Discontinuity in TexMecs, Goddag structures, and Rabbit/duck grammars

Michael Sperberg-McQueen
Claus Huitfeldt

Balisage: The Markup Conference
August 13, 2008, Montréal, Canada

Markup Discontinued

Discontinuity in TexMecs, Goddag structures, and Rabbit/duck grammars

Michael Sperberg-McQueen
Claus Huitfeldt

Balisage: The Markup Conference

August 13, 2008, Montréal, Canada

Background

Markup Languages for Complex Documents (MLCD)

- People involved*
- Problems addressed:
  - Overlap, concurrent hierarchies
  - Virtual elements, alternate orderings
  - ...discontinuous or interrupted elements
- Solutions proposed:
  - Notation: TexMecs
  - Data Structure: Goddag
  - Validation: Rabbit/duck grammars

The structure of this talk

1. The Problem of discontinuity
2. TexMecs notation
3. Goddag structures
4. Containment and dominance
5. Rabbit/duck grammars
6. Relation to other proposals
7. Conclusion

1. The Problem of discontinuity

Lewis Carroll, *Alice's Adventures in Wonderland*, p 83:

> "To begin with," said the Cat, "a dog's not mad. You grant that?"
> "I suppose so," said Alice.

In XML:

```
<p>
  <q>To begin with,</q>
 said the Cat,
  <q>a dog's not mad. You
  grant that?</q>
  <q>I suppose so,</q>
 said Alice.
</p>
```

Questions:

- How many quotations, two or three?.
- Does the phrase "To begin with, a dog's not mad" occur in the document?

This problem is not new...

The TEI P1
in.quot element

```
<p>
  <q>To begin with,<in.quot>
    said the Cat
    </in.quot> a dog's not mad. You
    grant that?</q>
  <q>I suppose so,</q>
 said Alice.
</p>
```

The TEI P2
part attribute

```
<p>
  <q part="I">To begin with,</q>
  said the Cat,
  <q part="F">a dog's not mad. You
  grant that?</q>
  <q part="N">I suppose so,</q>
 said Alice.
</p>
```

The TEI next and prev attributes

```
<p>
  <q id="q1" next="q2">To begin with,</q>
    said the Cat
  <q id="q2" prev="q1">a dog's not mad. You
    grant that?</q>
  <q>I suppose so,</q>
 said Alice.
</p>
```

The TEI join element

```
<join result="q" targets="q1 q2"
scope="branches">
```

...it is not limited to quotations (”wormholes in the universe of discourse”):

- stage directions, speaker attributions...?
- notes...??
- parenthetical remarks, comments, ...???

...and it is not limited to texts:

- annotation of multimedia objects
- digital forensics (e.g. hard disk recovery)
- gene transcription

2. TexMecs notation

```
<p|
  <q|To begin with,|q>
 said the Cat,
  <q|a dog's not mad. You
  grant that?|q>
  <q|I suppose so,|q>
 said Alice.
|p>
```

```
<p|
  <q|To begin with,|-q>
 said the Cat,
  <+q|a dog's not mad. You
  grant that?|q>
  <q|I suppose so,|q>
 said Alice.
|p>
```

3. Goddag structures [simplify this slide]

Informally:

- Goddags are **tangled trees** with shared substructures, where overlap simply is **multiple parentage**

More formally:

- A Goddag is a directed acyclic graph), where:
- Every node is either a *leaf* or a *non-terminal*.
- Each leaf is labeled with a string.
- Each non-terminal is labeled with an identifier.
- Directed arcs identify *parent*/*child* relation; paths identify *ancestor*/*descendant* relation.
- Node *n* is a leaf node iff *n* is not a parent.
- Node *n* is a non-terminal node iff *n* is not a leaf node.

Features:

- arcs = parent/child relations
- each parent's children are ordered
- simple inheritance
- additive meaning / override
- positional meaning

Kinds of Goddags:

- Restricted (overlap-only) Goddags
  - cf. Yves Marcoux, Balisage 2008
- Clean Goddags
  - clean Goddags have no empty string nodes
- Colored Goddags
  - Colored arcs, each color identifies a separate Goddag
- Unresolved issue:
  - Dominance implies containment, but does containment also imply dominance? (cf. Balisage 2008 paper)

Overlapping elements

Broadway Hit or Miss

Overlapping elements

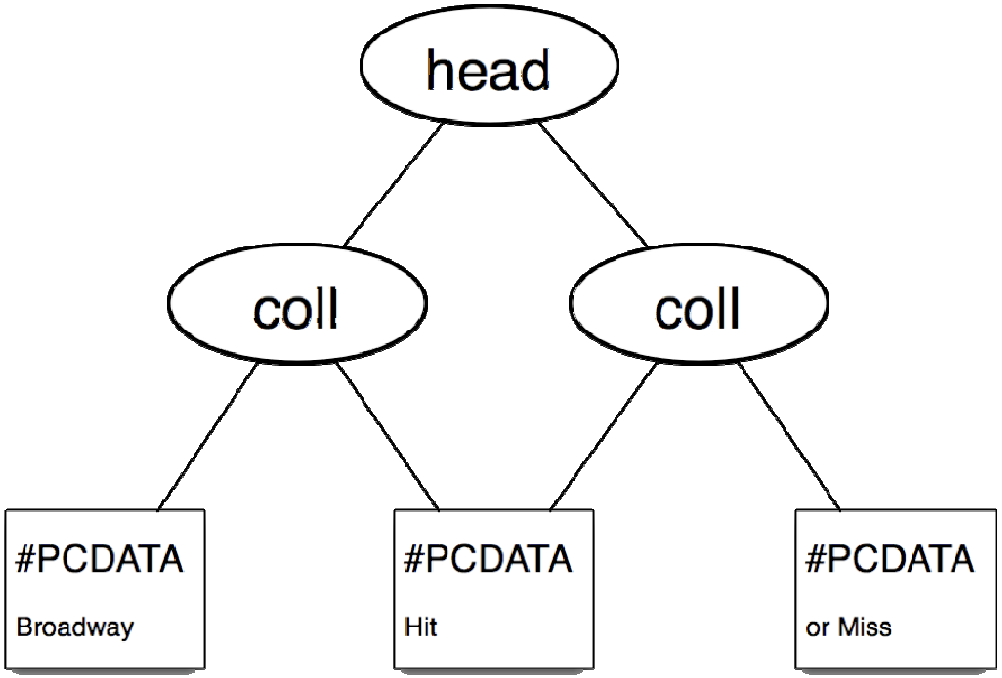
Broadway Hit or Miss

<head| <coll~1|Broadway <coll~2|Hit|coll~1> or Miss|coll~2> |head>

Overlapping elements

Broadway Hit or Miss

<head| <coll~1|Broadway <coll~2|Hit|coll~1> or Miss|coll~2> |head>



…virtual elements:

