Advancing Discourse-based Sentence Splitting

Bonnie Webber
School of Informatics
University of Edinburgh

Outline

- **Introduction**
  - Sentence splitting [Cripwell, Legrand & Gardent]
  - Exploiting recent work on
    - Intra-Sentential (Intra-S) discourse relations
    - Concurrent explicit and non-explicit relations

- **PDTB-3:** Intra-S discourse annotation
  - relations associated with linguistic constructions
  - Intra-S senses
  - multiple relations that hold between the same args.

  => Exploiting this in discourse-based sentence splitting

Sentence splitting [Cripwell, Legrand & Gardent]

- **Goal is to simplify text**
- Syntax-based splitting considers splits licensed by syntactic structure or by sentence-level semantics such as thematic roles.
- Discourse-based splitting considers splits that reflect relations between clauses, sentences and/or larger units of texts.
- Discourse relations are often signalled by explicit connectives such as conjunctions (coordinating and subordinating) and adverbials.

Sentence splitting [Cripwell, Legrand & Gardent]

Taking a cue from California, more politicians will launch their campaigns by backing initiatives [wsj_0120]

S: (NP-SBJ-1 (JJR more) (NNS politicians))
(VP (MD will)
(VP (VB launch)
(NP (PRP$ their) (NNS campaigns))
(PP (IN by)
(S-NOM (NP-SBJ (-NONE- *-1))
(VP (VBG backing) (NP (NNS initiatives))))))))

D: Taking a cue from California, more politicians will launch their campaigns by backing initiatives (ARG2-AS-MANNER)

=> Taking a cue from California, more politicians will launch their campaigns. The manner of doing so will be by backing initiatives.

Discourse Coherence and discourse relations

Discourse coherence reflects, in part, relations between eventualities and propositions (typically realized as clauses, sentences, or larger segments of text).

Relations can be signalled **explicity** or **implicitly**.

E.g., Relation of REASON:

- John did not eat the fish **because** he is vegetarian.
- John did not eat the fish. **That's because** he is vegetarian.
- John did not eat the fish. He is vegetarian.
- Being vegetarian, John did not eat the fish.

Some work aims to combine individual relations into more complex coherence structures spanning the entirety of a given text ➔ E.g., **RST**, **SDRT**

PDTB only annotates low-level relations, without combining them further.

- The idea was to see if high-level structure might emerge in some way from low-level discourse relations..
- **PDTB-2** was annotated over 40600 relations in the WSJ corpus and was released in 2008.

PDTB Annotation Basics

Text (Discourse)

John did not eat the fish because he is vegetarian

Identify individual relations, their explicit realization (if any) and their (two) arguments

John did not eat the fish **because** he is vegetarian.

Label arguments (Arg1/Arg2) and the sense of the relation

John did not eat the fish **because** he is vegetarian.

Arg1 Contingency.Cause.Reason Arg2

GUIDELINES

Definitions for identifying discourse relations (explicit/implicit) and arguments

Arg naming convention

Sense Classification (as hierarchy)

From PDTB-2 to PDTB-3

Limitations of PDTB-2

- Time/money constraints prevented all relations in a text from being annotated.
- Because annotation was being done for the first time on a large scale, guidelines needed to be improved in order to be more reliable and comprehensive

PDTB-3

- Addressed some major gaps in the corpus, primarily **intra-sentential relations** ➔ ~ 13K new relations
- Modified and extended annotation guidelines to make them more reliable and comprehensive
- Revised guidelines were applied to PDTB2, revising some earlier annotation
- Merging of revised PDTB-2 and new relations ➔ **PDTB-3** (~53K relations)

New Relations

PDTB-2 guidelines precluded various intra-sentential relations.

- Explicit relations lexicalized by discourse connectives, and implicit relations between paragraph-internal adjacent sentences and between (semi-) colon separated clauses within sentences.
- Discourse connectives drawn from the pre-defined syntactic classes
- Arguments realized as one or more clauses or sentences.

Precluded **subordinate clauses that can occur without lexical subordinators** while bearing an implicit relation to their matrix clause.

