Unification of Linguistic Resources - Benefits and Issues

Aravind Joshi and Alan Lee

Institute for Research in Cognitive Science &
Department of Computer and Information Science,
University of Pennsylvania

LREC Coscoda/Write Workshop,

Marrakech, Morocco

June 1, 2008

Unifying language resources

- Growing number of linguistic annotations covering phenomena of various levels (morpho-syntax, semantics, discourse, etc.).
- Often annotating the same underlying text.
- Obvious benefits of unifying these linguistic resources:
  - increased reusability
  - interoperability
  - better extensibility
  - etc.

Example - ULA

- Unified Linguistic Annotation (ULA)
- Participants:
  - Penn Treebank (U of Penn)
  - Penn Discourse Treebank (U of Penn)
  - Timebank (Brandeis)
  - Propbank (U of Colorado, Boulder)
  - Nombank (NYU)
  - MPQA Opinion Corpus (U of Pittsburgh)

How much unification?

- The features of one resource should not unncessarily *constrain* the annotation of another resource.

- Linguistic modules are often “layered” - a lower level (e.g. syntax) often projects to a higher level (e.g. discourse). Should the lower layer define the elementary units of the higher level?

A case study…

- How do we get from syntax to discourse?
- Common approach - syntactic clauses are considered the building blocks (elementary units) of discourse relations.
- But we observe that certain annotated features of the Penn Discourse Treebank (PDTB) do not match up neatly with annotations at the syntactic level.
- What do certain mismatches suggest for linguistic theory? How does this affect NLP applications?

Discourse relation

The federal government suspended sales of U.S. savings bonds **because** **Congress hasn't lifted the ceiling on government debt**.

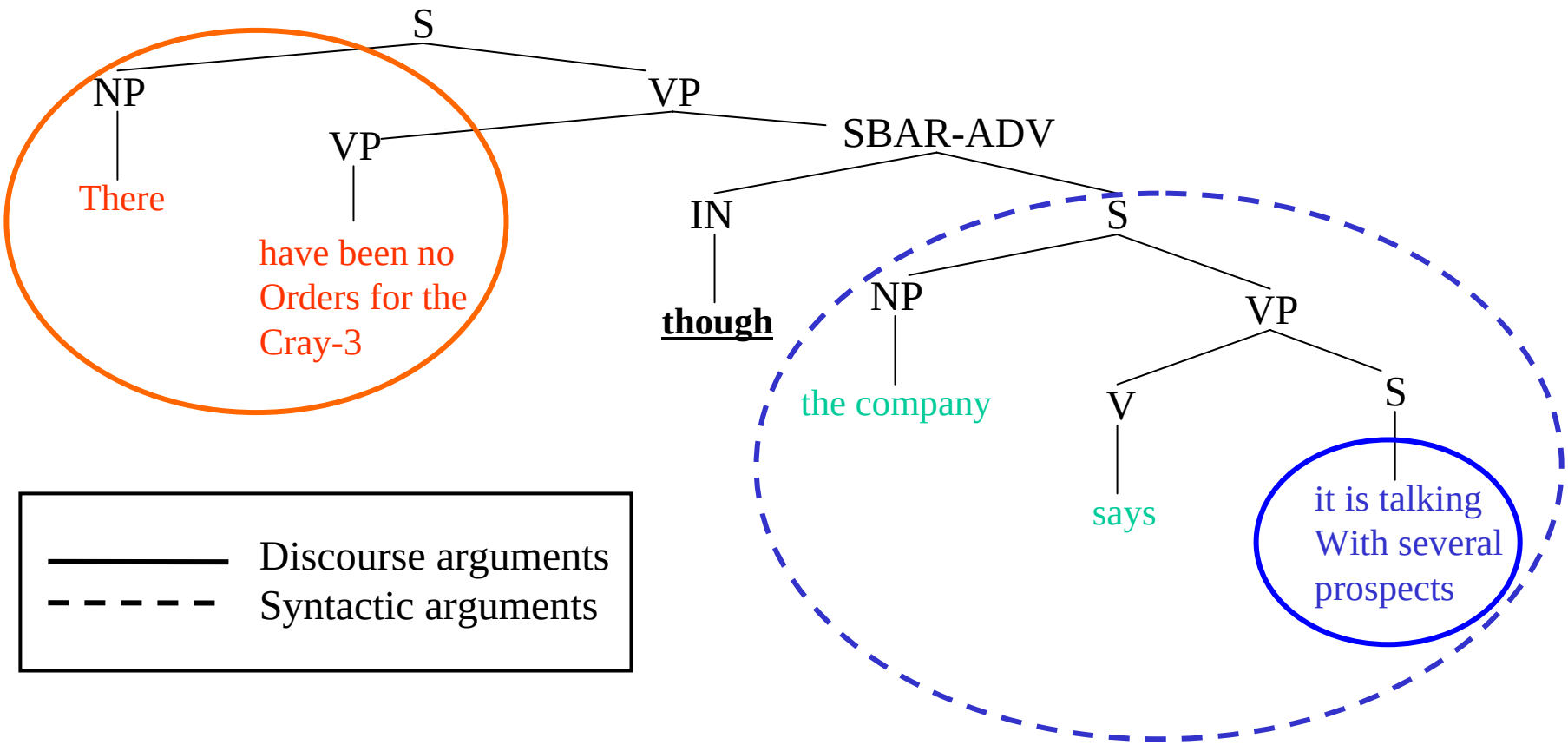
Relation between two “abstract objects” (facts, beliefs, eventualities, states - Asher 93).