(*Donald's house. Hughie, Louis, and Dewey.*)

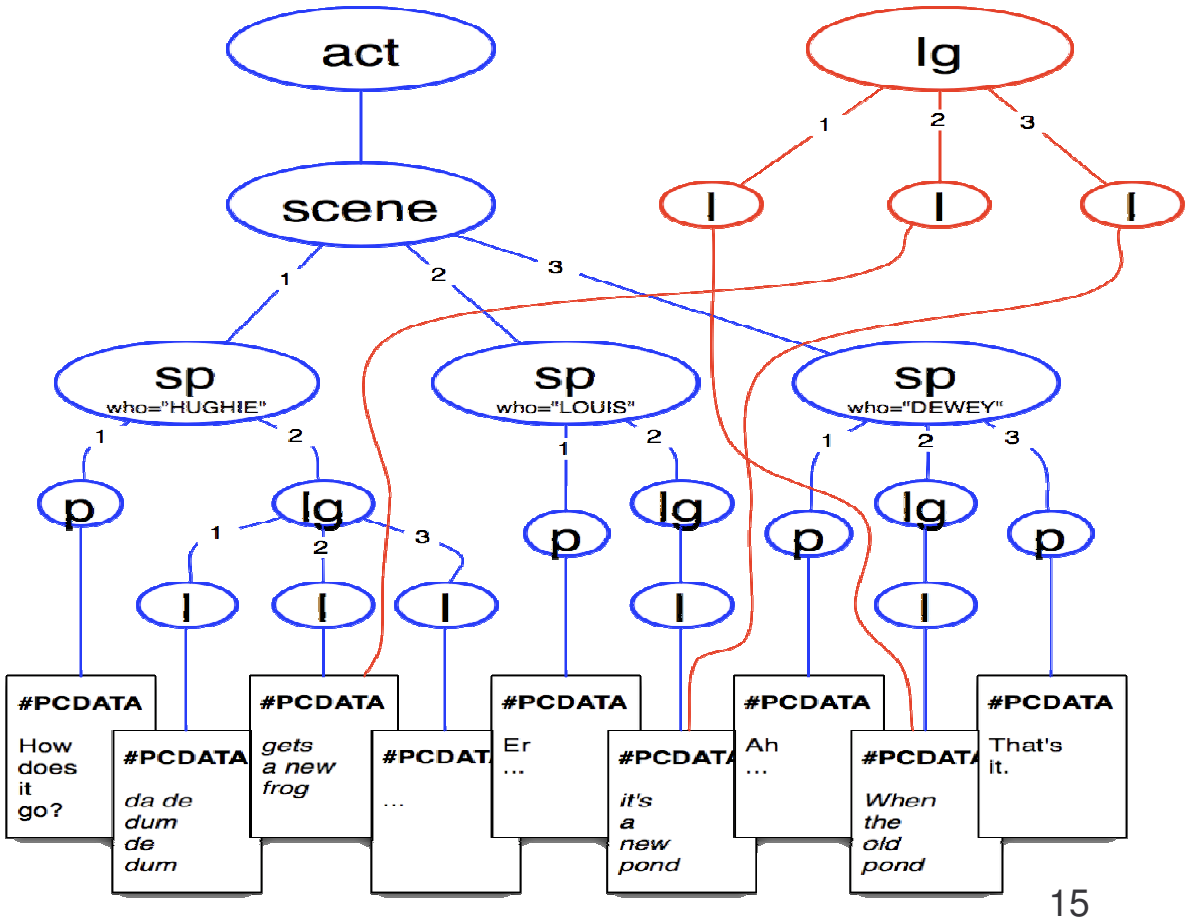
HUGHIE: How did that translation go?
da de dum de dum,
gets a new frog,
...
LOUIS: Er ...
it's a new pond.

DEWEY: Ah ...
When the old pond
Right. That's it.

Is this poem *in* the text?

