Finite-state linguistic structure building

Terry Langendoen
Linguistics Program, National Science Foundation
Department of Linguistics, University of Arizona

This material is based in part upon work supported while the author was serving at the National Science Foundation. Any opinion and conclusions are those of the author and do not necessarily reflect the views of the National Science Foundation.

Outline – Part 1

- Parataxis
  - Very small bound on degree of paratactic embedding (PE°)
    - Langendoen 1998
- Right- and left-embedding hypotaxis (R/LE)
  - Readjustment to pseudo-paratactic (ΨP) structure
    - Langendoen 1975, updated with traces
- Center-embedding hypotaxis (CE)
  - Motivating concept of zigzag embedding (ZE)
  - Finite-state modeling of bounded degree of ZE

Polysyndeton (more than one coordinator)

1. He is not quite journalist or carnival barker or orator or interlocutor or master of ceremonies or trained seal.
   - American Publishing House for the Blind (APHB) corpus
2. It's great to see millions on this earth who had nothing but a record of sadness and poverty and misery and hunger and disease have the chance to go up.
   - APHB corpus

Monosyndeton (exactly one coordinator)

1. It provides him with a ... method of getting a hearing ... in any federal income, gift or estate tax dispute ....
   - (APHB corpus)
2. Courts often need ... to settle manslaughter charges, inheritance claims, insurance proceeds, tax problems, and the disposition of jointly held money and property.
   - (APHB corpus)

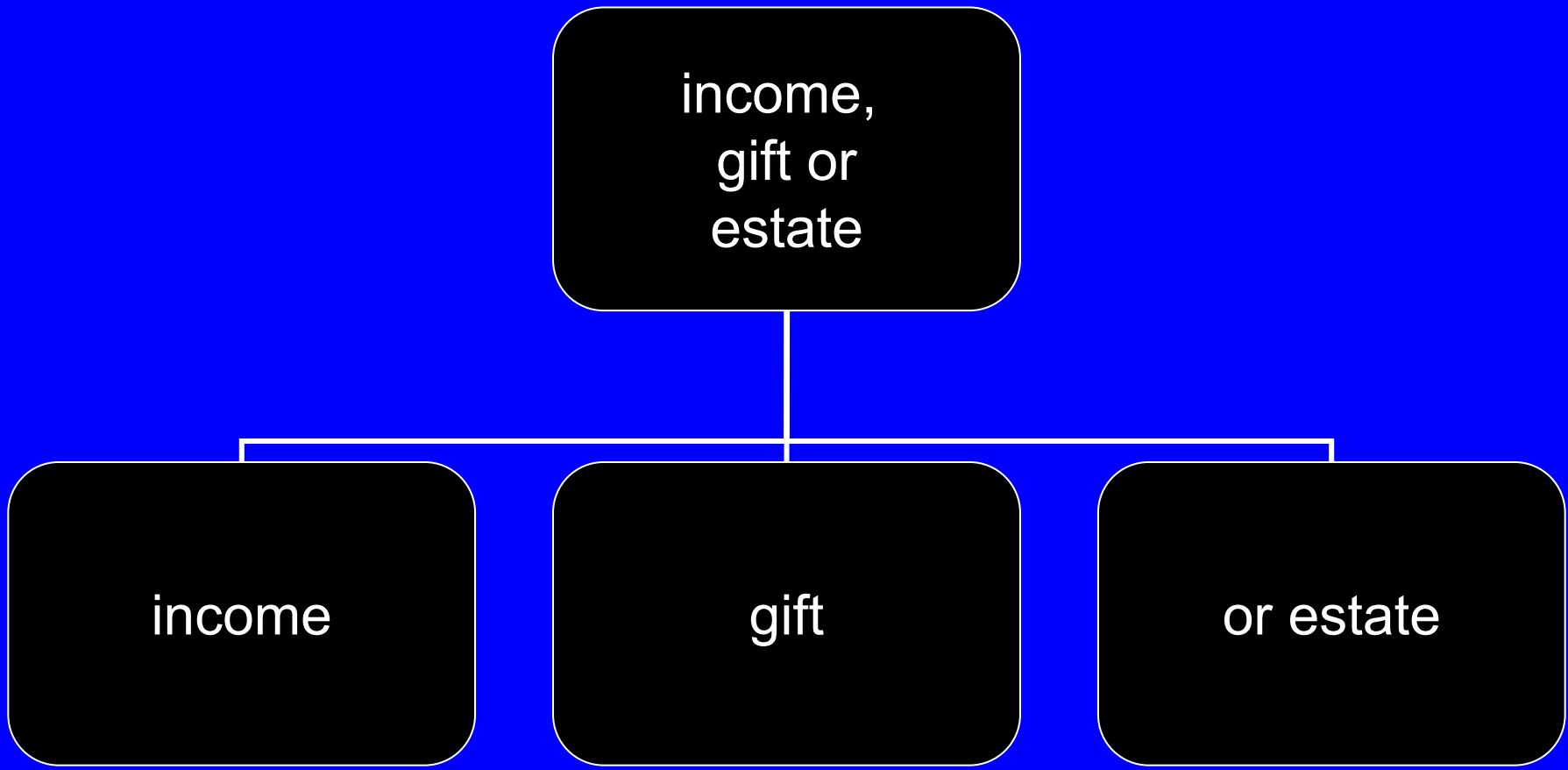
Asyndeton (no coordinator)

1. I felt exposed, unprotected, somehow afraid of what might happen.
   - SUSANNE corpus
2. The thistle, the nettle, the burdock, the belladonna / Have a future....
   - Czeslaw Milosz, “The Thistle, The Nettle,” translated from Polish by Czeslaw Milosz and Robert Hass, *The New Yorker*, April 30, 1990

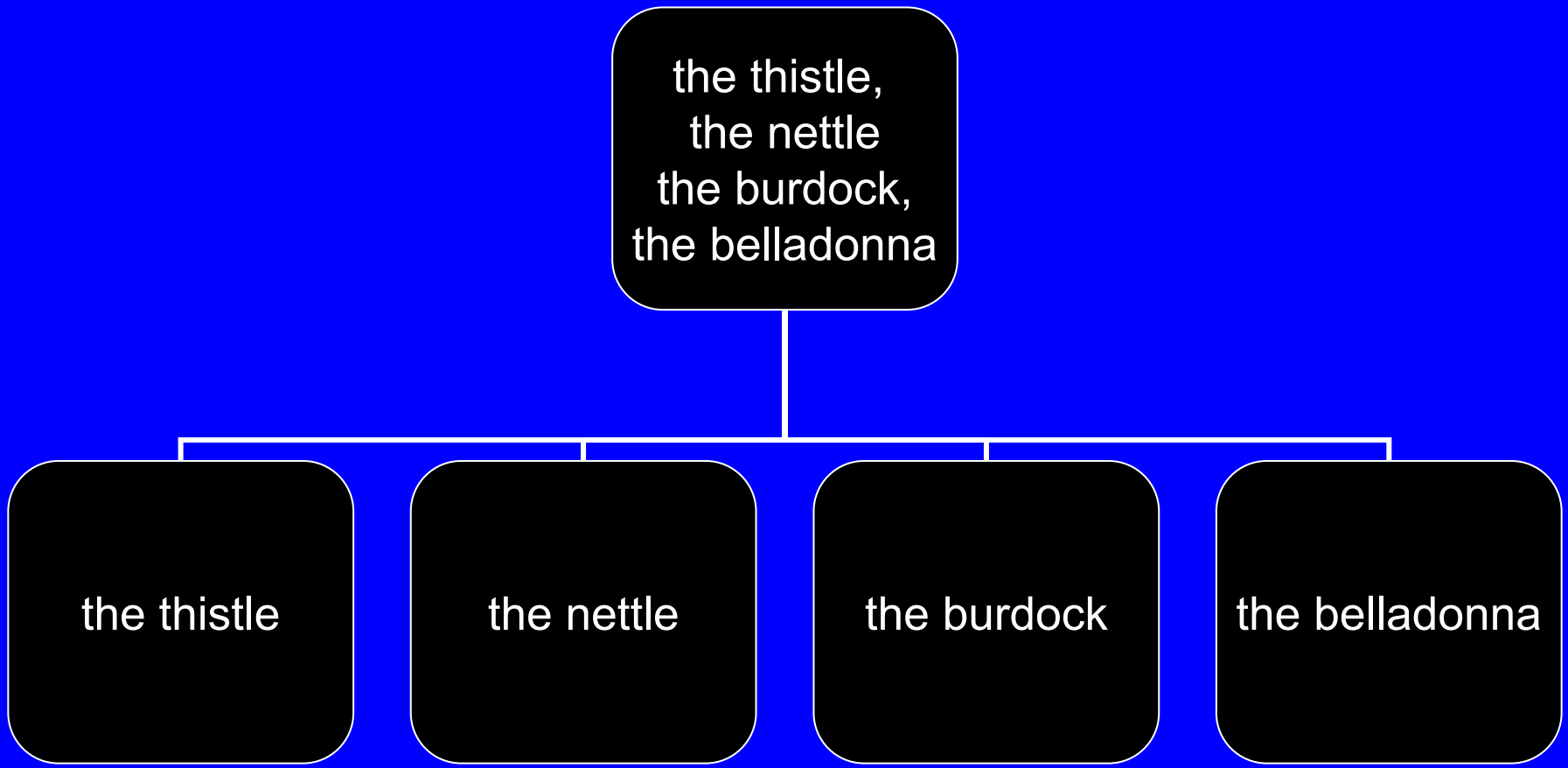
Nesting of paratactic structure -- not discussed further