Attribution

The company says it is talking with several prospects.

- Attribution is a relation between an agent (“the company”, in this case) and an event/state (“it is talking with several prospects”).

- There have been no orders for the Cray-3 so far, **though** **the company says it is talking with several prospects**.
  - ✓ Discourse semantics: contrary-to-expectation relation between “there being no orders for the Cray-3” and “there being a possibility of some prospects”.
  - ✖ Sentence semantics: contrary-to-expectation relation between “there being no orders for the Cray-3” and “the company saying something”.



Residual issue

- Attribution cannot always be excluded by default

➢ Advocates said the 90-cent-an-hour rise, to $4.25 an hour by April 1991, is too small for the working poor, **while** opponents argued that the increase will still hurt small business and cost many thousands of jobs.

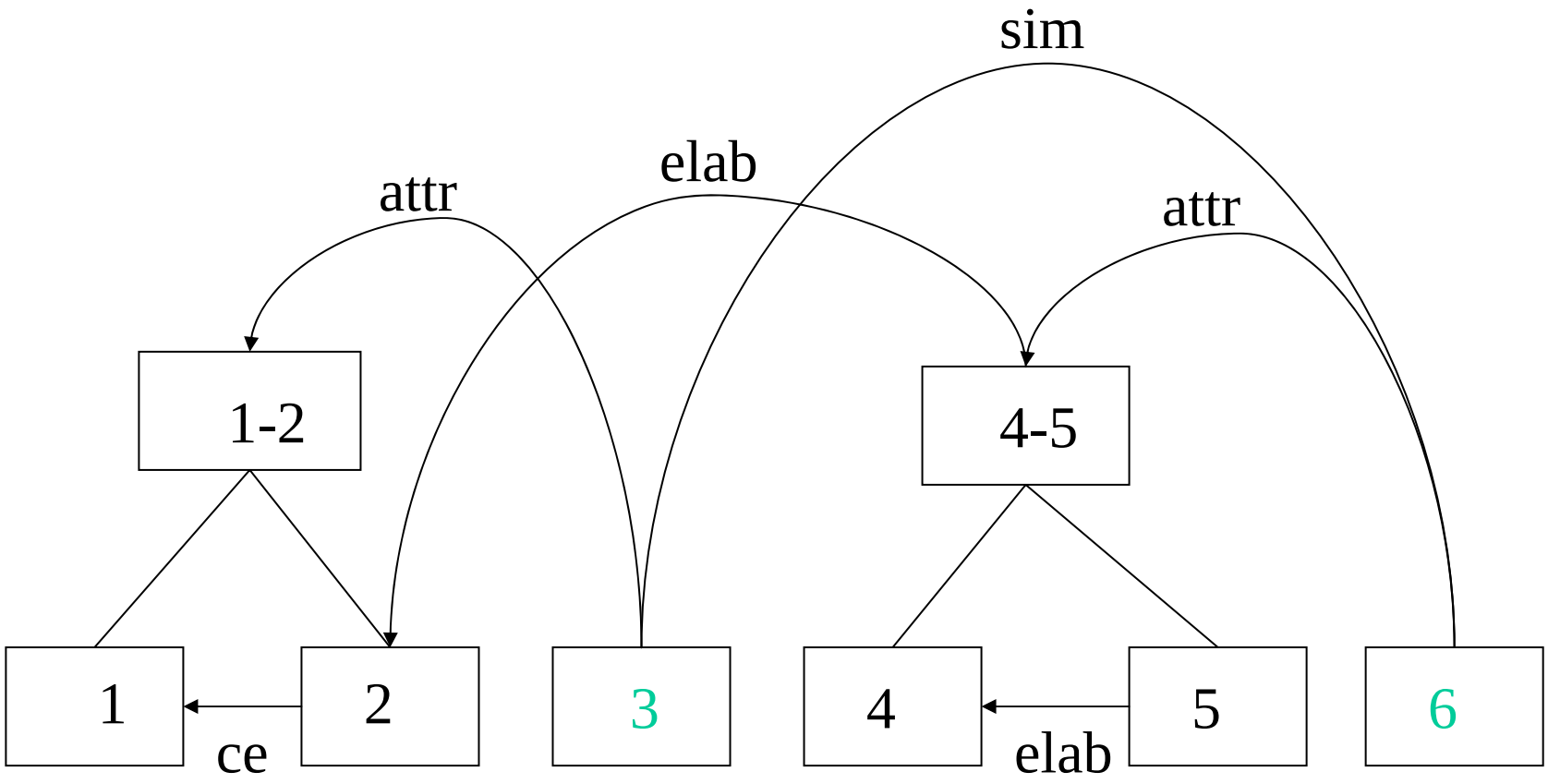
What implications does this have for the approach of treating attribution as an independent layer of discourse?

