The Biomedical Discourse Relation Bank (BioDRB) Annotation Guidelines

Rashmi Prasad

Institute for Research in Cognitive Science, University of Pennnsylvania
rjprasad@seas.upenn.edu

Susan Mcroy, Nadya Frid, Hong Yu

University of Wisconsin-Milwaukee
mcroy@tigger.cs.uwm.edu, nadyafrid@yahoo.com, hongyu@uwm.edu

April 28, 2010

Contents

1   Introduction

In the BioDRB (Biomedical Discourse Relation Bank), we are annotating a subset of the biomedical articles in the GENIA corpus with *discourse relations*. The BioDRB annotation guidelines are an adaptation of the PDTB (Penn Discourse Treebank Prasad *et al.* (2008)) guidelines PDTB-Group (2008)).

1.1   Terms and Definitions

The key terms and definitions associated with discourse relations are as follows. They will be followed closely in the annotation.

- Discourse relations are binary relations, which denote *senses* (meanings) such as 'cause', 'contrast', and 'temporal sequence'.

- Discourse relations are *binary* relations such that each relation holds between two and only objects.

- The objects related by discourse relations are *abstract objects* (AOs), such as events, states, facts and propositions, and they are referred to as the relation's *arguments*, named Arg1 and Arg2.

- Discourse relations can be *explicit*, i.e., expressed in the text with lexical items (words or phrases), but they can also be *implicit* such that the reader must infer them from the prior context in the text.

- Discourse relations are realized explicitly as either *connectives*, which are lexical items from well-defined syntactic classes (subordinating conjunctions, coordinating conjunctions, adverbs, subordinators), or *alternative lexicalizations (AltLex)*, which are other expressions that cannot be classified as connectives.

- When relations are implicit, it should be possible to *insert a connective* to express the inferred relation (except for relations which are always unrealized as connectives, e.g., continuation, background, and circumstance). When insertion of a connective leads to the connective sounding *redundant*, this is due to the presence of an AltLex expression that realizes the relation.

- The Arg1 and Arg2 arguments of a discourse relation are named on the basis of *syntactic and linear ordering conventions*. In the case of explicit connectives and AltLex's, Arg2 is the argument to which the relation expression is syntactically associated, whereas in the case of implicit relations, Arg2 is second argument in the linear sequence of the two arguments.

Based on the definitions above, there are primarily three ways in which discourse relations may be realized: (i) as *explicit connectives*, (ii) as *AltLex*, and (iii) as *implicit relations*. These types represent a high-level classification of discourse relations.

A fourth type called *NoRel* stands for cases where a sentence seems unrelated to anything else in the text. We only used NoRel in two specific cases. The first kind of NoRel is annotated within the "Abstract" section of the articles, some of which are partitioned into "Background", "Case Presentation", "Results", "Conclusion", etc. These "Abstract" sections are not separated by any paragraph boundary, but we treat them as such and indicate these boundaries with the NoRel label. The second kind of NoRel was annotated for typological errors that led, for example, to some sentences being duplicated in the article. Since we didn't want to admit a non-semantic repetition relation, these were annotated as NoRel. Such cases are rare.

1.2   Examples of Types of Discourse Relations

The following examples illustrate the three major discourse relation types from the corpus. In all examples in this document, the explicit connective/AltLex/implicit connective is underlined, and the sense of the relation is shown in parentheses at the end of the example. Arg2 is shown in bold, and Arg1 is shown in italics.

In Ex. (1), the explicit connective *while* expresses contrasting observations with respect to IL-2 production in RA patients and in healthy controls, in the presence of 1ng/ml IL-10. The connective is a subordinating conjunction which is syntactically associated with the second subordinate clause labelled as Arg2. The main clause is labelled as Arg1. The sense label assigned to the relation is 'contrast'.