- ... the purified soybean meal is [$_{cAP0}$ colored, flavored, pressed, shaped and cut into bits that [$_{cVP1}$ [$_{cV2}$ look and taste] like [$_{cNP2}$ bacon [$_{cN3}$ chips or strips], pork sausage, ground beef, sliced ham or chicken] and are [$_{cAP2}$ cheaper and just as nourishing as the real thing]]].
  - APHB corpus

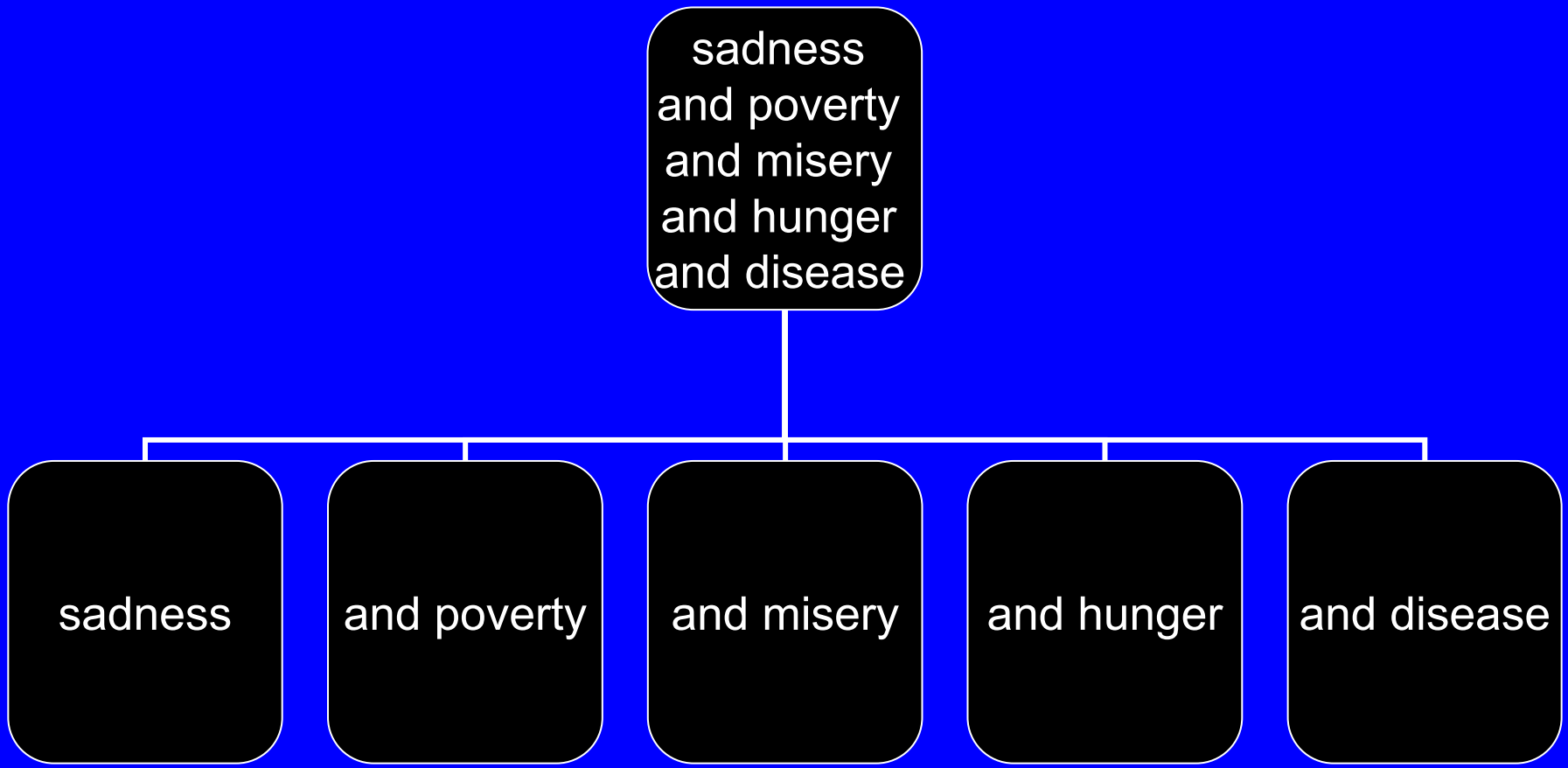
Monosyndetons have ‘flat’ structure



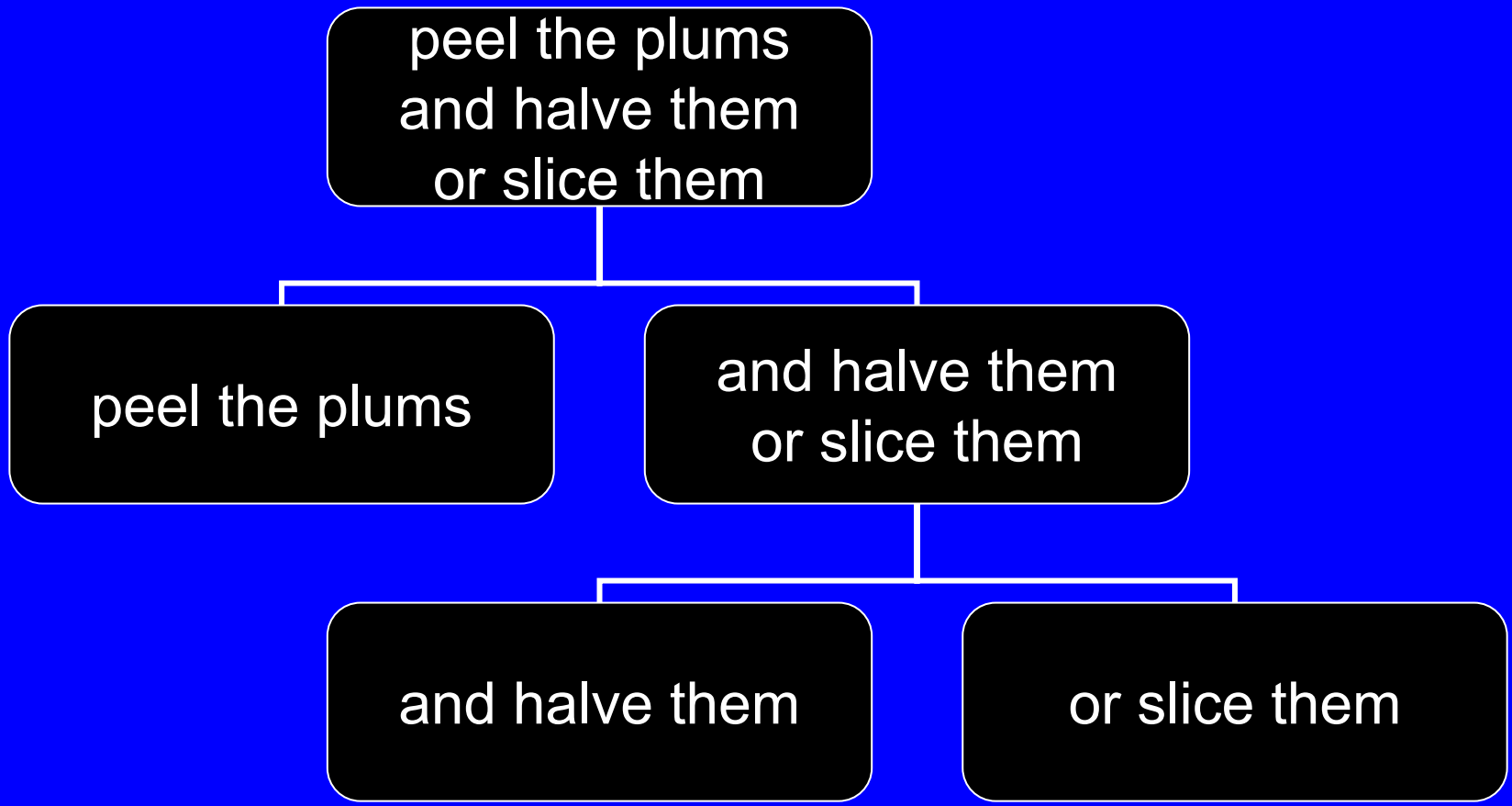
So do asyndetons



And so *can* polysyndetons



But sometimes polysyndetons show paratactic embedding (PE)



Factors resulting in PE

A. Members introduced by different connectives.
B. Members introduced by different junctures -- not considered further here.
C. A member introduced by a connective followed by one that is not.
D. Lexical, pragmatic or semantic considerations.

