

**Discourse Relations:
Perhaps, a New Kind of
MWE's?**

**Aravind K Joshi
University of Pennsylvania
Philadelphia PA
USA**

**MWE Workshop on Multiword Expressions
COLING 2010
Beijing, China
August 28 2010**

Outline

- **What are MWE's?**
- **A very quick description of the task of annotating discourse relations**
- **Alternate Lexicalizations (AltLex)**
 - **Possibly, a kind of MWE's serving as discourse relations**
 - **Open, Closed, or Partially Open Class?**
 - **Cross Linguistic Implications**
- **Summary**

What are MWE's ?

- MWE's are expressions whose structure and meaning cannot be derived from their component words as they occur independently
- MWE's can be more frequent than their single word paraphrases, for example, single words for elementary spatial relations, e.g.,
 - in front of before
 - next to beside(Alternate Lexicalizations)
- Fundamentality of expressions as opposed to words, as basic units (Filmore 2003)

What are MWE's ?

- MWE's with figurative meanings, sometimes allowing modifications and variations in sentence forms, e.g.,
spill some of the beans
which beans were spilled?
- MWE's as possibly large non-compositional chunks (semantically or syntactically non-compositional) as single units, thus simplifying
 - large scale semantic annotation tasks
- Such expressions arise in the task of annotating discourse relations
- Should one treat them as certain kinds of MWE's or not is an open question?

**A very quick introduction to a discourse
annotation project:
Penn Discourse Annotation Project (PDTB)**

What is a discourse relation?

The meaning and coherence of a discourse results partly from how its constituents relate to each other.

- **Reference relations**

- **Discourse relations**
 - Relation between abstract entities like events, states, and propositions (*Abstract Objects* (Asher 1993))
 - E.g., CONTRAST, CAUSE, CONDITIONAL, TEMPORAL

(1) She hasn't played any music since the earthquake hit.

(Temporal relation between
an event of the earthquake hitting and
a state where no music is played by a certain woman)

How are Discourse Relations Triggered?

Broadly, two ways in which discourse relations may be declared in text:

- **Lexically:**

Discourse Relations can be **grounded in lexical items**. Abstract Objects related by lexically anchored discourse relations can be adjacent or non-adjacent in the text.

(2) She has not played any music since the earthquake hit

- **Structurally through Adjacency:**

Discourse Relations can be triggered by structure underlying adjacency. Such relations have to be **inferred** (but may be partly supported by text).

(3) John left. He was tired. (inference of CAUSAL relation)

The Penn Discourse Treebank (PDTB)