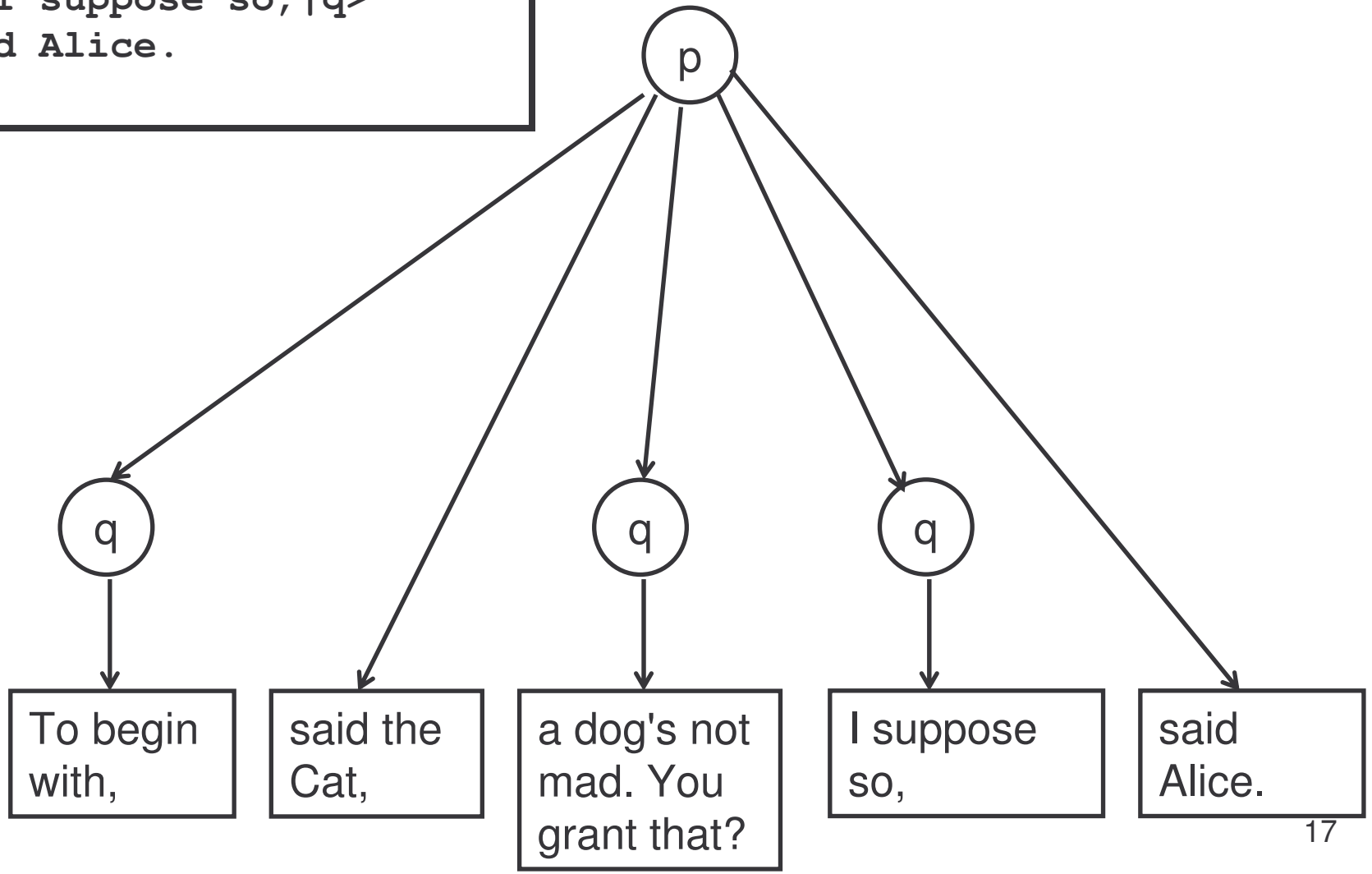
When the old pond
gets a new frog,
it's a new pond.

```
<act|<scene|
<sp who="HUGHIE"|
<p|How did that translation go?|p>
<Lg type="haiku"|
  <L|da de dum de dum,|L>
  <L@frog|gets a new frog,|L>
  <L|...|L>|Lg> |sp>
<sp who="LOUIS"|
<p|Er ...|p> <Lg|
  <L@new|it's a new pond.|L>|Lg> |sp>
<sp who="DEWEY"|
  <p|Ah ...|p> <Lg|
  <L@pond|When the old pond|L>|Lg>
  <p|Right. That's it.|p> |sp>
|act>|scene>
<Lg|<^L^pond><^L^frog><^L^new>|Lg>
```