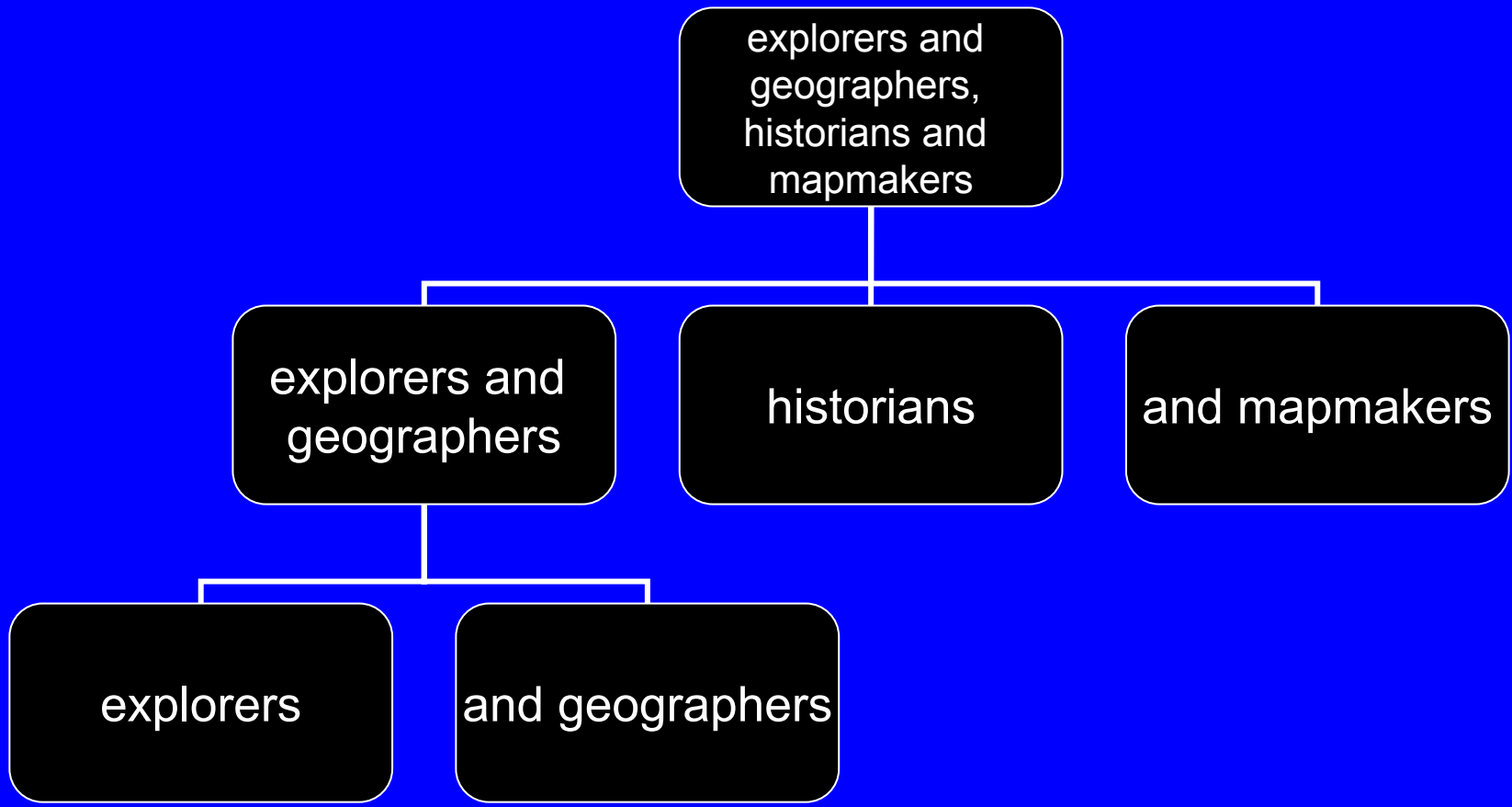
Other examples of Factor A

1. Caroline was going into the possibility of the Pope or his priests and the nuns.
   - APHB corpus
2. The bush babies cluster together and groom each other, or run through the trees in gangs.
   - APHB corpus

Examples of Factor C

1. ... an employer has a right to refuse to hire a man if he doesn't like the color of his tie, or his diction, his shifty eyes, or his having taken the Fifth Amendment.
   - APHB corpus
2. *... The Travels of Marco Polo* became an indispensable book for explorers and geographers, historians and mapmakers and the delight of all who travel or dream of doing so.
   - APHB corpus

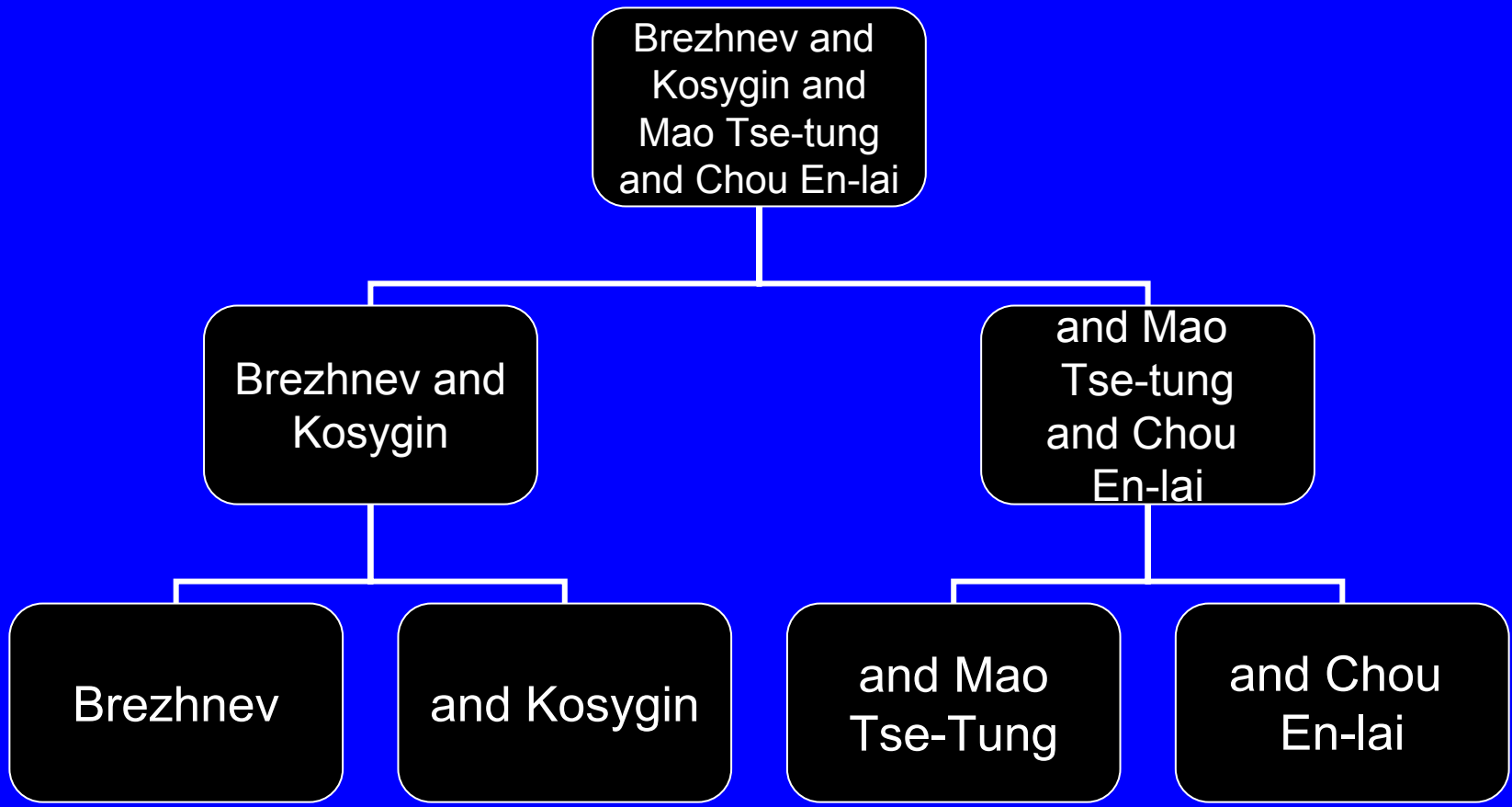
One PE structure resulting from Factor C



Examples of Factor D

1. That's our job and that's the job of Brezhnev and Kosygin and Mao Tse-tung and Chou En-lai.
   - APHB corpus)
2. Then Dr. White and his faculty and students could assemble and throw rocks at each other and play with matches and burn things down.
   - APHB corpus