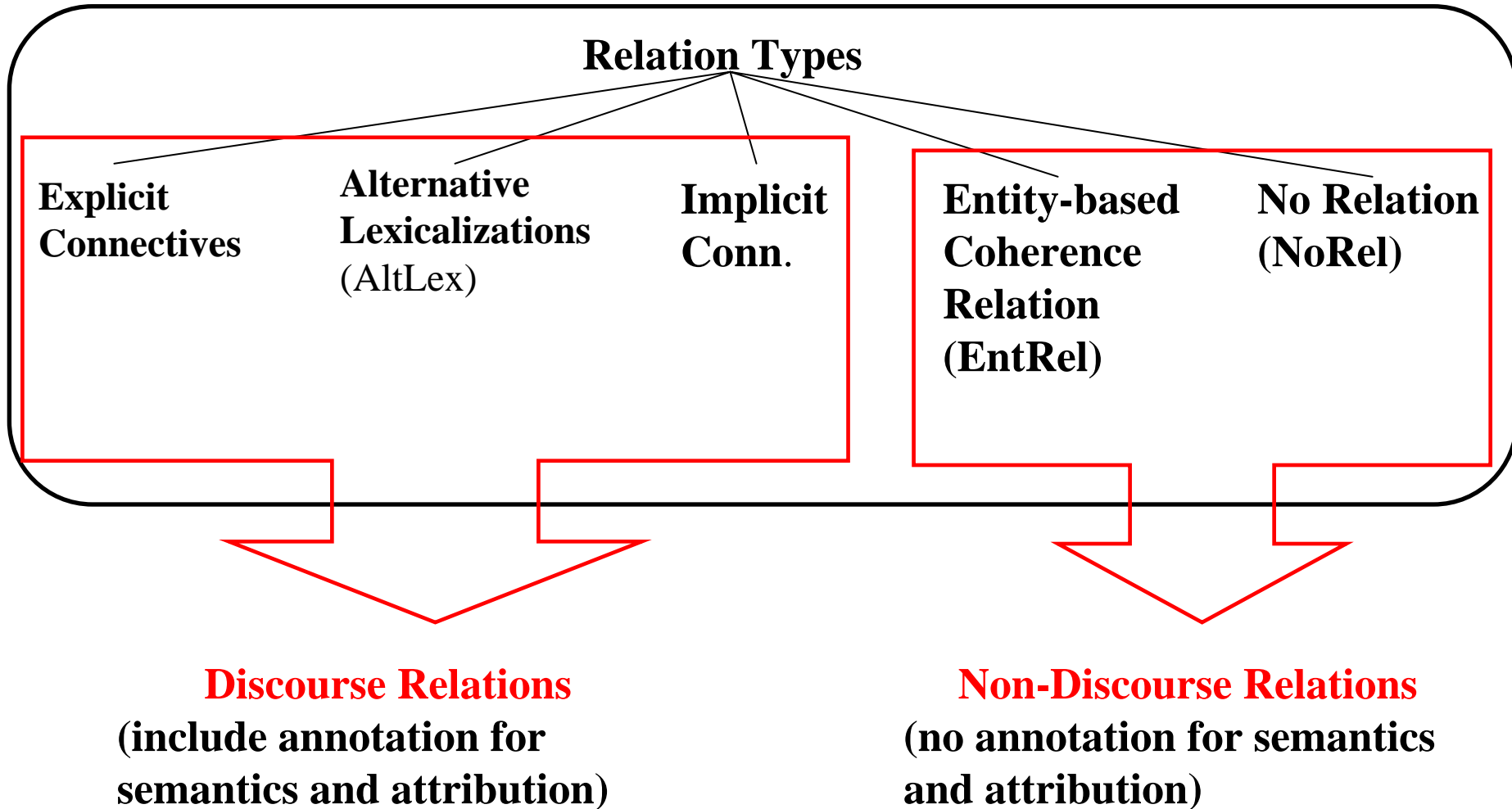
The PDTB provides annotations of both lexically-triggered discourse relations as well as inferred discourse relations triggered via structural adjacency.

Corpus: 2159 Wall Street Journal texts (~1M words) in the Penn Treebank Corpus (Marcus et al., 1993)

What is annotated:

- **Discourse relations** and their **arguments**, as their anchoring text spans
(inferred relations have a dummy anchor)
- **Semantics** of relations, as features
- **Attribution** of relations and their arguments, as text spans anchoring attribution phrases, and features capturing the attribution semantics.

PDTB Annotation Overview



Explicit Connectives

Explicit connectives are drawn from well-defined syntactic classes:

- Subordinating conjunctions (e.g., *when, because, although*, etc.)
 - *The federal government suspended sales of U.S. savings bonds* because Congress hasn't lifted the ceiling on government debt.
- Coordinating conjunctions (e.g., *and, or, so, nor*, etc.)
 - *The subject will be written into the plots of prime-time shows,* and viewers will be given a 900 number to call.
- Discourse adverbials (e.g., *then, however, as a result*, etc.)
 - *In the past, the socialist policies of the government strictly limited the size of ... industrial concerns to conserve resources and restrict the profits businessmen could make.* As a result, industry operated out of small, expensive, highly inefficient industrial units.

N.B. Discourse relations have two and only 2 AO arguments:

Arg2 is the clause with which connective is syntactically associated

Arg1 is the other argument

Argument Labels and Linear Order

- **Arg2** is the sentence/clause with which connective is syntactically associated.
- **Arg1** is the other argument.
- No constraints on relative order. Discontinuous annotation is allowed.
 - **Linear:**
 - *The federal government suspended sales of U.S. savings bonds* because Congress hasn't lifted the ceiling on government debt.
 - **Interposed:**
 - *Most oil companies*, when they set exploration and production budgets for this year, *forecast revenue of \$15 for each barrel of crude produced.*
 - *The chief culprits*, he says, *are big companies and business groups that buy huge amounts of land "not for their corporate use, but for resale at huge profit."* ... The Ministry of Finance, as a result, has proposed a series of measures that would restrict business investment in real estate even more tightly than restrictions aimed at individuals.