Free adjuncts

- Treasurys opened lower, **Implicit=as a result of** reacting negatively to news that the producer price index – a measure of inflation on the wholesale level – accelerated in September. (CONTINGENCY.CAUSE.REASON)

Free to-infinitives

- Banks need a competitive edge **Implicit=if (they are)** to sell their products. (CONTINGENCY.CONDITION.ARG2-AS-CONDITION)

New Relations

PDTB-2 guidelines precluded various intra-sentential relations

- Explicit relations lexicalized by discourse connectives, and implicit relations between paragraph-internal adjacent sentences and between (semi-) colon separated clauses within sentences.
- Discourse connectives drawn from the pre-defined syntactic classes
- Arguments realized as one or more clauses or sentences.

Precluded relations triggered by **subordinators** like **for, by, instead of**, etc., that can take clausal complements.

- But **with** foreign companies snapping up U.S. movie studios, the networks are pressing their fight harder than ever. (CONTINGENCY.CAUSE.REASON)
- But on reflection, Mr. Oka says, he concluded that Nissan is being prudent in following its slow-startup strategy **instead of** simply copying Lexus. (EXPANSION.SUBSTITUTION.ARG1-AS-SUBST)

Consequence of including subordinators in PDTB-3

Connective	Conns/All Tokens	Most frequent senses
by	435/5194	Arg2-as-manner, Arg2-as-manner/Arg1-as-goal, Arg2-as-manner/Reason, Arg2-as-manner/Arg2-as-cond
with	299/4927 (6%)	Arg2-as-detail, Reason, Conjunction
without	94/339	Arg2-as-manner, Arg2-as-denier, Arg2-as-negCond
in	70/17764 (0.3%)	Synchronous, Arg2-as-detail, Arg2-as-manner
for	60/9029 (0.6%)	Reason, Arg2-as-goal, Arg2-as-cond
in order	55/59	Arg2-as-goal, Arg2-as-cond
instead of	43/99	Arg1-as-subst
rather than	40/152	Arg1-as-subst
just as	34/57 (60%)	Similarity (19), Synchronous (15)
so that	31/37	Arg2-as-goal (21), Result (10)

New Relations

PDTB-2 guidelines precluded various intra-sentential relations

- Explicit relations lexicalized by discourse connectives, and implicit relations between paragraph-internal adjacent sentences and between (semi-) colon separated clauses within sentences.
- Discourse connectives drawn from the pre-defined syntactic classes
- Arguments realized as one or more clauses or sentences.

Precluded relations between **conjoined verb phrases** (Webber et al., 2016).

Exceptions allowed VPs to be arguments of connectives

- She became an abortionist accidentally, and continued **because** it enabled her to buy jam, cocoa and other war rationed goodies. (CONTINGENCY.CAUSE.REASON)

but not arguments of the VP conjunction.