How to address mismatch?

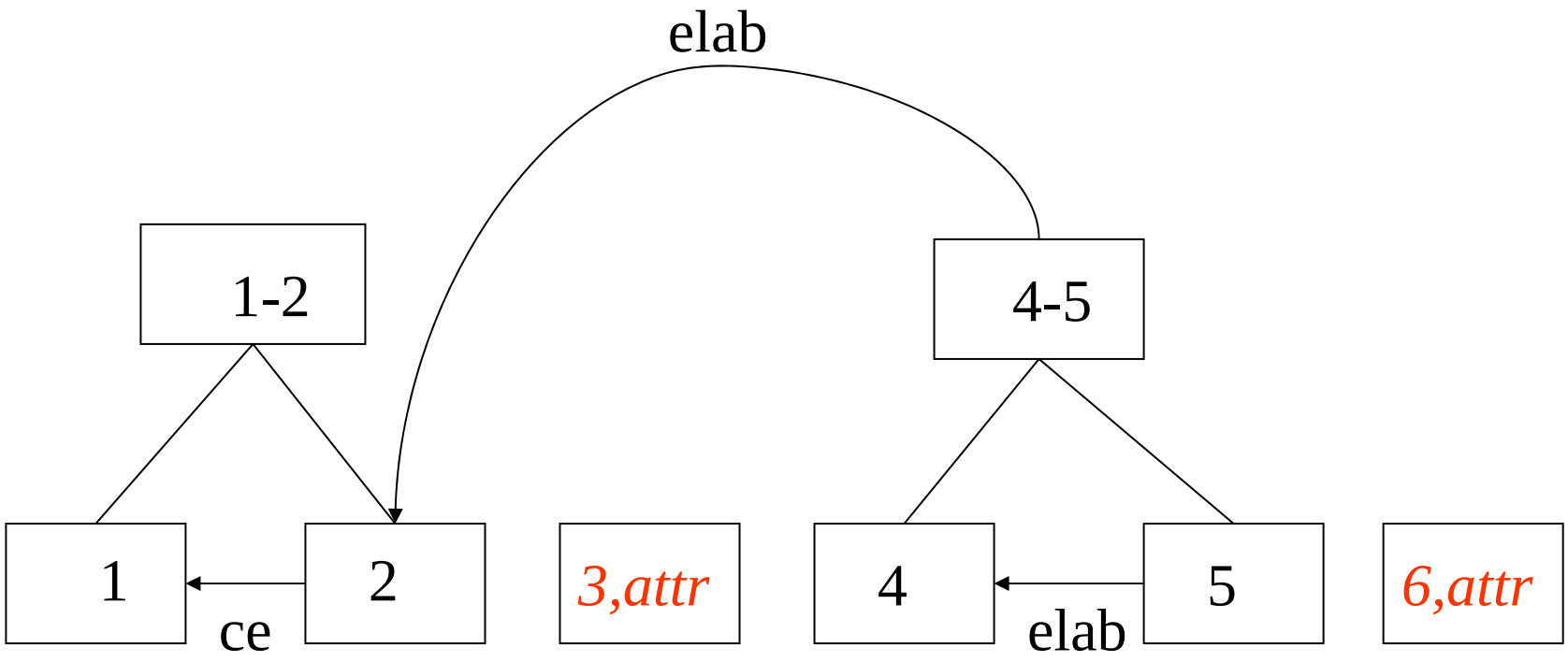
- One possibility - treat attribution as a different layer of structure in discourse. (and also in syntax?)
- This has the effect of reducing the complexity of the discourse structure.