Location of Arg1

- Same sentence as Conn and Arg2:
 - *The federal government suspended sales of U.S. savings bonds because Congress hasn't lifted the ceiling on government debt.*
- Sentence immediately previous to Conn and Arg2:
 - *Why do local real-estate markets overreact to regional economic cycles? Because real-estate purchases and leases are such major long-term commitments that most companies and individuals make these decisions only when confident of future economic stability and growth.*
- Previous sentence non-contiguous to Conn and Arg2:
 - Mr. Robinson ... said *Plant Genetic's success in creating genetically engineered male steriles doesn't automatically mean it would be simple to create hybrids in all crops.* That's because pollination, while easy in corn because the carrier is wind, is more complex and involves insects as carriers in crops such as cotton. "It's one thing to say you can sterilize, and another to then successfully pollinate the plant," he said. Nevertheless, he said, **he is negotiating with Plant Genetic to acquire the technology to try breeding hybrid cotton.**

Multiple Clauses: Minimality Principle

- Any number of clauses can be selected as arguments:
 - *Here in this new center for Japanese assembly plants just across the border from San Diego, turnover is dizzying, infrastructure shoddy, bureaucracy intense. Even after-hours drag; "karaoke" bars, where Japanese revelers sing over recorded music, are prohibited by Mexico's powerful musicians union. Still, 20 Japanese companies, including giants such as Sanyo Industries Corp., Matsushita Electronics Components Corp. and Sony Corp. have set up shop in the state of Northern Baja California.*

But, the selection is constrained by a **Minimality Principle**:

- Only as many clauses and/or sentences should be included as are minimally required for interpreting the relation. Any other span of text that is perceived to be relevant (but not necessary) should be annotated as **supplementary information**:
 - **Sup1** for material supplementary to **Arg1**
 - **Sup2** for material supplementary to **Arg2**

Supplements to Arguments

Example of **Sup1**:

Mr. Robinson of Delta & Pine, the seed producer in Scott, Miss., said ***Plant Genetic's success in creating genetically engineered male steriles doesn't automatically mean it would be simple to create hybrids in all crops.*** That's because pollination, while easy in corn because the carrier is wind, is more complex and involves insects as carriers in crops such as cotton. "It's one thing to say you can sterilize, and another to then successfully pollinate the plant," he said. **Nevertheless**, he said, **he is negotiating with Plant Genetic to acquire the technology to try breeding hybrid cotton.**

Annotation Overview: Explicit Connectives

- All WSJ sections (25 sections; 2159 texts)
- 100 distinct types
 - Subordinating conjunctions – 31 types
 - Coordinating conjunctions – 7 types
 - Discourse Adverbials – 62 types
- 18,459 distinct tokens

Implicit Connectives

When there is no Explicit connective present to relate adjacent sentences, it may be possible to **infer** a discourse relation between them **due to adjacency**.

- *Some have raised their cash positions to record levels.*
Implicit=because High cash positions help buffer a fund when the market falls.
- *The projects already under construction will increase Las Vegas's supply of hotel rooms by 11,795, or nearly 20%, to 75,500.* **Implicit=so** By a rule of thumb of 1.5 new jobs for each new hotel room, Clark County will have nearly 18,000 new jobs.

Such discourse relations are annotated by **inserting an “Implicit connective” that “best” captures the relation.**

- Sentence delimiters are: period, semi-colon, colon
- Left character offset of Arg2 is “placeholder” for these implicit connectives.

Non-insertability of Implicit Connectives

There are three types of cases where **Implicit connectives cannot be inserted** between adjacent sentences.

- **AltLex**: A discourse relation is inferred, but insertion of an Implicit connective leads to redundancy because the relation is **alternatively lexicalized** by some non-connective expression:
 - *A few years ago, the company offered two round-trip tickets on Trans World Airlines to buyers of its Riviera luxury car.* The promotion helped Riviera sales exceed the division's forecast by more than 10%, Buick said at the time.