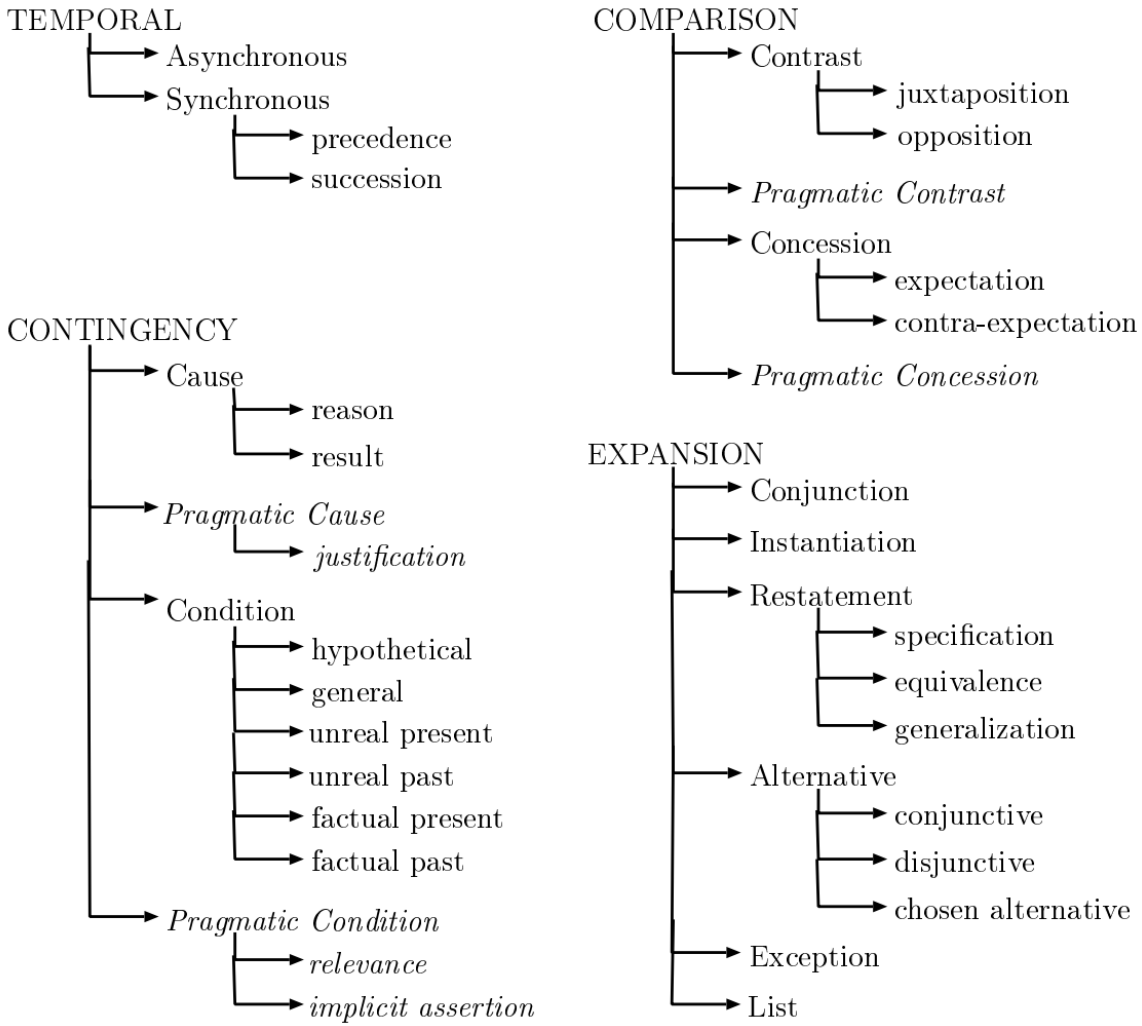
- She became an abortionist accidentally, **and** continued because it enabled her to buy jam, cocoa and other war rationed goodies. (EXPANSION.CONJUNCTION)
- Stocks closed higher in Hong Kong, Manila, Singapore, Sydney and Wellington, **but** were lower in Seoul. (COMPARISON.CONTRAST)

New PDTB-3 Relations: Distribution

RelType	Intra-S PDTB-2	Intra-S PDTB-3	Diff	Inter-S PDTB-2	Inter-S PDTB-3	Diff
Explicit	11209	16908	5699	7243	7332	89
Implicit	531	6234	5703	15516	15593	77
AltLex	–	785	785	624	713	89
AltLexC	–	134	134	–	6	6
EntRel	53	358	305	5155	5177	22
Hypophora	–	4	4	–	142	142
NoRel	–	–	–	252	286	34
Total	11794	24416	–	28789	29242	–

Total annotated relations: 53676

PDTB2 Sense Hierarchy



PDTB3 Sense Hierarchy

Temporal	Synchronous	--
	Asynchronous	Precedence
		Succession

Contingency	Cause +/-β, +/-ξ	Reason
		Result
		Negative-result*
	Condition +/-ξ	Arg1-as-cond
		Arg2-as-cond
	Negative condition +/-ξ	Arg1-as-negcond
		Arg2-as-negcond
	Purpose	Arg1-as-goal
		Arg2-as-goal
		Arg2-as-negGoal

Comparison	Contrast	--
	Similarity	--
	Concession +/-ξ	Arg1-as-denier*
		Arg2-as-denier

Expansion	Conjunction	--
	Disjunction	--
	Equivalence	--
	Instantiation	Arg1-as-instance
		Arg2-as-instance
	Level-of-detail	Arg1-as-detail
		Arg2-as-detail
	Substitution	Arg1-as-subst
		Arg2-as-subst
	Exception	Arg1-as-excpt
		Arg2-as-excpt
	Manner	Arg1-as-manner
		Arg2-as-manner

Simplifications: senses at Level-3 now only encode directionality of the arguments, and so only appear with **asymmetric** Level-2 senses

PDTB3 Sense Hierarchy

Temporal	Synchronous	--
	Asynchronous	Precedence
		Succession

Contingency	Cause +/-β, +/-ξ	Reason
		Result
		Negative-result*
	Condition +/-ξ	Arg1-as-cond
		Arg2-as-cond
	Negative condition +/-ξ	Arg1-as-negcond
		Arg2-as-negcond
	Purpose	Arg1-as-goal
		Arg2-as-goal
		Arg2-as-negGoal