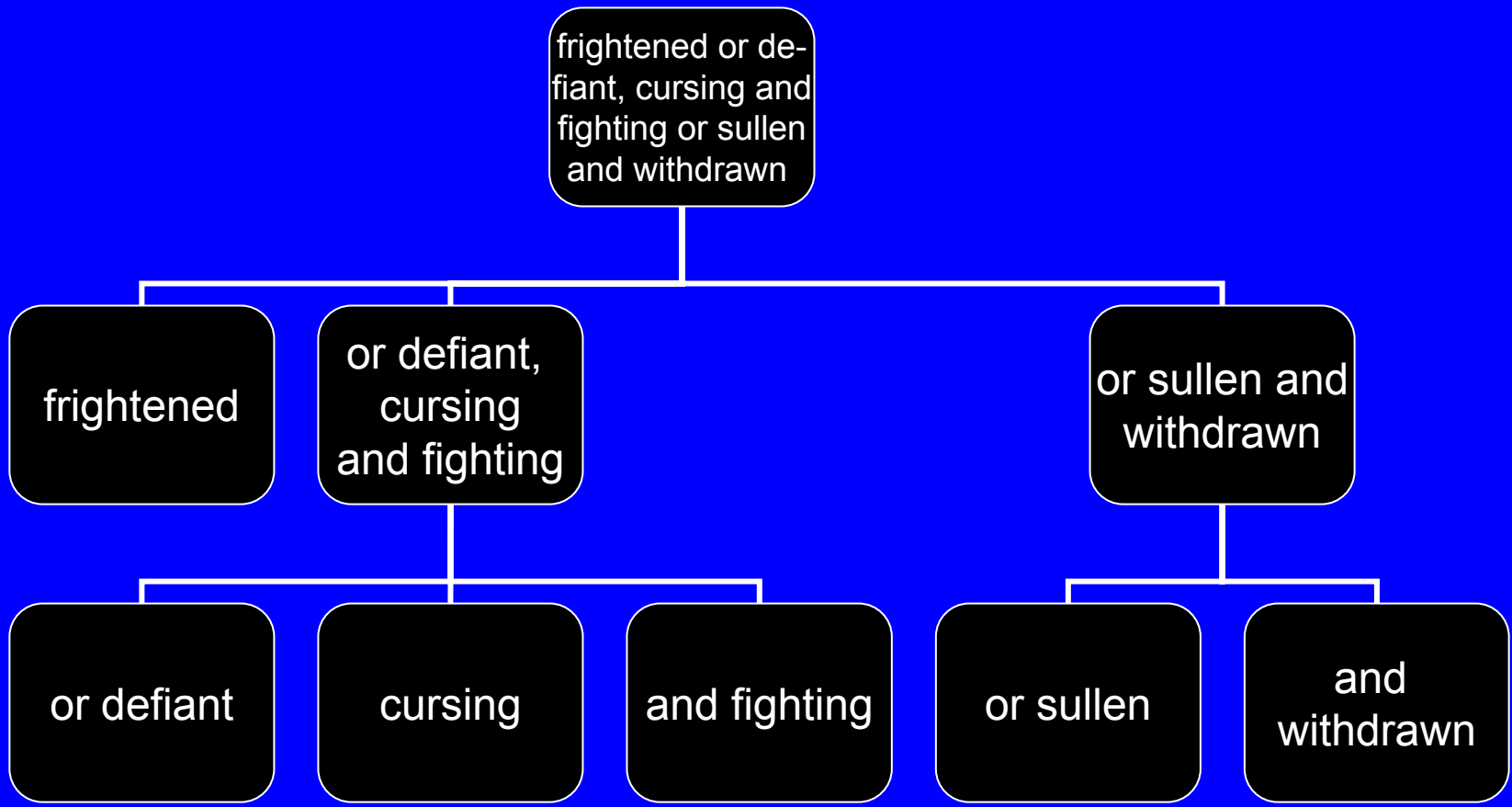
One PE structure resulting from Factor D



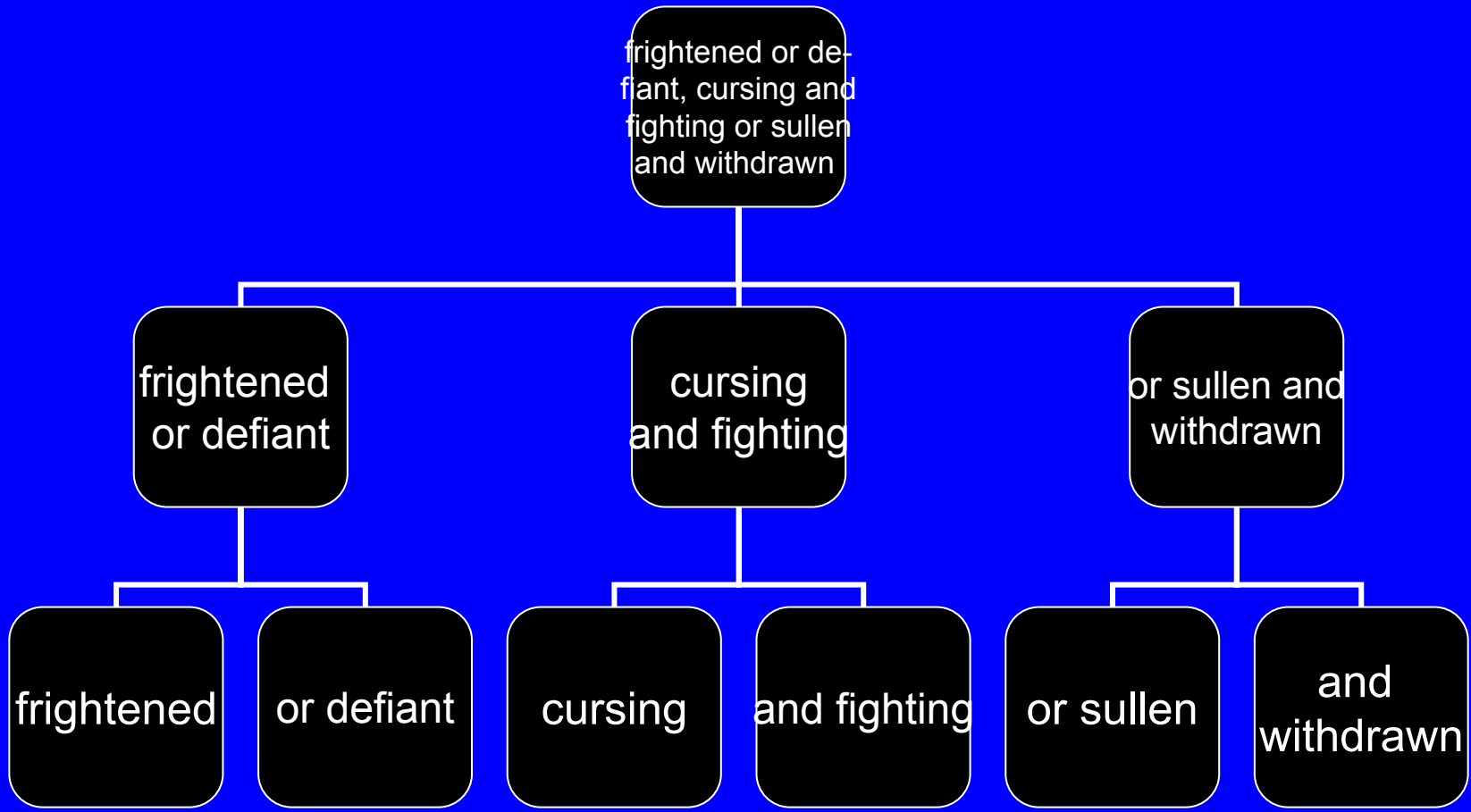
Combination of Factors

1. a company of persons gathered for deliberation and legislation, worship or entertainment
   - *Webster's Collegiate Dictionary* 7th ed., definition of *assembly*
2. The girls are brought in frightened or defiant, cursing and fighting or sullen and withdrawn.
   - APHB corpus
3. Combine grapefruit with bananas, strawberries and bananas, bananas and melon balls, raspberries or strawberries and melon balls, seedless white grapes and melon balls, or pineapple cubes and orange slices.
   - *The James Beard Cookbook*, recipe for grapefruit salad

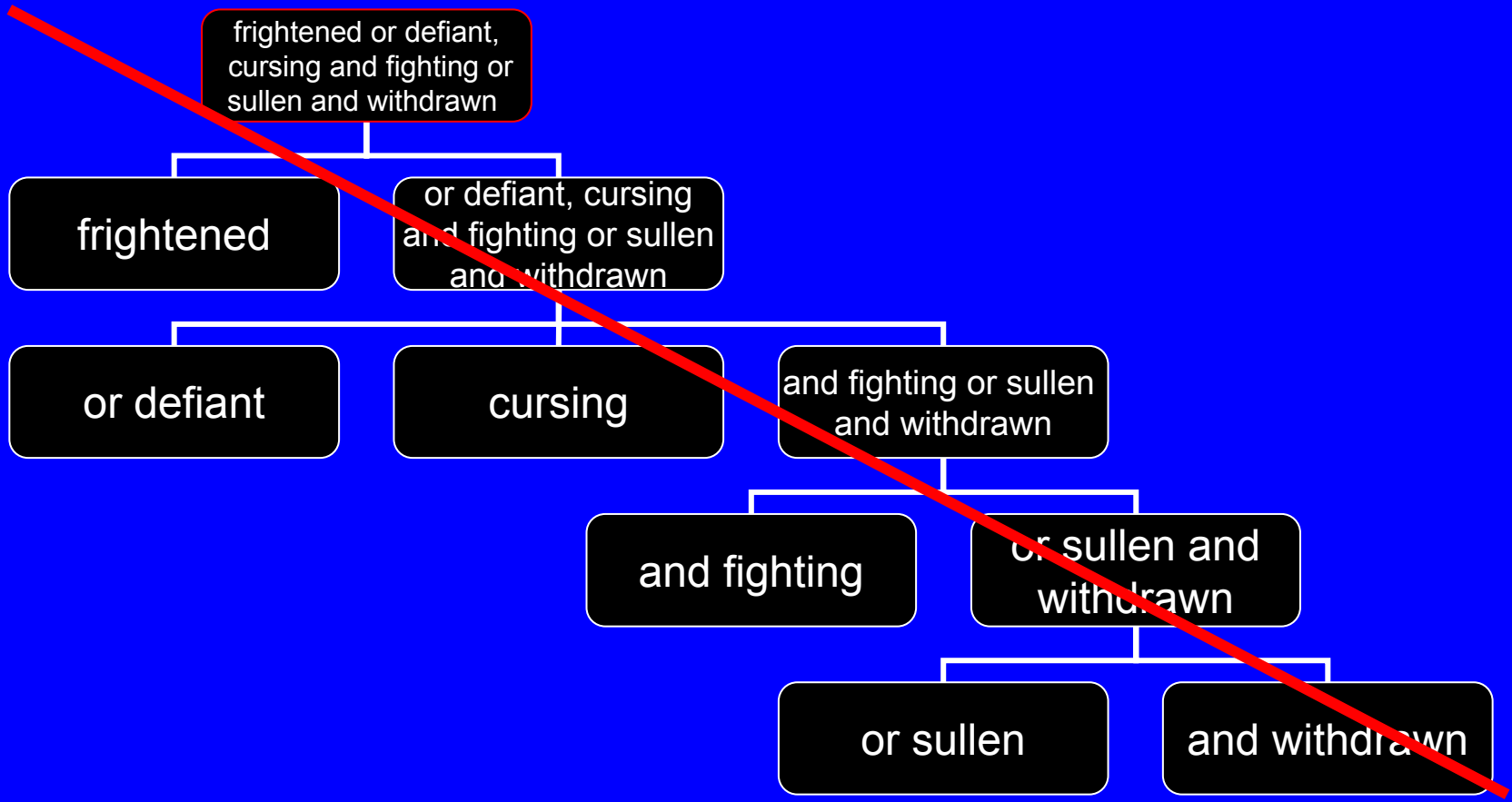
One PE structure resulting from Factors A and C



