serial kind of dependencies higher up in the hierarchy: while Context-Free Grammar (CFG) are powerful enough to capture nested dependencies, a more powerful grammar is needed for capturing serial dependencies. This might indicate at least two things: first, the cognitive "complexity" of a grammar could not be straightforwardly predictable from Chomsky's Hierarchy (Chesi & Moro 2014); second, it might be the case that recursion (hence hierarchy) is not tested in these experiments, but just a three-level dependency that simply compares sequences of objects mimicking the effect of the application of a serial vs. nested recursive dependency formation rule. This is a general problem shared by many distributional-based approaches to linguistic performance (Tomasello 2009). One way to solve this uncertainty is to focus on specific linguistic constructions that share similar distributions but that are processed differently (as the nested vs serial dependencies) or linguistic facts that show different distributional patterns, but that do not present any genuine asymmetry in performance, like anaphoric binding.

2.3 Wrong expectations and the role of hierarchy in anaphoric binding

One simple linguistic case suggesting that processing just word sequences results in wrong expectations is binding: a reflexive pronoun (e.g. *herself*) requires a preceding local noun phrase (*Eva*) to be coreferent with it[3] (*Eva*, in the example (1.a)). From (1.a) we might conclude that precedence is the correct property, but this intuition is in contrast with (1.b) where *compagno* and not "Eva" can be coreferent with the second reflexive *si*, though Eva just precedes and is even closer to it. (1.c) confirms that not even immediate precedence is sufficient for picking up the correct binder. This proves that hierar-

[3] Subscripts indicate co-reference: a_i and b_i are coreferent; a_i and b_j are not coreferent. When start (*) prefixes a sentence where the noun phrase and the reflexive are coindexed (i.e. *a_i ... b_i ...) , coreference is impossible.

chy preempts “time” (namely linear order). These constraints are expressed by C(onstituent)-command idea: the first node dominating a noun phrase should dominate the coreferent reflexive (Reinhart 1976)[4].

(1) a. **[**[*Eva*$_i$] [*si*$_{i/*j}$ presentò] **[**e [*il compagno*$_j$] [*si*$_{*i/j}$ offese]**]]**
E. *him/her-self* introduced and the partner *him/her-self* upset (lit.)
“E. introduced him/her-self and the partner got upset”
b. **[**[*il compagno*$_j$] **[**a cui [*Eva*$_i$] [*si*$_{i/*j}$ presentò]**]** [*si*$_{*i/j}$ offese]**]**
“the partner whom E. introduced him/her-self got upset”
c. **[**[*il compagno*$_j$] [di [*Eva*$_i$]] [*si*$_{*i/j}$ presentò] [e [*si*$_{*i/j}$ offese]]**]**
“the partner of E. introduced him/her-self and got upset”

If we would expect a binder to always (immediately) precede the reflexive, we would not be able to interpret correctly the sentence (1.b) and (1.c). We decide to run a little experiment to verify the consistency of a “usage-based” approach in this special case.

3. Distributional frequencies in reflexive binding

If distributional frequencies were sufficient to learn subtle structural phenomena, we would expect an evidence about the fact that coreference in reflexive binding is equally attested in any structural configuration. So we queried Repubblica corpus (380M tokens, Baroni et al. 2004) for the distribution of the sequences “NP *si* intransitive_pronominal_verb”, “NP PP$_{\text{genitive}}$ *si* intransitive_pronominal_verb” and “NP a cui NP *si* intransitive_pronominal_ verb”. Among the 674.057 occurrences found, about 46% of occurrences where of the local binding kind (e.g. “[la camera]$_i$ si$_i$ appresta”) and only 16% of the NP PP kind (e.g. “[[l’articolo]$_i$ [di Ajello]] si$_i$ presenta”). Just a bunch of occurrences were of the “NP a cui NP *si*” type, all the rest conforms to the pattern “PP/NP si” but coreference is not at issue (e.g. “secondo indiscrezioni *si* tratterebbe di…”). Such distributional asymmetries do not correlate with any difficulty/ambiguity perceived for the [NP$_i$ [PP]] *si*$_i$ kind of binding nor for the “*NP*$_i$ a cui NP *si*$_i$” type. Moreover, the majority of cases conform with a distributional pattern that is of the “NP si” kind without involving any coreference between the NP and the reflexive. How this can be learned/explained simply on the basis of linear distribution?

[4] In (1), bold squared brackets indicated the first node dominating the noun phrase (the corresponding closed bold squared bracket indicates the end of the binding domain). Inclusion among brackets indicates hierarchical dominance. Subscripts indicate possible and impossible (*) coreference.

4. Discussion

On the basis of "next-word prediction" SRN experiment and anaphoric binding facts, we argued that the necessary asymmetry (Moro 2000) created by "time" in linguistic processing is necessary for ordering the relevant expectations that drive the interpretation of the sentence (Chesi 2015). But we also stressed the fact that restrictions on binding suggest that hierarchy preempts time and phrase structure can not be predictable on the basis of simple distributional evidence.

Bibliografia

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G., & Mazzoleni, M. (2004) Introducing the La Repubblica corpus. 2:5-163.

Chesi, C. (2015) On directionality of phrase structure building. Journal of psycholinguistic research, 44(1):65-89.

Chesi, C., Moro, A. (2014) Computational complexity in the brain. Measuring grammatical complexity. 264-280.

Chomsky, N. (1995) The minimalist program. Cambridge, MA: MIT press.

Christiansen, Chater (1999) Toward a Connectionist Model of Recursion in Human Linguistic Performance. Cognitive Science 23(2):157-205

Elman, J. (1993) Learning and development in neural networks: the importance of starting small. Cognition 48:71-99

Fox, D., Pesetsky, D. (2005) Cyclic linearization of syntactic structure. Theoretical linguistics, 31(1-2):1-45.

Moro, A. (2000) Dynamic antisymmetry. MIT press.

Napoli, D. J., & Sutton-Spence, R. (2010). Limitations on simultaneity in sign language. Language, 86(3):647-662.

Reinhart, T. (1976) The syntactic domain of anaphora. Ph.D. Thesis MIT.

Rohde, Plaut (2001) Less is less in language acquisition. In Quinlin Connectionist modelling of cognitive development.

Tomasello, M. (2009) Constructing a language: A usage-based theory of language acquisition. Harvard university press.