Comparison	Contrast	--
	Similarity	--
	Concession +/-ξ	Arg1-as-denier*
		Arg2-as-denier

Expansion	Conjunction	--
	Disjunction	--
	Equivalence	--
	Instantiation	Arg1-as-instance
		Arg2-as-instance
	Level-of-detail	Arg1-as-detail
		Arg2-as-detail
	Substitution	Arg1-as-subst
		Arg2-as-subst
	Exception	Arg1-as-excpt
		Arg2-as-excpt
	Manner	Arg1-as-manner
		Arg2-as-manner

Simplifications: Some (asymmetric) Level-2 senses were discovered to permit arguments in either order, rather than the single order assumed in the PDTB2.

ARG1-AS-COND: *Call Jim Wright's office in downtown Fort Worth, Texas, these days* and **the receptionist still answers the phone, "Speaker Wright's office.** [wsj_0909]

ARG1-AS-SUBST: "The primary purpose of a railing is *to contain a vehicle* **and** not **to provide a scenic view**," [wsj_0102]

ARG1-AS-EXCPT: *Twenty-five years ago the poet Richard Wilbur modernized this 17th-century comedy merely by avoiding "the zounds sort of thing," as he wrote in his introduction.* Otherwise, **the scene remained Celimene's house in 1666**. [wsj_1936]

ARG1-AS-INSTANCE: *In a country where a bribe is needed to get a phone, a job, and even into a school, the name Bofors has become a potent rallying cry against the government.* **That illustrates the kind of disappointment many Indians feel toward Mr. Gandhi, whom they zestfully elected and enthusiastically supported in his first two years in power**. [wsj_2041]

"That illustrates" is called an Alternative Lexicalization (**AltLex**). It conveys the sense of the relation while not being an explicit discourse connective.

PDTB3 Sense Hierarchy

Temporal	Synchronous	--
	Asynchronous	Precedence
		Succession

Contingency	Cause +/-β, +/-ξ	Reason
		Result
		Negative-result*
	Condition +/-ξ	Arg1-as-cond
		Arg2-as-cond
	Negative condition +/-ξ	Arg1-as-negcond
		Arg2-as-negcond
	Purpose	Arg1-as-goal
		Arg2-as-goal
		Arg2-as-negGoal

Comparison	Contrast	--
	Similarity	--
	Concession +/-ξ	Arg1-as-denier*
		Arg2-as-denier

Expansion	Conjunction	--
	Disjunction	--
	Equivalence	--
	Instantiation	Arg1-as-instance
		Arg2-as-instance
	Level-of-detail	Arg1-as-detail
		Arg2-as-detail
	Substitution	Arg1-as-subst
		Arg2-as-subst
	Exception	Arg1-as-excpt
		Arg2-as-excpt
	Manner	Arg1-as-manner
		Arg2-as-manner

Simplification: Level-2 pragmatic senses have been replaced with features that can be attached to a relation token to indicate an inference of implicit belief or of a speech act associated with arguments.

PDTB3 Sense Hierarchy

Temporal	Synchronous	--
	Asynchronous	Precedence
		Succession

Contingency	Cause +/-β, +/-ξ	Reason
		Result
		Negative-result*
	Condition +/-ξ	Arg1-as-cond
		Arg2-as-cond
	Negative condition +/-ξ	Arg1-as-negcond
		Arg2-as-negcond
	Purpose	Arg1-as-goal
		Arg2-as-goal
		Arg2-as-negGoal

Comparison	Contrast	--
	Similarity	--
	Concession +/-ξ	Arg1-as-denier*
		Arg2-as-denier

Expansion	Conjunction	--
	Disjunction	--
	Equivalence	--
	Instantiation	Arg1-as-instance
		Arg2-as-instance
	Level-of-detail	Arg1-as-detail
		Arg2-as-detail
	Substitution	Arg1-as-subst
		Arg2-as-subst
	Exception	Arg1-as-excpt
		Arg2-as-excpt
	Manner	Arg1-as-manner
		Arg2-as-manner

Additional senses were introduced to annotate Intra-S relations, where they hadn't been noticed as needed for Inter-S relations.