Non-insertability of Implicit Connectives

- **EntRel:** the coherence is due to an entity-based relation.
 - *Hale Milgrim, 41 years old, senior vice president, marketing at Elektra Entertainment Inc., was named president of Capitol Records Inc., a unit of this entertainment concern.* EntRel Mr. Milgrim succeeds David Berman, who resigned last month.
- **NoRel:** Neither discourse nor entity-based relation is inferred.
 - *Jacobs is an international engineering and construction concern.* NoRel Total capital investment at the site could be as much as \$400 million, according to Intel.

Annotation overview: Implicit Connectives

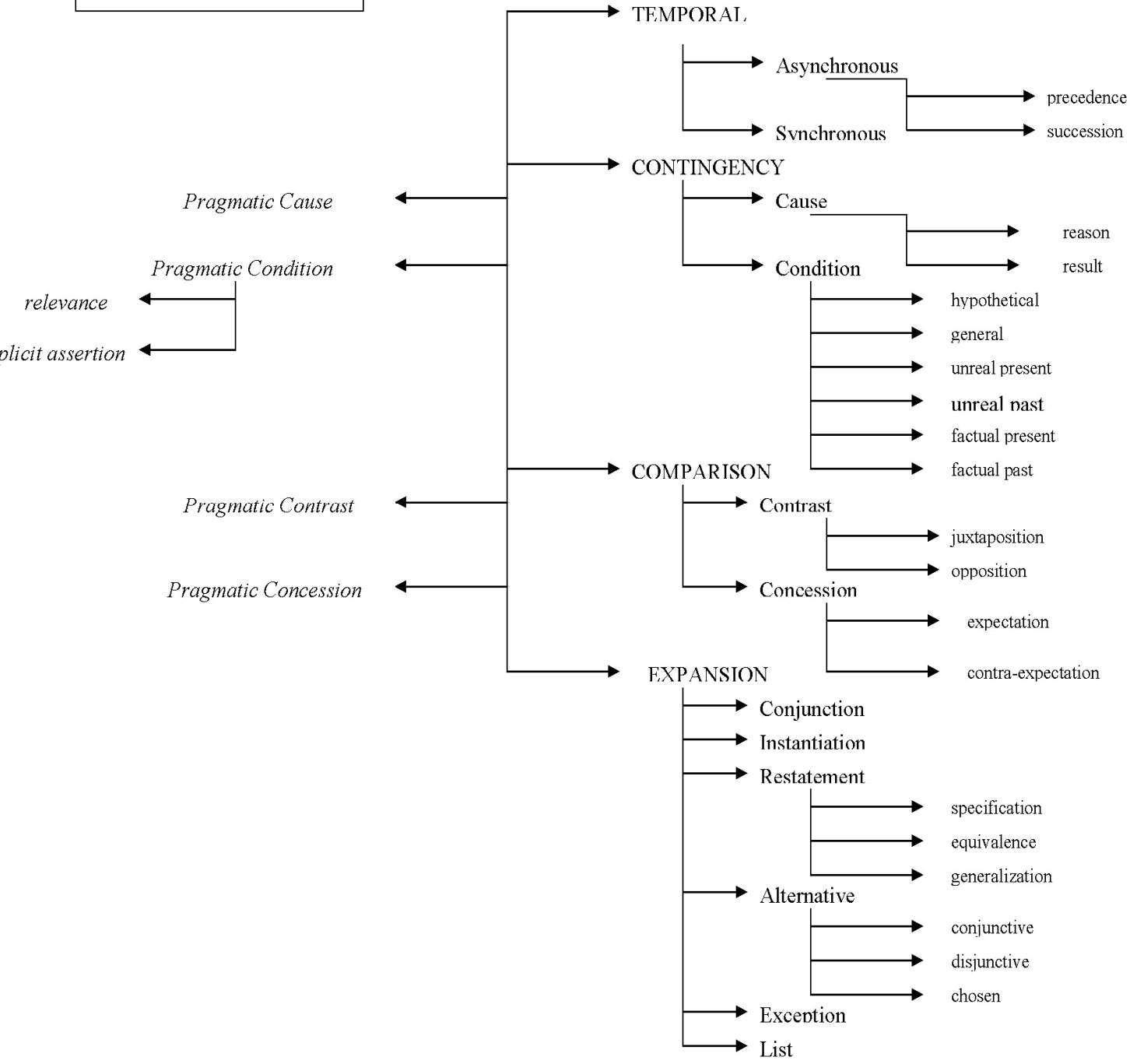
About 22,000 tokens

- **Implicit Connectives:** about 16, 053 tokens
- **AltLex:** 624 tokens
- **EntRel:** 5210 tokens
- **NoRel:** 254 tokens