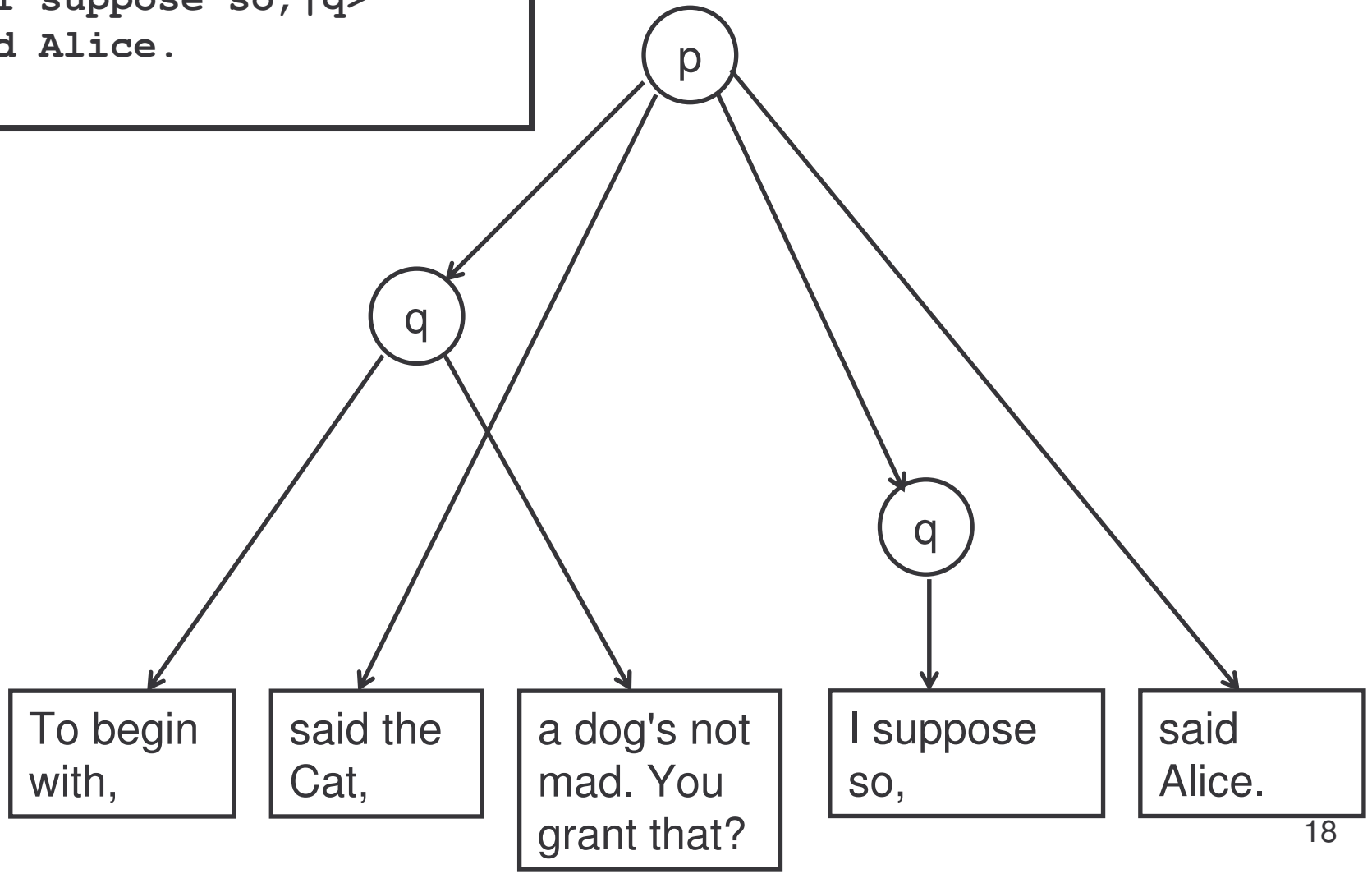


...back to Alice...

```
<p|
  <q|To begin with,|q>
  said the Cat,
  <q|a dog's not mad. You
  grant that?|q>
  <q|I suppose so,|q>
  said Alice.
|p>
```



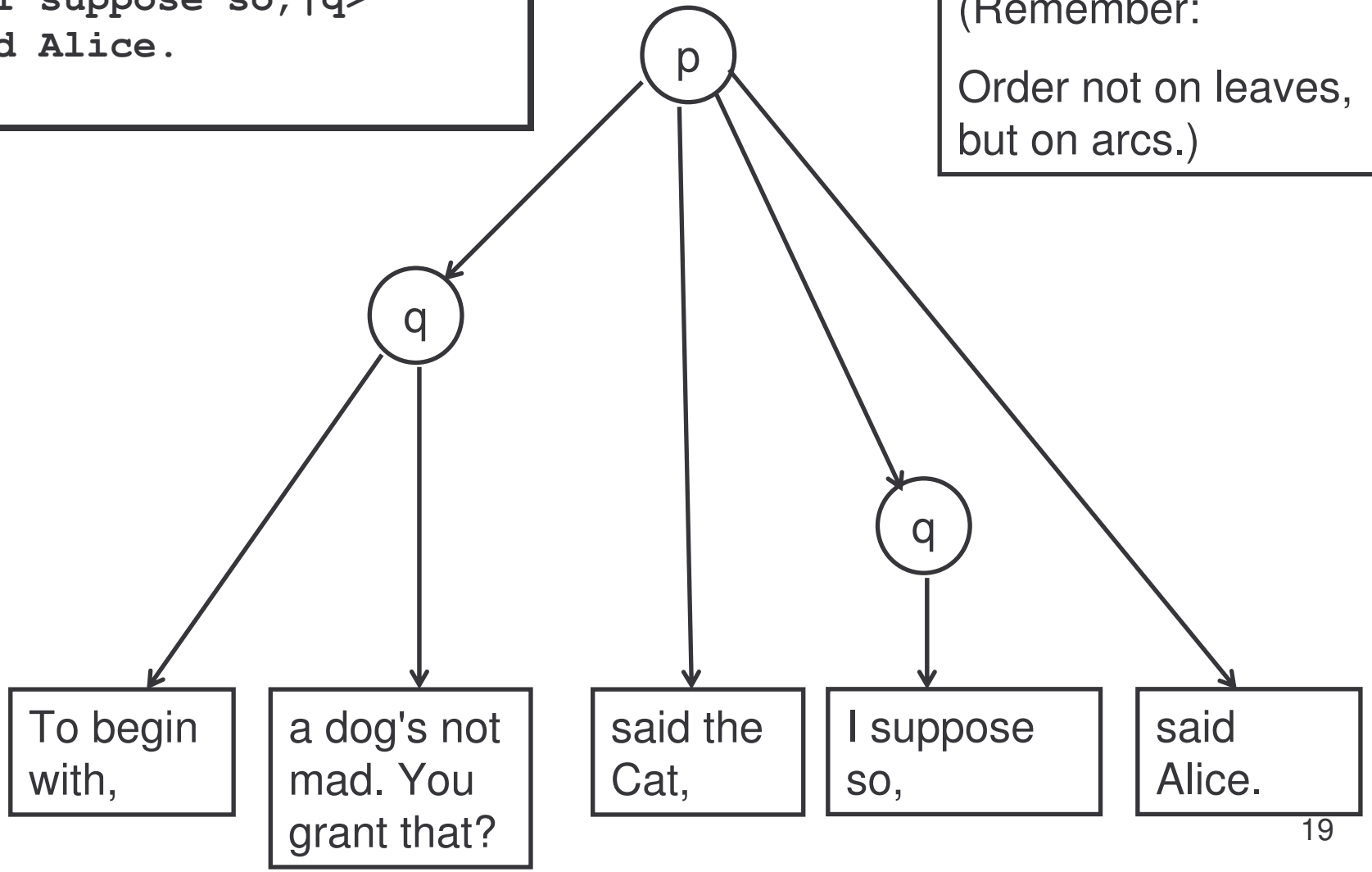
```
<p|
  <q|To begin with,|-q>
  said the Cat,
  <+q|a dog's not mad. You
  grant that?|q>
  <q|I suppose so,|q>
  said Alice.
|p>
```



```
<p|
  <q|To begin with,|-q>
  said the Cat,
  <+q|a dog's not mad. You
  grant that?|q>
  <q|I suppose so,|q>
  said Alice.
|p>
```

(Remember:

Order not on leaves, but on arcs.)



4. Containment and dominance

We had thought that

If p dominates q, then p must contain q.

If p contains* q, then p dominates q.

But perhaps the latter doesn't hold?

Or, in other words: we had thought that

containment* implies dominance

— but perhaps it doesn't ?