SIMILARITY: One or more similarities between *Arg1* and **Arg2** are highlighted with respect to what each argument predicates as a whole or to some entities it mentions.

. . ., *the Straits Times index is up 24% this year*, so investors who bailed out generally did so profitably. Similarly, **Kuala Lumpur's composite index yesterday ended 27.5% above its 1988 close**. [wsj_2230]

CAUSE:NEGATIVE RESULT: *Arg1* gives the reason/explanation/justification for why **Arg2** does not result.

A search party soon found the unscathed aircraft in a forest clearing much too small **to have allowed a conventional landing**.

MANNER: The situation described by one argument presents *how* (i.e., the manner in which) the situation described by other argument has happened or is done.

ARG1-AS-MANNER: He argued that program-trading by roughly 15 big institutions is *pushing around the markets* Implicit=thereby **and scaring individual investors**. [wsj_0987]

ARG2-AS-MANNER: A native of the area, he is back now after riding the oil-field boom to the top, *then surviving the bust* Implicit=by **running an Oklahoma City convenience store**. [wsj_0725]

NEGATIVE CONDITION: One argument describes a situation presented as unrealized (the antecedent or condition), which if it doesn't occur, would lead to the situation described by the other argument (the consequent).

ARG1-AS-NEGCOND: In Singapore, a new law requires smokers to *put out their cigarettes before entering restaurants, department stores and sports centers* or **face a $250 fine**. [wsj_0037]

ARG2-AS-NEGCOND: Unless **the Federal Reserve eases interest rates soon to stimulate the economy**, *profits could remain disappointing*. [wsj_0322]

PURPOSE: One argument presents an action that an agent undertakes with the purpose (intention) of achieving the goal conveyed by the other argument.

ARG1-AS-GOAL: She ordered *the foyer done in a different plaid planting*, Implicit=for that purpose **and made the landscape architects study a book on tartans**. [wsj_0984]

ARG2-AS-GOAL: *Skilled ringers use their wrists to advance or retard the next swing*, so that **one bell can swap places with another in the following change**. [wsj_0089]

ARG2-AS-NEGGOAL: We can applaud Mr. Pryor's moment of epiphany, even as we understand *that he and his confreres need restraint* lest **they kill again**. [wsj_1698]

PDTB-3 Adopted Syntax-based Argument Labeling

More fine-grained reference to syntactic structure, regardless of realization type, in order to avoid inconsistencies, while not requiring any change to existing labels in PDTB-2.

- Arguments to **inter-sentential discourse relations** remain labeled by position: Arg1 is first (lefthand) argument and Arg2, the second (righthand) argument.

- Arguments of **intra-sentential coordinating structures** are also labeled by position: Arg1 is the first conjunct and Arg2, the second conjunct.

- Arguments of **intra-sentential subordinating structures** are determined syntactically. The subordinate structure is always labeled Arg2, and the structure to which it is subordinate is labeled Arg1.

Extensions to AltLex Identification

AltLex: In the absence of an explicit connective, annotators who inferred a relation between sentences but felt that it would be redundant to insert an implicit connective, were asked to identify as the **AltLex,** whatever **non-connective expression in Arg2** they saw as the source of the redundancy.

(1) In PDTB3, AltLex can include material from both Arg1 and Arg2.

- Some of the proposals are **so close** **that** non-financial issues such as timing may play a more important role.
  (CONTINGENCY.CAUSE.RESULT)

- Things have gone **too far** **for** the government to stop them now.
  (CONTINGENCY.CAUSE.RESULT)

Extensions to AltLex Identification

AltLex: In the absence of an explicit connective, annotators who inferred a relation between sentences but felt that it would be redundant to insert an implicit connective, were asked to identify as the **AltLex,** whatever **non-connective expression in Arg2** they saw as the source of the redundancy.

(2) In PDTB3, syntactic constructions can serve as AltLex: *AltLexC*.

- Crude as they were, these early PCs triggered explosive product development in desktop models for the home and office.
  (COMPARISON.CONCESSION.ARG1-AS-DENIER)

  Predicate Inversion

- Had the contest gone a full seven games, ABC could have reaped an extra $10 million in ad sales on the seventh game alone, compared with the ad take it would have received for regular prime-time shows.
  (CONTINGENCY.CONDITION.ARG2-AS-CONDITION)

  AUX Inversion

Multiple relations between discourse arguments

If there is >1 discourse connective, there may be >1 discourse relation:

- It's too far to walk. So instead let's take the bus.