(1)   *Similarly, IL-2 production was not affected by 1 ng/ml IL-10 in RA patients (percentage decrease, -2.1 ± 13.8%)* <u>while</u> **it was significantly reduced in healthy controls (61.1 ± 13.7%; P < 0.05).** (Contrast)

In Ex. (2), what we learn from the explicit connective *As a consequence* is that the possibility of co-expressed gene groups escaping detection is a result of the fact that the tissues in the dataset represent samples of limited morphological resolution at a fixed time point. The connective is an adverbial which is syntactically associated with the second sentence labelled as Arg2. The prior sentence is labelled as Arg1. The sense label assigned to the relation is 'cause:result'.

(2)   *The tissues in the dataset represent samples of limited morphological resolution at a fixed time point.* <u>As a consequence,</u> **co-expressed gene groups that are active under specific developmental phases, under specific environmental stresses, or in small and localized anatomical structures are likely to escape detection.** (Cause:Result)

In Ex. (3), the claim in second sentence is presented as a conclusion that the authors have arrived at on the basis of the evidence presented in the previous sentence. This CLAIM relation is expressed explicitly in the second sentence not with a connective, but an AltLex expression *These results suggest that*, which is composed of a subject and a verb. The Arg2 for the AltLex is the second sentence within which the AltLex occurs, and Arg1 is the other previous sentence. Note that unlike explicit connectives, AltLex's spans are contained within the Arg2 spans. This is reflected in the simultaneous boldface and underlining of the AltLex.

(3)　*As shown in Fig. 3a,3b, the intensity of IL-10R1 expression on CD4+ T cells was significantly increased in RA patients compared with in healthy controls.* **These results suggest that the intracellular signal transduction pathway of IL-10 may be impaired in CD4+ T cells of active RA.** (Cause:Claim)

Finally, Ex. (4) shows that we can infer a temporal relation between the events expressed by the two sentences, although there is no explicit connective or AltLex present to express this relation. It is possible to insert a temporal connective like *then* to express the relation. In implicit relations such as this, Arg2 is the second sentence in the sequence while Arg1 is the first sentence. The temporal ordering of the events is one of *precedence*, indicating that the Arg1 event occurs before the Arg2 event.

(4)　*Motifs represented as position frequency matrices (PFMs) were downloaded from the TRANS-FAC [55] and JASPAR [25] databases.* Implicit=then **Non-vertebrate matrices were filtered out.** (Temporal:Precedence)

2　Identifying Connectives

Discourse connectives in BioDRB are considered to belong to four broad grammatical classes: (a) subordinating conjunctions (e.g., *because, when, although*), (b) coordinating conjunctions (e.g., *and, or, but*), (c) adverbials and prepositional phrases (e.g., *however, as a result, then*), and (d) subordinators (e.g., *by, to, on, for*).

The primary criterion for identifying discourse connectives follows from the definition of discourse relations. That is, connectives should denote discourse relations that relate two and only two AOs. Thus, adverbials that do not denote relations between two AOs should not been annotated as discourse connectives. For example, adverbials called "cue phrases" or "discourse markers" such as *of course, anyway, now*, etc., have not been annotated since they serve to signal the organizational or focus structure of the discourse, rather than relate AOs. And clausal adverbials such as *notably, interestingly, in all likelihood* etc. are also not annotated as discourse connectives since they take a single AO as argument, rather than two.

Not all tokens of words and phrases that can serve as explicit connectives actually do so: Some tokens serve other functions, such as to relate non-AO entities (e.g., the use of *and* to conjoin noun phrases in Ex. (5) or the use of *also* to signal presupposed information in Ex. (6). Such expressions are not annotated as discourse connectives.

(5) Sera from RA patients induced signal transducer **and** activator of transcription 3 phosphorylation in normal CD4+ T cells, which was mostly abolished by neutralizing anti-IL-6 antibody.

(6) To determine whether CD123+ cells in synovial tissue were **also** nuclear RelB+, formalin-fixed tissue was double-stained for RelB and CD123 without hematoxylin counterstaining.