Annotations of Senses in PDTB

- Sense annotations provided for explicit, implicit and altlex tokens
- Total: 35,312 tokens

Hierarchy of sense tags



Implicit Discourse Connectives: Examples

>> *Some have raised their cash positions to record levels.*
Implicit=because (cause-reason) High cash positions help buffer a fund when the market falls. [wsj_0983]

>> *The projects already under construction will increase Las Vegas's supply of hotel rooms by 11,795, or nearly 20%, to 75,500.* Implicit=so (cause-result) By a rule of thumb of 1.5 new jobs for each new hotel room, Clark County will have nearly 18,000 new jobs. [wsj_0994]

Arg2 is the second sentence.

Arg1 is the first sentence.

When Implicit Connectives could NOT be Inserted

In annotating implicit relations in these adjacent sentence contexts, annotators were not able to insert connectives in many cases! They inserted "NONE" as the connective.

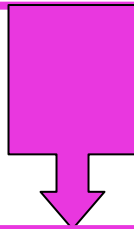
In a later phase of the annotation, NONE tokens (approx 6000) were analyzed further.

About 10% of these tokens did in fact express a discourse relation, but annotators were nevertheless unable to insert a connective!

When Implicit Connectives could NOT be Inserted

Annotators were unable to insert a connective despite the inference of a discourse relation because there was a perceived redundancy after insertion of the connective, and thus the connective did not meet the fluency criteria.

Source of Redundancy: Discourse relation was realized by an expression that had not been pre-classified as a discourse connective.



Alternative Lexicalizations (AltLex)

Examples of AltLex in PDTB

AltLex Expression	(Syntax)	Sense	Adv.Connective Counterpart
Trouble is	(NP-SBJ V)	Concession	However
At the other end of the spectrum	(PP-LOC)	Contrast	??
The reason:	(NP)	Cause-Reason	NONE?
That means	(NP-SBJ V)	Cause-Result	As a result
Beyond that	(PP)	Conjunction	In addition
Probably the most egregious example	(ADV NP-SBJ V)	Instantiation	??
Putting it all Together	(S-ADV)	Restatement	In sum
That was followed by	(NP-SBJ V V P)	Temporal	Then

AltLex Analysis: A Somewhat Closed-class

Some AltLex's are somewhat closed class expressions and potential connectives once propositional anaphoric pronouns referring to Arg1 are also allowed to be part of the connective phrase.

Examples from Knott (1996):

- After that, after this,
- That's why, that is why, this is why,
- This means, that means

We have found many new items that don't appear in previous lists:

- trouble (with that) is, the idea (behind that) is,
the problem (regarding that) is, the reason (for that) is,
the result (of that) is, etc.

Closed-class AltLex: Examples

>> *Certainly, the Oct. 13 sell-off didn't settle any stomachs. Beyond that (conjunction), money managers and analysts see other problems.* [wsj_0359]

>> *She spent a month at an Aetna school in Gettysburg, Pa., learning all about the construction trade, including masonry, plumbing and electrical wiring. That was followed by (temporal) three months at the Aetna Institute in Hartford, where she was immersed in learning how to read and interpret policies.* [wsj_0766]

>> *Mr. Payson, an art dealer and collector, sold Vincent van Gogh's "Iris" at a Sotheby's auction in November 1987 to Australian businessman Alan Bond. Trouble is (Concession), Mr. Bond has yet to pay up, and until he does, Sotheby's has the painting under lock and key.* [wsj_2113]

Closed-class AltLex: Examples

>> *In addition, Unisys must deal with its increasingly oppressive debt load. Debt has risen to around \$4 billion, or about 50% of total capitalization.* That means (cause-result) Unisys must pay about \$100 million in interest every quarter, on top of \$27 million in dividends on preferred stock. [wsj_0568]

>> *Both are in great need of foreign exchange, and South Africa is also under pressure to meet foreign loan commitments,* he said. "Putting it all together (restatement), we have a negative scenario that doesn't look like it will improve overnight," he said. [wsj_1687]

AltLex Analysis: Cause-Reason is Special

Cause-Reason is the only listed sense for which there are no attested adverbial counterparts in English. The preferred way to realize this relation inter-sententially is as an AltLex.