But the same meaning can be conveyed without multiple discourse connectives:

- It's too far to walk. So let's take the bus.
- It's too far to walk. Instead let's take the bus.
- It's too far to walk. Let's take the bus.

Even if a discourse relation is explicitly cued, additional **implicit relations** may be inferred.

Multiple relations between discourse arguments

- We can refer to Implicit discourse relations, AltLex and AltLexC relations, and Entity relations as **Non-explicit relations**.

- Non-explicit relations within a sentence can either
  - Stand-alone
  - Co-occur with an Explicit relation

- Non-explicit relations that share arguments with explicit relations can be said to be **linked** to those relations.

Implicit Relations Linked to Explicit Connectives

- CONJUNCTION (and) with Implicit RESULT
  . . . opponents argued that the increase will still hurt small business and cost many thousands of jobs [wsj 0098]

- CONTRAST (but) with Implicit ARG2-AS-SUBST
  Volatility surrounding his trades occurs not because of index arbitrage, but because his is a large addition or subtraction to a widget market with finite liquidity [wsj 0118]

- ARG2-AS-GOAL (in order) with Implicit ARG1-AS-MANNER
  Government officials tried throughout the weekend to render a business-as-usual appearance in order to avoid any sense of panic [wsj 2413]

AltLex Relations Linked to Explicit Connectives

- CONCESSION.ARG1-AS-DENIER (though) with AltLex LEVEL-OF-DETAIL.ARG1-AS-DETAIL (in general)
  The Pentagon's recently issued Soviet Military Power repeated the Sverdlovsk assessment, though in general adopting a softer line [wsj_1143]

- CAUSE.RESULT (thus) with AltLex CAUSE.RESULT (giving)
  Under the guise of "healing the wounds of the nation" President Carter pardoned thousands of draft evaders, thus giving dignity to their allegations of the war's "immorality" [wsj_0290]

- COMPARISON.SIMILARITY (as well) with AltLex CAUSE.RESULT (giving)
  The space shuttle Altantis boosted the Galileo spacecraft on its way to Jupiter, giving a big lift as well to an ambitious U.S. program of space exploration [wsj_1817]

AltLexC Relations Linked to Explicit Connectives

- CONTRAST (but) with ALTLEXC SIMILARITY
  All independent media activity is now illegal, which perhaps is not surprising, but so is the manufacture of perfume, cosmetics, household chemicals and sand candles [wsj_0439]

- ARG2-AS-DENIER (but) with AltLexC SIMILARITY
  Hassan comes to a bad end, but so does almost everyone else in the book [wsj_0790]

AltLexC Relations Linked to Explicit Connectives

- CONTRAST (but) with ALTLEXC SIMILARITY
  All independent media activity is now illegal, which perhaps is not surprising, but so is the manufacture of perfume, cosmetics, household chemicals and sand candles [wsj_0439]

- ARG2-AS-DENIER (but) with AltLexC SIMILARITY
  Hassan comes to a bad end, but so does almost everyone else in the book [wsj_0790]

Syntax-based vs. Discourse-based splitting

- Syntax-based splitting: licensed by syntactic tree-structure and/or by argument roles.
- Discourse-based splitting: licensed by presence of low-level discourse relations within a sentence.
- What I've shown is that the PDTB-3 makes heavy use of syntax in recognizing Intra-S discourse relations so the difference is less distinct.
- Suggest that an interesting source of rephrasing splits is AltLex patterns.

Syntax-based vs. Discourse-based splitting

- Quickly suggest that an interesting source of rephrasing splits is AltLex patterns – e.g.
  - This/that means/meant: 27 tokens
  - This/that <be> because: 23 tokens
  - The result <be>: 11 tokens
  - This/that compares/compared with: 10 tokens
  - This/that <be> why: 9 tokens
  - The reason <be>: 8 tokens

Conclusion

There is clearly more to say and do about using discourse relations for simplification.

So thanks to the organizers for the opportunity to start thinking about it.

References

- Cripwell, L., Legrand, J. & Gardent, C. (2021). Discourse-based Sentence Splitting. **Findings of the ACL** (EMNLP), pp. 261-293.

  Cripwell, L., Legrand, J. & Gardent, C. (2022). Controllable Sentence Simplification via Operation Classification. **Findings of the ACL** (NAACL), pp. 2091-2103.