Another PE structure resulting from Factors A and C



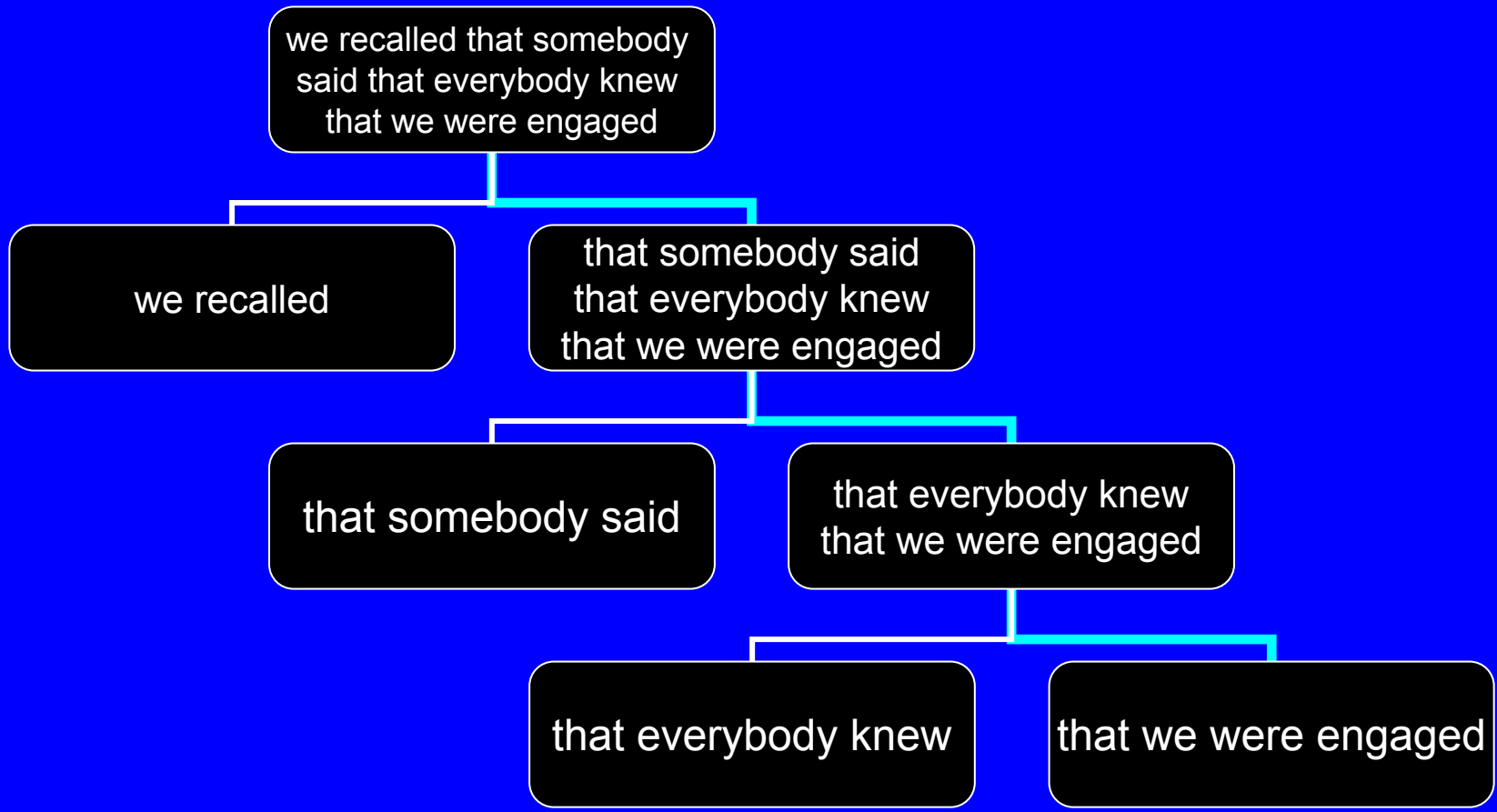
PE° > 2 is hardly ever attested



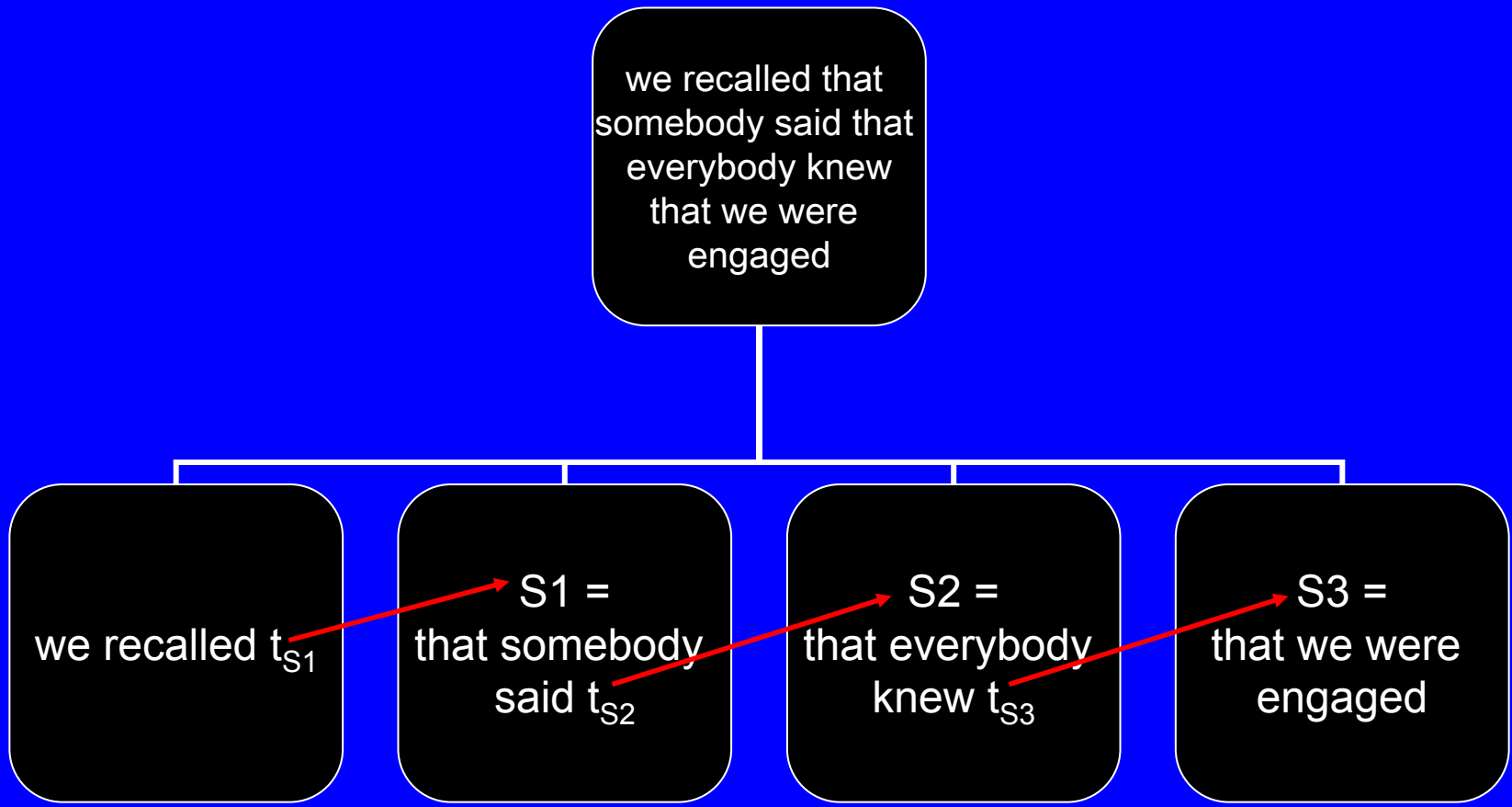
Outline – Part 2

- Parataxis
  - Very small bound on degree of paratactic embedding (PE°)
    - Langendoen 1998
- ➢ Right- and left-embedding hypotaxis (R/LE)
  - Readjustment to pseudo-paratactic (ΨP) structure
    - Langendoen 1975 updated with traces
- Center-embedding hypotaxis (CE)
  - Motivating concept of zigzag embedding (ZE)
  - Finite-state modeling of bounded degree of ZE