Let's try...

Originally, we said:

R2 No node dominates another node both directly and indirectly.

R1 One node dominates a second if and only if the second node is completely contained within the first.

But now we think about saying (instead of R1) only that:

R5 Given two nodes *A* and *B*, if *A* dominates *B*, then *A* contains *B*.

Details: OK, we originally said:

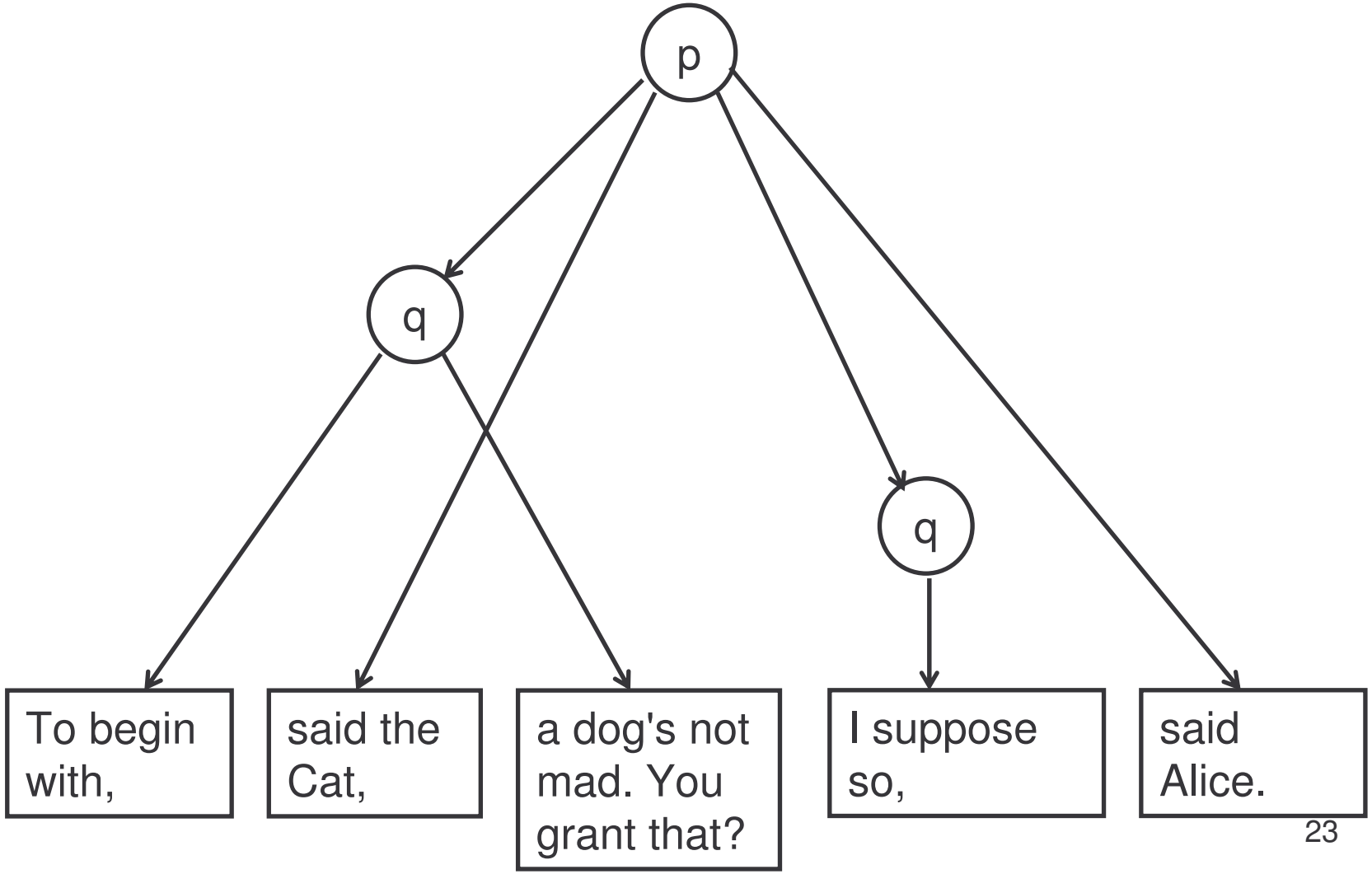
R1 One node dominates a second if and only if the second node is completely contained within the first.

We should have been more precise, i.e. instead of R1 we should have said:

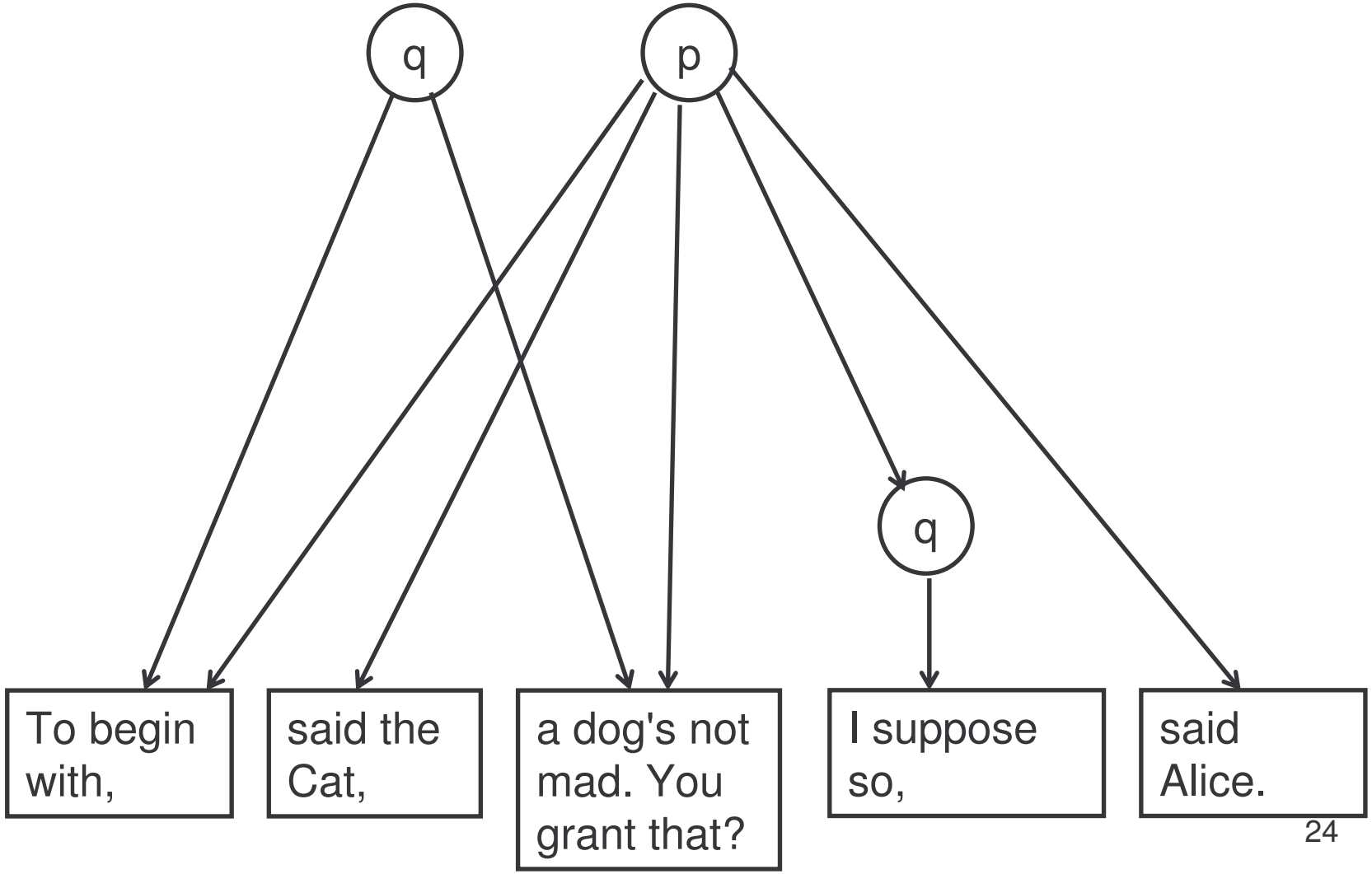
R3 If two nodes *A* and *B* each contain the other (i.e. they dominate the same set of leaves), then either *A* dominates *B* or *B* dominates *A*.

R4 Given two nodes *A* and *B* which do not dominate the same set of leaves, *A* dominates *B* if and only if *A* contains *B*.

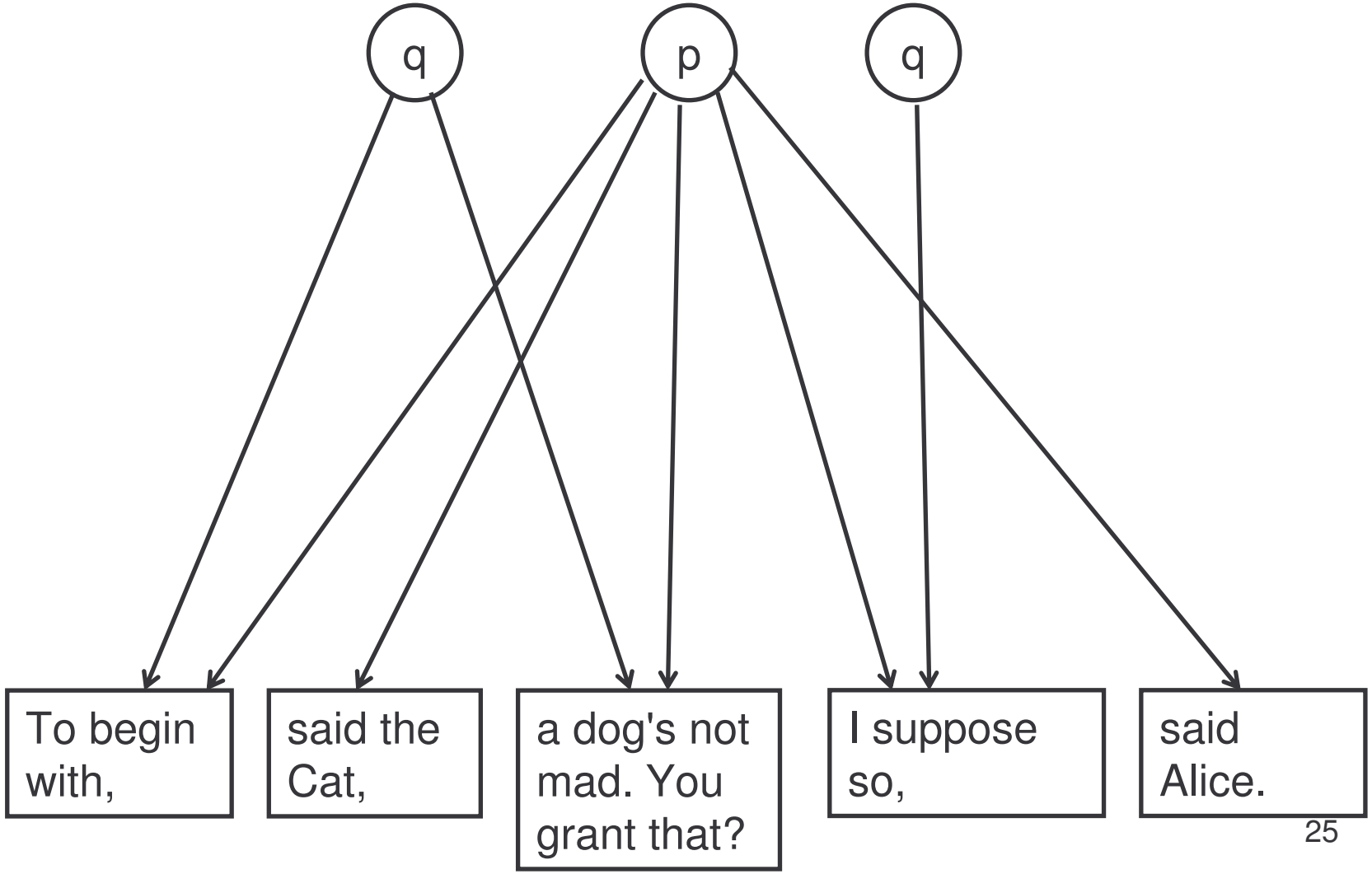
Instead of this...