Following the PDTB, we do not annotate coordinating conjunctions that conjoin verb phrases. However, when a discourse connective that is not a conjunction, such as then or thus, occurs between two verb phrases, we do annotate the relation. Thus, the VP conjunction *and* is annotated in Ex. (7) but not in Ex. (8).

(7) *Critical technical parameters were evaluated,* <u>and</u> **the results and the methods provide a valuable resource for future experimental design**. (CONJUNCTION)

(8) A key step in the elucidation of gene battery-like regulation is to detect **and** functionally test the cis regulatory elements that mediate the co-regulation.

3 Arguments of Discourse Relations

The smallest syntactic unit for the realization of an AO argument of a discourse relation is a clause, tensed (Ex. 9) or non-tensed (Ex. 10). Verb phrases can also be legal arguments when the connectives are not verb phrase conjunctions themselves (Ex. 11).

(9) *A receptor that binds several CXC and CC chemokines, vGCR appears to be constitutively active [33],* <u>although</u> **some studies indicate that it is expressed only during lytic replication [34]**. (Contrast)

(10) *These beacons generate a fluorescent signal during amplification* <u>when</u> **hybridising with their target**. (Temporal:Synchronous)

(11)

(12) Though DNA is an obvious choice because HHV-8 is a DNA virus, we have chosen RNA, *because it indicates activity of the virus* **and may** <u>thus</u> **be a better prognostic marker**. (Cause:Claim)

In addition, because discourse relations hold between AOs, nominalizations are also allowed (Exs. 13-14) as arguments, since they denote AOs, such as events and states.

(13) *She was originally considered to be at high risk* <u>due to</u> **the familial occurrence of breast and other types of cancer**, (Cause:Reason)

(14) <u>Despite</u> **her status as a BRCA1 mutation carrier, and her mammographically dense breast tissue**, *we did not find increased cell proliferation or deficient differentiation potential in breast epithelial cells from this patient which might have contributed to her cancer susceptibility.* (Concession:Expectation)

There are no syntactic constraints on how many clauses or sentences an argument can extend over. An argument can be a single clause, or multiple clauses within the same sentence, or multiple sentences. Semantically, however, arguments are required to be *minimal* in that "only as much should be selected as an argument as is necessary for interpreting the relation". Ex. (15) shows Arg1 as well as Arg2 spanning over multiple sentences for the AltLex generalization relation. However, for both Arg1 and Arg2, all the included sentences are necessary and sufficient because for the generalization relation in question, the specific details as well as the generalization of the details are distributed across exactly these multiple sentences.

(15) *We show here that mice lacking ITK have much reduced IL-2 production and T cell expansion in response to SEB in vitro and in vivo. We also show that SEB induced the activation of the JNK MAPK pathway in responding T cells in vivo, and that ITK null T cells were defective in the activation of this pathway in vivo. However, toxicity analysis indicated that both WT and ITK null animals were similarly affected by SEB exposure.* <u>**Our data suggest that ITK is required for full IL-2 secretion following SEB exposure, and that this may be due to the regulation of the JNK pathway by ITK in vivo. However, reducing T cell signals does not necessarily lead to better physiological responses to SEB exposure**</u>. (Restatement:Generalization)

Finally, except for NoRel (see below), there are also no constraints on how far away a relation's Arg1 and Arg2 arguments can be from each other. That is, they need not be adjacent. Ex. (16) shows Arg1 and Arg2 in non-adjacent sentences for the explicit connective *However*.