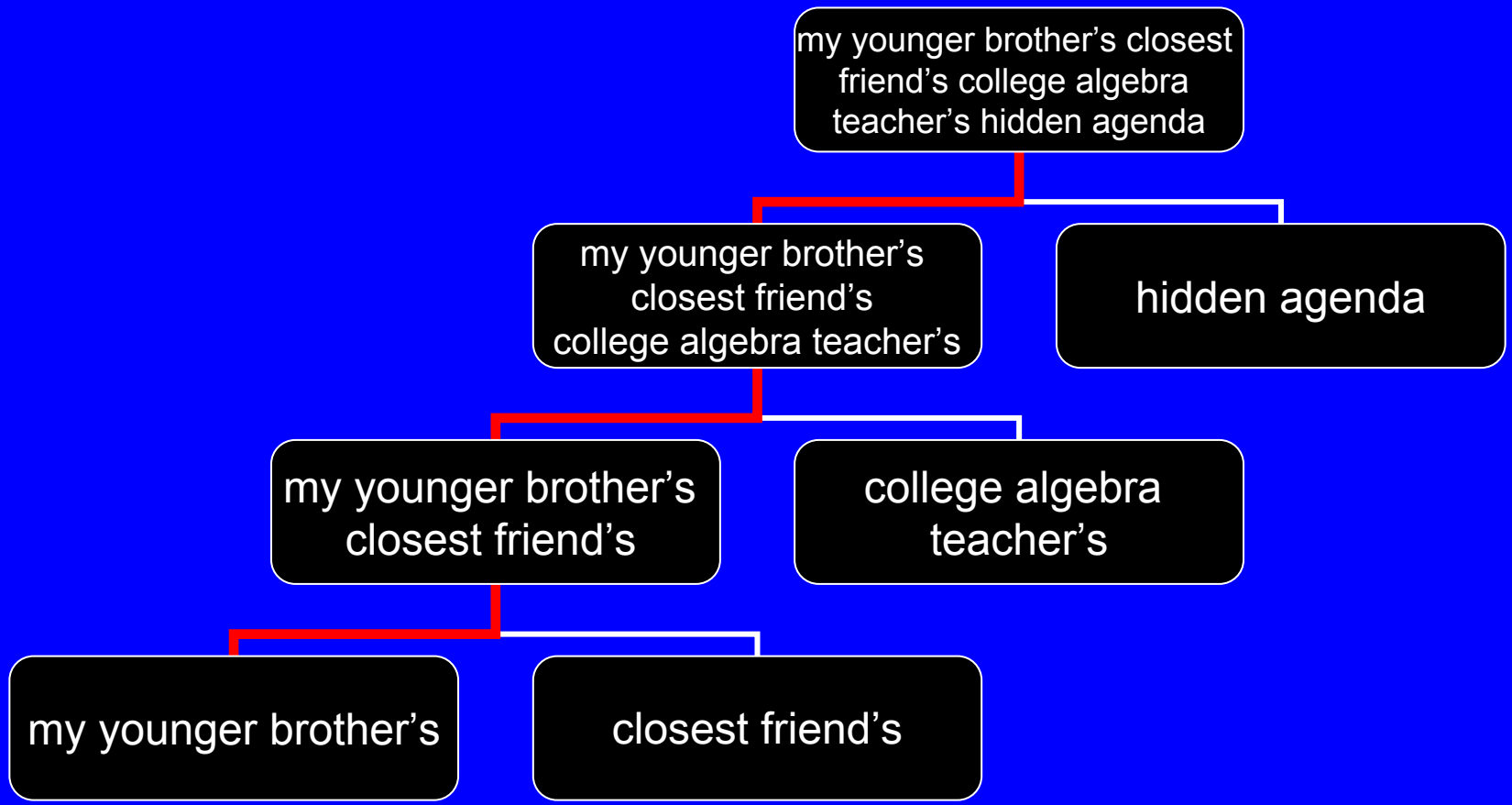
A right-embedding (RE) hypotactic structure



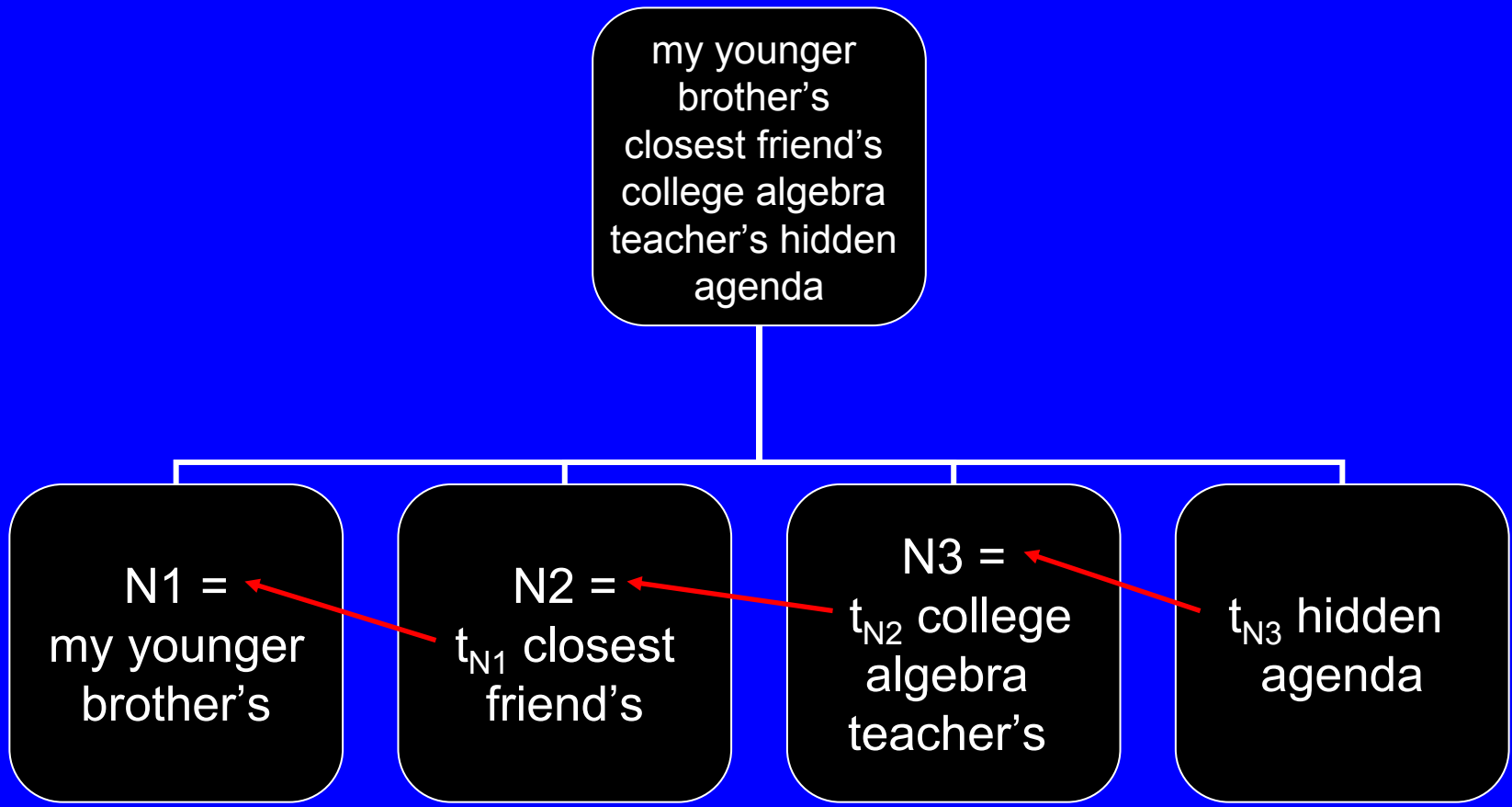
Equivalent pseudo-paratactic (ΨP) structure