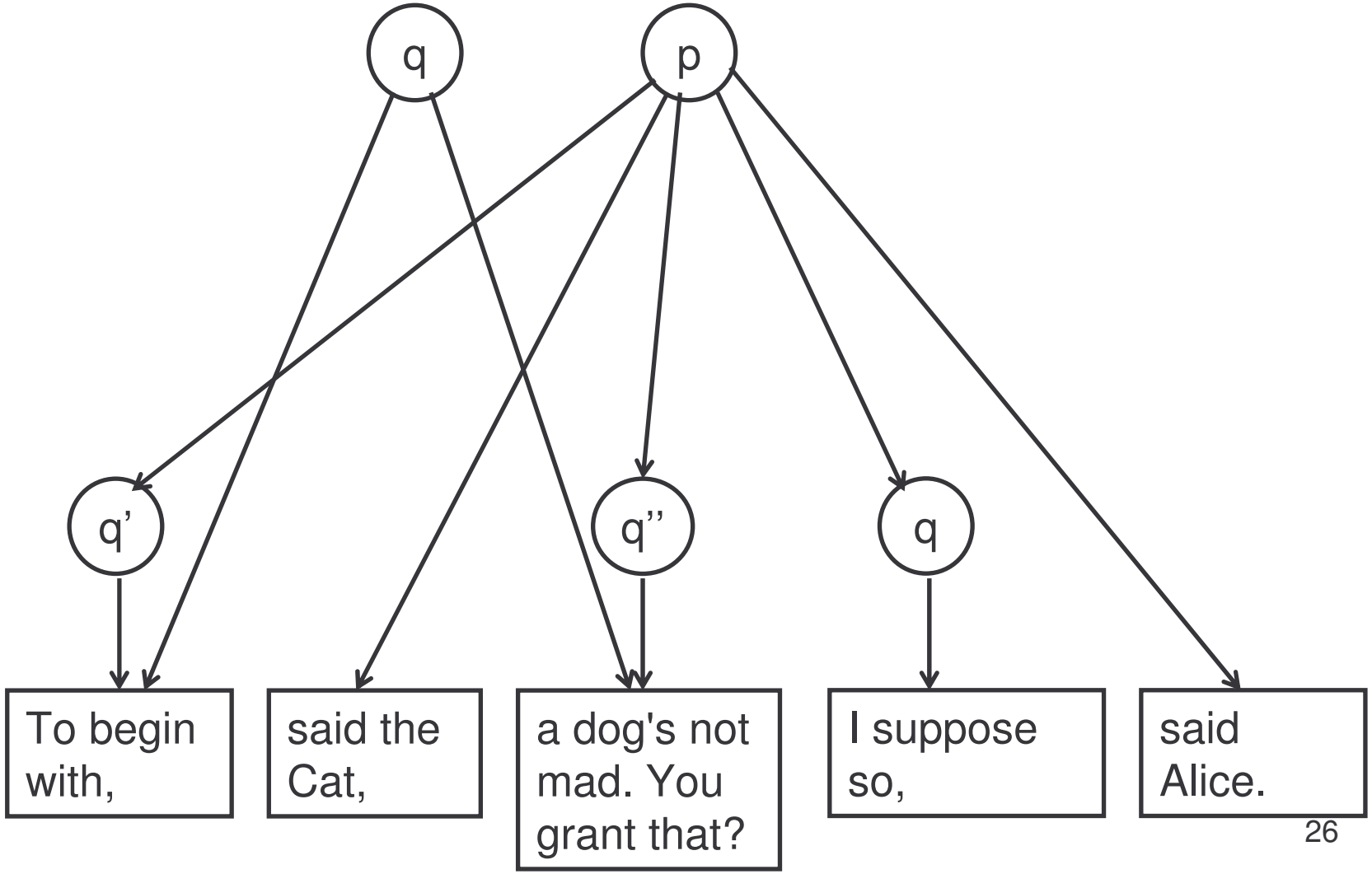
we may get this...