>> *After trading at an average discount of more than 20% in late 1987 and part of last year, country funds currently trade at an average premium of 6%. The reason: (cause-reason) Share prices of many of these funds this year have climbed much more sharply than the foreign stocks they hold.* [wsj_0034]

Is this specific to English, or a linguistic universal?

Hindi, Czech, Turkish, Italian, Arabic

AltLex Analysis: Cause-Reason is Special

There are 7 out of 858 instances of because observed in PDTB that appear as adverbs.

4 of these are in QA contexts

>> *"Why was containment so successful?
Because it had bipartisan support."* [wsj_0771]

“Because” as an Adverb?

Are the remaining 3 simply stylistic aberrations, or evidence of because emerging as an adverb as well?

>> Many of us are suckers. *But what we may not know is just what makes somebody a sucker.* What makes people blurt out their credit-card numbers to a caller they've never heard of? Do they really believe that the number just for verification and is simply a formality on the road to being a grand-prize winner? What makes a person buy an oil well from some stranger knocking on the screen door? Or an interest in a retirement community in Nevada that will knock your socks off, once it is built?

Because in the end, these people always wind up asking themselves the same question: "How could I be so stupid?" [wsj_1572]

N.B: There is other evidence of connectives behaving similarly, e.g., *so*, *but*, *and* (but their adverbial use alternates with their use as coordinating, not subordinating conjunction).

“Because” as an Adverb?

>> Players ran out on the field way below, and the stands began to reverberate. It must be a local custom, I thought, stamping feet to welcome the team. But then the noise turned into a roar. And no one was shouting. *No one around me was saying anything.* Because we all were busy riding a wave. Sixty thousand surfers atop a concrete wall, waiting for the wipeout. [wsj_1643]

>> President Bush told reporters: "Whether that {the leadership change} reflects a change in East-West relations, *I don't think so.* Because Mr. Krenz has been very much in accord with the policies of Honecker." [wsj_1875]

AltLex Analysis: Not so closed-class

>> *Inflation is expected to be highest in Greece, where it is projected at 14.25%, and Portugal, at 13%.*

At the other end of the spectrum (contrast), West German inflation was forecast at 3% in 1989 and 2.75% in 1990.

>> *Typically, these laws seek to prevent executive branch officials from inquiring into whether certain federal programs make any economic sense or proposing more market-oriented alternatives to regulations.* Probably the most egregious example is (instantiation) a proviso in the appropriations bill for the executive office that prevents the president's Office of Management and Budget from subjecting agricultural marketing orders to any cost-benefit scrutiny. There is something inherently suspect about Congress's prohibiting the executive from even studying whether public funds are being wasted in some favored program or other. [wsj_0112]

When do the open-ended AltLex's Occur?

Relation modification to convey more than the bare connective can convey

We do have "modified connectives": e.g., possibly because

But many adverbial connective forms do not allow
Modification - #possibly for example

Modification is possible only after Altlexification!
- a possible example (NP)

-Eventually, some of these may get grammaticized in much the same manner as some current day adverbials - cf. therefore

AltLex Analysis: Not so closed-class

>> *Inflation is expected to be highest in Greece, where it is projected at 14.25%, and Portugal, at 13%.*

At the other end of the spectrum (contrast), West German inflation was forecast at 3% in 1989 and 2.75% in 1990.

>> *Typically, these laws seek to prevent executive branch officials from inquiring into whether certain federal programs make any economic sense or proposing more market-oriented alternatives to regulations. Probably the most egregious example is (instantiation)*

a proviso in the appropriations bill for the executive office that prevents the president's Office of Management and Budget from subjecting agricultural marketing orders to any cost-benefit scrutiny. There is something inherently suspect about Congress's prohibiting the executive from even studying whether public funds are being wasted in some favored program or other. [wsj_0112]

Summary

- Discourse Relations as a possible source of a new kind of MWE's (AltLex items)
- How do AltLex arise?
- AltLex items are a new kind of MWE's
 - They are semantically compositional but not necessarily syntactically
- Is AltLex an open or a closed class, or partially open?
- Are there impossible AltLex?
 - Impossible adverbial AltLex with the sense "cause-reason" ?
 - Is this a universal?