A left-embedding (LE) hypotactic structure



Equivalent ΨPS



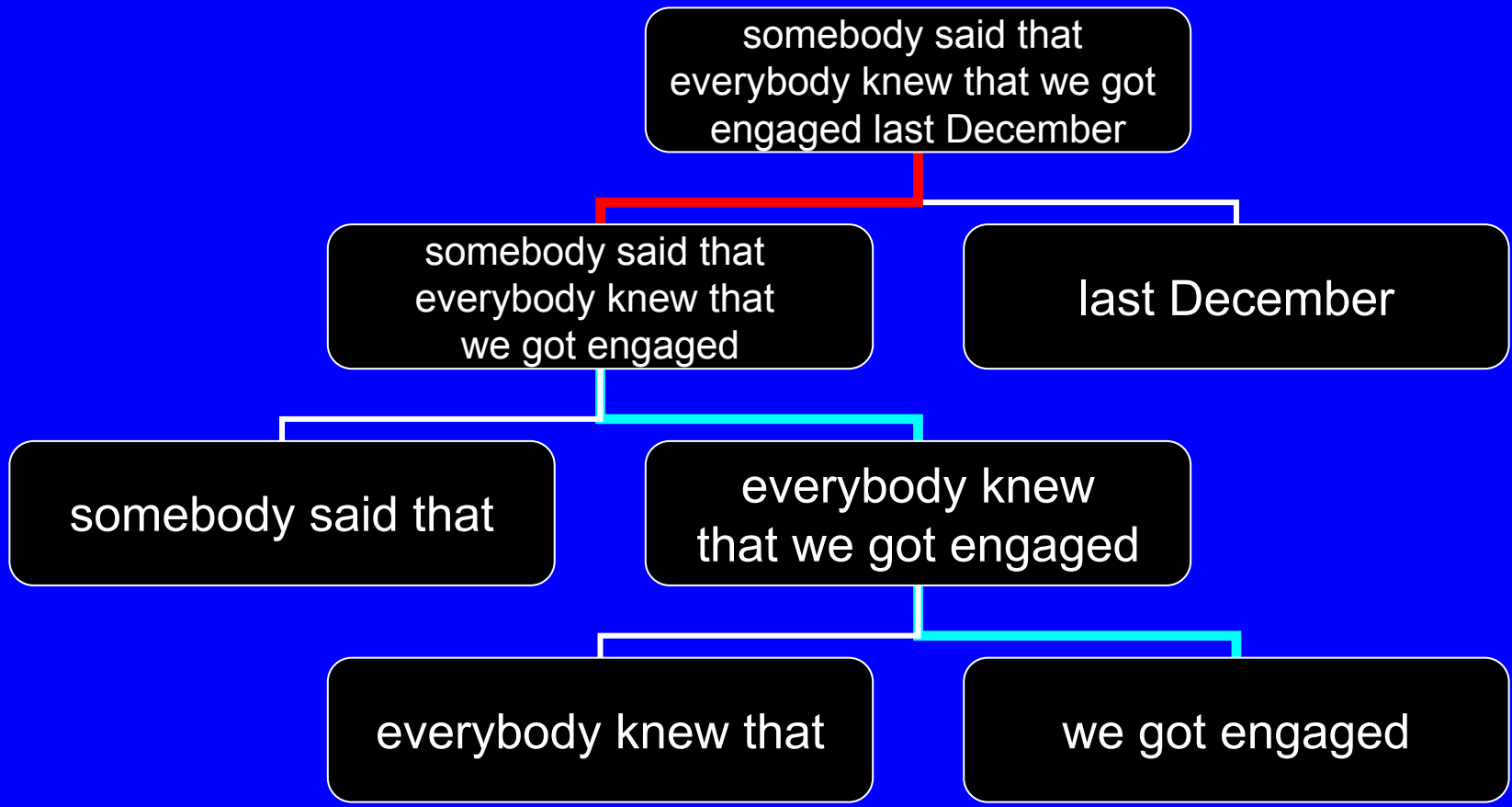
Outline – Part 3

- Parataxis
  - Very small bound on degree of paratactic embedding (PE°)
    - Langendoen 1998
- Right- and left-embedding hypotaxis (R/LE)
  - Readjustment to pseudo-paratactic (ΨP) structure
    - Langendoen 1975 updated with traces
- ➢ Center-embedding hypotaxis (CE)
  - Motivating concept of zigzag embedding (ZE)
  - Finite-state modeling of bounded degree of ZE

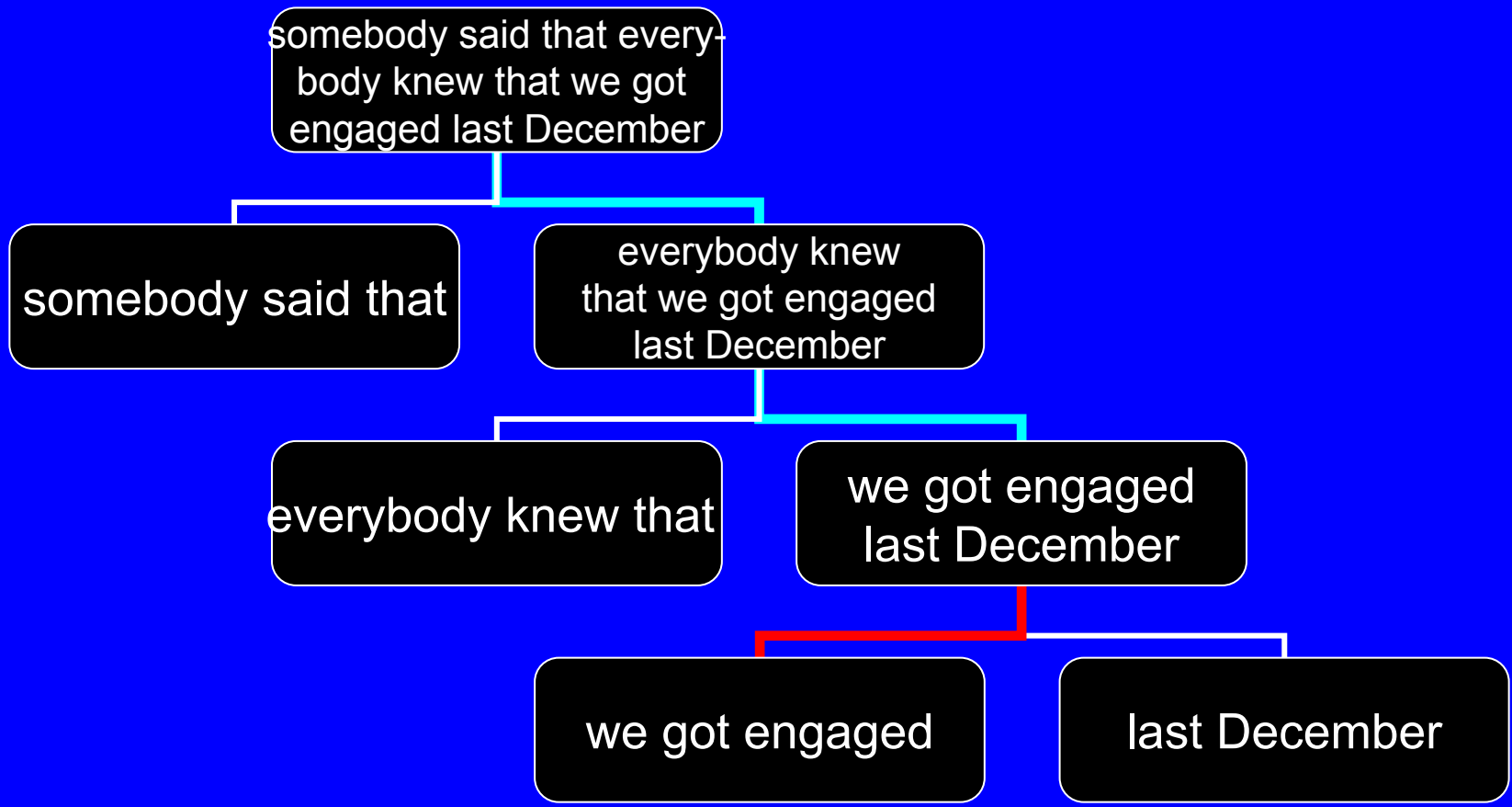
Adverbial attachment to an RE structure results in CE

- The sentence Somebody said that everybody knew that we got engaged last December is structurally ambiguous, depending on which clause last December attaches to:
  - High: somebody said that ...
  - Low: we got engaged
  - Middle: everybody knew that ...
- Each structure exhibits CE° = 1.

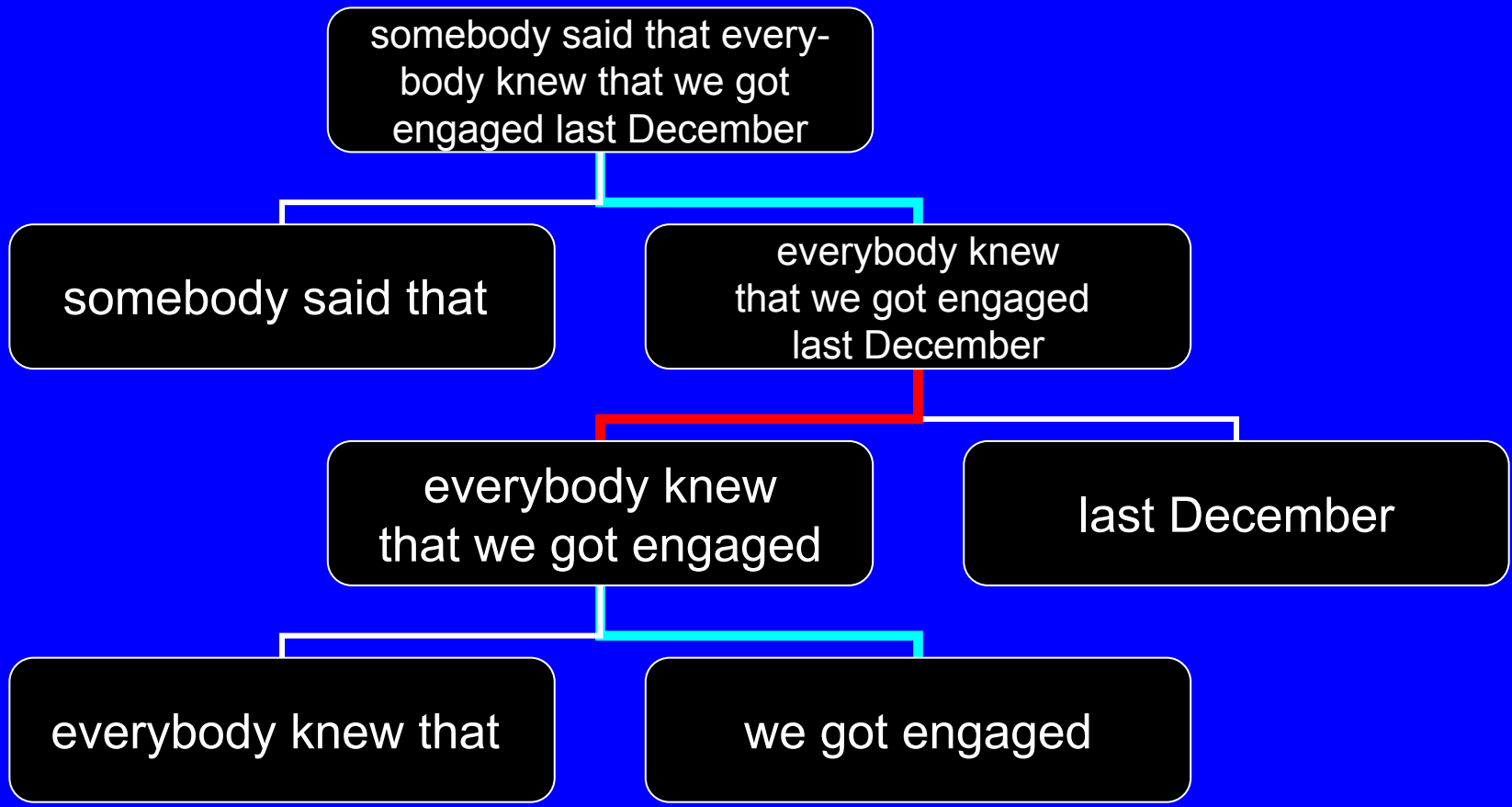
High attachment (L then R)