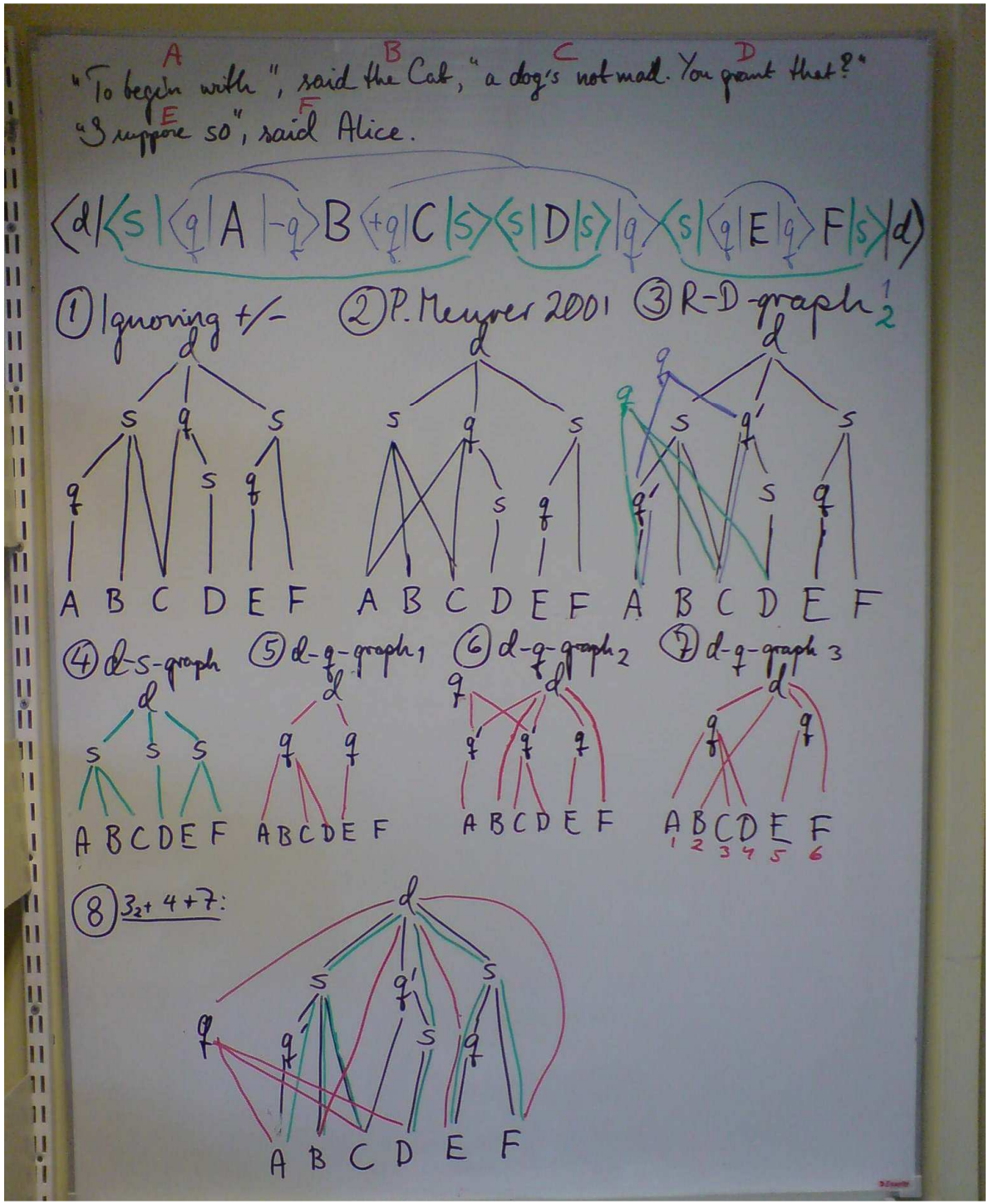


or even this...


q
p
q
To begin with,
said the Cat,
a dog's not mad. You grant that?
I suppose so,
said Alice.

or perhaps rather this...

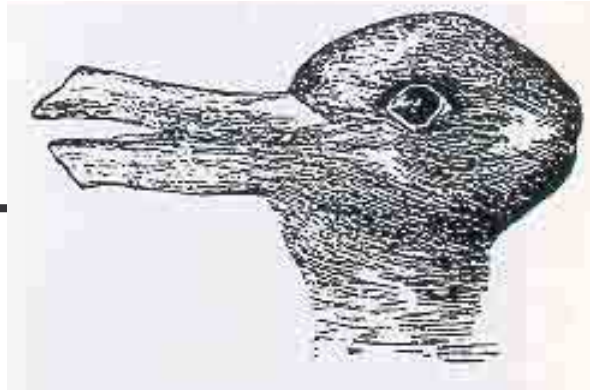



A B C D E F
"To begin with", said the Cat, "a dog's not mad. You grant that?"
"I suppose so", said Alice.
⟨d|⟨s|⟨q|A|-q⟩B⟨+q|C|s⟩⟨s|D|s⟩|q⟩⟨s|⟨q|E|q⟩F|s⟩|d⟩
① Ignoring +/-
② P. Meurer 2001
③ R-D-graph 1/2
④ d-s-graph
⑤ d-q-graph 1
⑥ d-q-graph 2
⑦ d-q-graph 3
⑧ 3₂+ 4+7:

- While this points us to a solution, some mysteries remain. Such as:
  - If p contains q, what decides whether p dominates q?
    - Purely empirical matters of syntax, such as overlap, discontinuity, syntactic flag, etc?
    - Reference to a priori, i.e. external, information? (e.g. a FTSD ?)
  - We don't know

- Next: Our story about validation

5. Rabbit/duck grammars



- Define L as

  $L_1 \cap L_2 \cap \ldots \cap L_n$

  (a context-sensitive language as the intersection of a set of context-free languages)
- Each grammar in a set partitions the vocabulary:
  - **First-class** elements (”normal”)
  - **Second-class** (”milestones”)
  - **Third-class** (”transparent”)
  - **Fourth-class** (”opaque”)
- Each GI in a vocabulary must be first-class in at least one grammar …
- … but may be 2d, 3d, 4th-class in others.

Two invariants

1 Each rule in each grammar is enforced wherever it applies.

2 When two elements overlap there are multiple readings: each element

- is parsed as 1st class in at least one reading
- is 2nd or 3d class in other(s)

Implementation

- filter into *n* XML documents, validate each
- parse against each grammar directly; lexical scanner distinguishes elements by class
- parse in parallel

In our (or Alices') case:

p (#PCDATA | q) *
clearly won't do.

So perhaps this:
p (#PCDATA | q | #frag(q)) *

But then, how does #frag(q) relate to the content model of q?

...or, how do #frag(text) and #frag(body) relate to the content models of text and body?

Assume

text (front, body, back?)

... how do #frag(text) and #frag(body) relate to the content models of text and body?

```
<doc|
   <p|...|p>
   <p|...
         <text|
                  <front| ... |front>
                  <body|... ...
                  |-body>|-text>
                           <p|Just then, we were interrupted.
                           ... |p>
                  <+text|<+body|...
                  |body>
         |text>
   ...|p>
|doc>
```

6. Relation to other proposals

- LMNL
  - dominance solely at application level
- XCONCUR
  - either p dominates q, or not. Basta.
- Trojan Horses
  - Convenient notation, unclear about data structure, no validation.

7. Conclusion

Yes, more work remains to be done in this area...

Thank you...

http://teksttek.aksis.uib.no/projects/mlcd