(16) The studies concerning the functional interaction between the NF-κB pathway and members of the steroid hormone receptor family, and their role in synovial inflammation, have advanced significantly, *although with controversial results [10,11]*. In particular, after binding with E2, oestrogen receptors have been shown to interact with NF-κB factors, via transcriptional co-factors, resulting in mutual or non-mutual antagonism. Other studies hypothesize that, since oestrogen receptors may repress both constitutive and inducible

NF-κB, the overexpression of NF-κB-inducible genes in oestrogen receptor-negative cells might contribute to malignant cell growth and chemotherapeutic resistance [12,13]. On the contrary, further studies report that E2 blocks the transcriptional activity of p65 in macrophages [14]. <u>However</u>, **these opposite observations arise using different cell lines (human/animals) and culture conditions as well as different hormone concentrations [15]. In addition, multiple mechanisms concerning the interaction between oestrogen receptors an NF-κB have been proposed, such as repression of NF-κB DNA binding by physical association with oestrogen receptors and the regulation of IκB-α expression by oestrogens [16,17].**

Arguments can also be non-adjacent for implicit and Altlex relation types, as shown in Exs. (17) and (18). This is a departure from the PDTB, where implicit and AltLex relation types are annotated only between adjacent sentences.

(17) *The precision of the assays was measured as the standard deviation and the confidence interval (fig. 2).* We used 5 replicates of 7 different concentrations of the standard curve. <u>Implicit=and</u> **Robustness was determined by having different people perform the experiments on different days**. (Conjunction)

(18) *However, to our surprise the two separated lobes from Hoxa3+/-Pax1-/- compound mutants did not always develop in predictable ways.* In fact, separated lobes cultured under identical conditions in the absence of any added cytokines had widely differing results. Changes in thymocyte cell number and differentiation after culture were highly variable; for example, cell numbers could differ stochastically after culture by as much as 10-fold between two lobes from the same compound mutant embryo. This result was never obtained with separately cultured thymic lobes from any other genotype, which always developed identically when cultured without cytokines. **<u>This property of the compound mutant thymic lobes probably reflects</u> phenotypic variability previously seen in our histological and gross morphological analysis of these mutants [29,34]**. (Cause:Reason)

For NoRel relation types, the two arguments are by default immediately adjacent and complete sentences. Ex. (19) illustrates a NoRel annotation from the "Abstract" section of an article.

(19) Background: CC Chemokine Receptor 3 (CCR3), the major chemokine receptor expressed on eosinophils, binds promiscuously to several ligands including eotaxins 1, 2, and 3. (...) *It is therefore important to elucidate the molecular mechanisms regulating receptor expression.* <u>Implicit=NoRel</u> **Results: In order to define regions responsible for CCR3 transcription, a DNAse hypersensitive site was identified in the vicinity of exon 1**.

3.1 Conventions for Attributions

The PDTB conventions specify that only the minimally necessary amount of content necessary to understand a discourse relation should be included in the arguments. Extra modifying information is annotated separately as supplements to the argument, or as an attribution (the definitions and examples of supplements and attribution are given below). In our effort, we do not annotate supplements or attributions separately and so we have developed some conventions for how modifiers should be included. We have also defined some conventions to help annotators deal with cases that may appear to be ambiguous.

Attribution is a relation of "ownership" between abstract objects and individuals or agents. That is, attribution has to do with ascribing beliefs and assertions expressed in text to the agent(s) holding or making them. Attribution can scope over the entire relation, i.e. over the discourse connective and its argument (Ex. 20) or over just one argument (Arg2 in Ex. 21). In the examples square brackets are used to mark expressions of attribution.

(20) Koehl (4) has written an excellent review of the various methods used to detect structural relationships. [He concluded that] <u>though</u> **significant progress has been made over the past decade**, *a fast, reliable and convergent method for protein structural alignment is not yet available.*

(21) *Heterologous anti-human lymphocyte or thymocyte antisera have been used successfully for immunosuppression in human organ transplantation* <u>and</u> [most researchers have concluded that] **they are effective by virtue of elimination of T cells by complement-mediated lysis (27, 28)**.