Discourse Graphbank (Wolf & Gibson 2005)

1. Farm prices in October edged up 0.7% from September
2. as raw milk prices continued their rise,
3. the Agriculture Department said.
4. Milk sold to the nation's dairy plants and dealers averaged $14.50 for each hundred pounds,
5. up 50 cents from September and up $1.50 from October 1988,
6. the department said.



ce - cause/effect; elab - elaboration;
sim - similiarity; attr - atribution



ce - cause/effect; elab - elaboration;

[sim - similiarity; attr - atribution]

Alternative Lexicalization (AltLex)

A discourse relation is inferred between two sentences which do not contain an Explicit connective, but insertion of an Implicit connective leads to redundancy. This is because the relation is **alternatively lexicalized** by some non-connective expression:

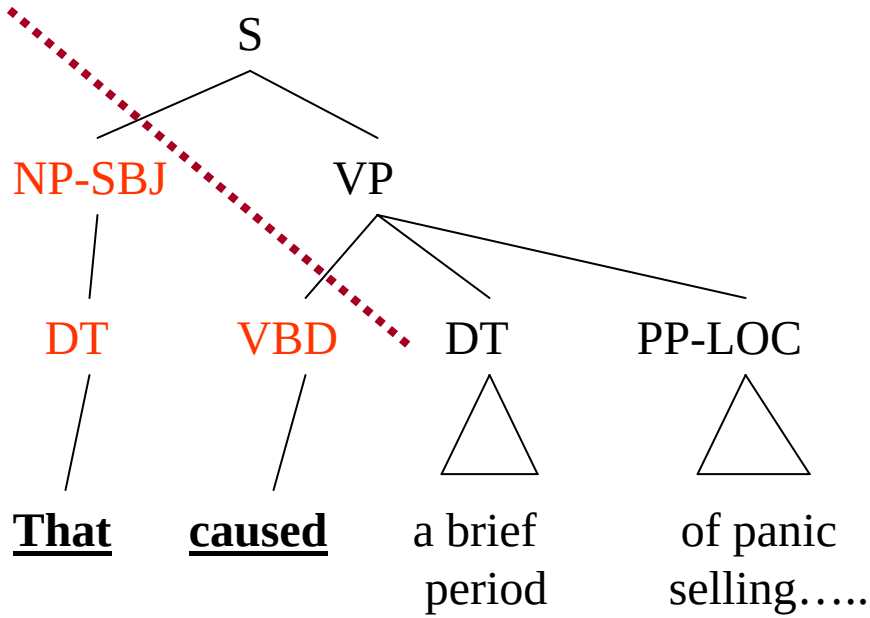
- ***Under a post-1987 crash reform, the Chicago Mercantile Exchange wouldn't permit the December S&P futures to fall further than 12 points for a half hour***. **AltLex = (consequence)** **That caused a brief period of panic selling of stocks on the Big Board**.

Discourse Connectives and Syntactic Constituency

- Discourse connectives were believed initially to correspond to syntactic constituencies. E.g. (“because” IN, “but” CC, “as a result” PP, etc.)

AltLex expressions do not correspond to syntactic constituencies.

Under a post-1987 crash reform, the Chicago Mercantile Exchange wouldn't permit the December S&P futures to fall further than 12 points for a half hour. **AltLex = (consequence) That caused a brief period of panic selling of stocks on the Big Board**.



Conclusion.....

Theoretical issues at linguistic “interfaces” (e.g. the syntax-semantics or the syntax-discourse interfaces) are active areas of research.

Allow for a degree of *independence* in the building of each resource while taking advantage of the benefits of unification. This will allow for a systematic investigation of interactions across linguistic modules, and of the theoretical implications thereof.

Standardization and unification of linguistic annotations of different levels should *illuminate* issues at these interfaces, rather than *eliminate* them.

.....and some possible ideas

- Annotation units based on some data structure identified over the actual source material (text, signals, etc.) rather than over linguistic units. Better consensus over source materials.
- A very modular and constrained approach to designing APIs and tools. How reusable across different annotations or different languages?
- Identify target audience (Are standards meant for language technologists? Linguists? Both?)
- Ongoing interaction between technologists and possible target users... Bridge the divide!