Low attachment (R then L)



Middle attachment (R then L then R)

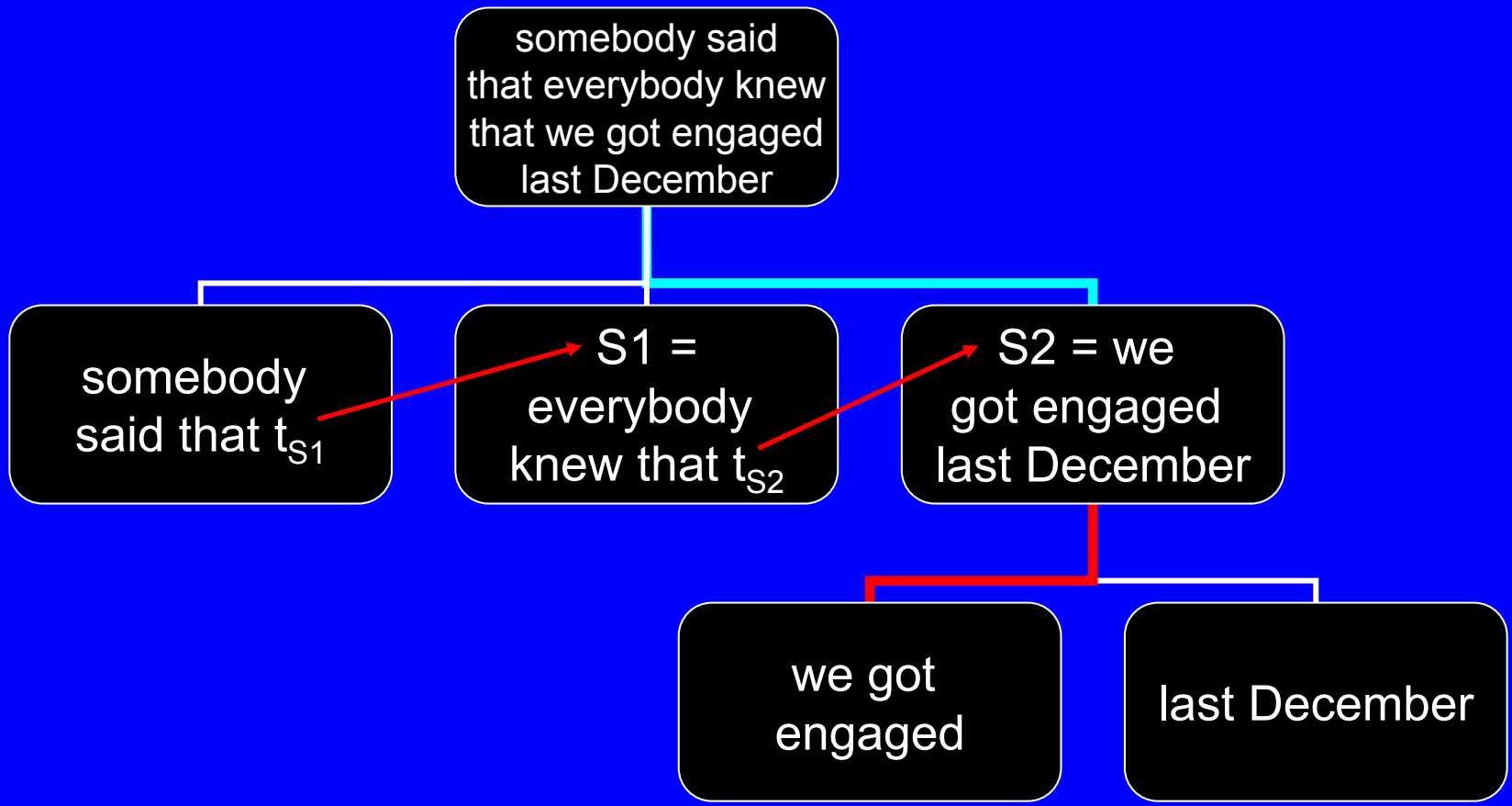


Preference for high and low to middle attachment

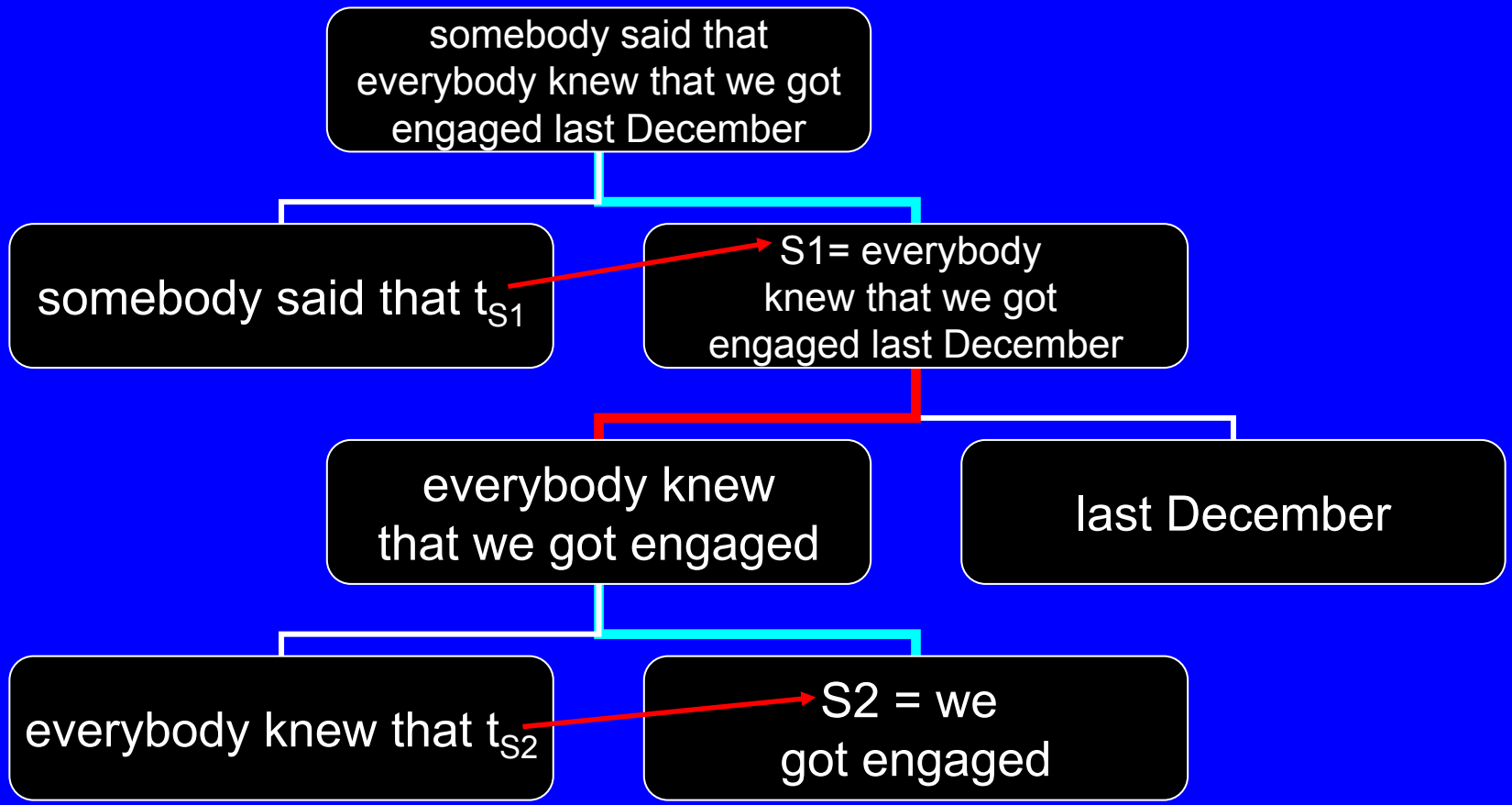
- High and low attachments are preferred to middle attachment.
- Degree of zigzag embedding (ZE°) correlates with these preferences.
  - High and low attachment structures have ZE° = 1.
  - Middle attachment structures have ZE° = 2.

ΨP structure for high attachment

ΨP structure for low attachment



ΨP structure for middle attachment



Significance of ZE

- ZE is a more sensitive measure of structural complexity than CE.
  - $CE^\circ \leq ZE^\circ \leq 2*CE^\circ$
- ZE° is preserved under mapping to ΨP structure.

How regular natural languages might grow onto- or phylo-genetically

1. Start with nonembedding structures, including paratactic ones.
2. Next, recognize or build ΨP structures rather than R/LE structures.
3. Next, recognize or build ΨP structures with 1 ZE° = 1.
4. Continue with ZE° = 2, etc. as needed.

Conclusion

- The procedure approximates but never achieves the coverage of a CE phrase-structure grammar.
- The upper bound on ZE° (and hence also on CE°) arises organically from the construction. It is not an arbitrarily imposed restriction.