Although we do not annotate the attribution separately like PDTB does, we do exclude attributions from argument spans when they lie outside the scope of the relation, as shown by the annotations above. However, when part of an argument is modified by reference to a citation or a reference to a figure, we include the reference within the argument (Exs. 22-23).

(22) <u>After</u> **birth**, *in the vestibule, the strongest signal was associated with the saccular and utricular otoconia (Fig. 4D)*

(23) *Saccular and utricular maculae from seven adult mice were microdissected in artificial endolymph (21) under a stereomicroscope,* <u>and</u> **the otoconial complexes were harvested with a siliconized pulled Pasteur pipette**.

Passivized attributions are also never excluded. Attributions are also not excluded when they include some modality or negation that is needed for the interpretation of the argument. Finally, attribution scopes are easier to judge when the arguments of a relation as well as the attribution

are all in the same sentence, as compared to when arguments appear in different sentences. For the latter case, therefore, annotators were instructed to include the attributions when in doubt.

4   Sense Labeling

All explicit, implicit and AltLex relations are annotated with sense labels that indicate their semantics. These senses are organized into two tiers, with the second "subtype" tier specifying further refinements to the sense "types" in the top tier. The complete BioDRB sense classification is shown in Table 1.

For any relation, sense annotation consists of selection of a sense subtype label when this is available for a type. Thus, for the "Cause" sense, the annotator is required to select one of its four subtypes, i.e., the type level label cannot be chosen. When a sense does not have a subtype available, such as "Contrast", this type level label is selected.

Refinements at the subtype level are of two kinds: some specify refinements of the semantics, whereas others specify the directionality of the arguments. Thus, for example, the three subtypes of the "Condition" sense type specify in more detail the nature of the conditional dependence between Arg1 (antecedent) and Arg2 (consequence), by indicating whether the antecedent describes a hypothetical situation ("Hypothetical"), an assumed fact ("Factual"), or a non-fact ("Non-Factual"). On the other hand, for the sense type "Concession" - in which one argument creates an expectation denied by the other argument - the two subtypes indicate the directionality of the concession: In the "Contra-expectation" subtype, Arg1 raises the expectation which Arg2 denies, while in the "Expectation" subtype, Arg2 raises the expectation which Arg1 denies.

With some connectives, more than one sense can be inferred. Annotators are allowed to admit two senses for a connective. In Ex. (24), for instance, the connective *as* is annotated with a temporal as well as a causal sense.

(24)   Tumors detected by this new technology could have unique etiologies and/or presentations, *and may represent an increasing proportion of clinical practice* <u>as</u> **new screening methods are validated and applied**. (Temporal:Synchronous/Cause:Justification)

<u>Cause</u>

The sense "Cause" is used when the two arguments of the relation are related causally and are not in a conditional relation. "Reason" and "Result" hold when the situation described in one of the arguments is the cause of the situation described in the other argument, the difference simply being the directionality of the causality: "Reason" when Arg2 is the cause, and Arg1 the effect (Ex. 25) and "Result" when Arg1 is the cause and Arg2 the effect (Ex. 26). "Claim" and "Justification" on the other hand, hold when the situation described by one of the arguments

Type	Subtype	Type	Subtype
CAUSE	Reason Result Claim Justification	CONDITION	Hypothetical Factual Non-Factual
PURPOSE	Goal Enablement	TEMPORAL	Synchronous Precedence Succession
CONCESSION	Contra-Expecation Expectation	ALTERNATIVE	Chosen-Alternative Conjunctive Disjunctive
CONTRAST		INSTANTIATION	
CONJUNCTION		EXCEPTION	
SIMILARITY		CONTINUATION	
CIRCUMSTANCE	Forward-Circumstance Backward-Circumstance	BACKGROUND	Forward-Background Backward-Background
RESTATEMENT	Equivalence Generalization Specification	REINFORCEMENT	

Table 1: BioDRB Sense Classification for Discourse Relations

is the cause not for the situation described by the other argument, but rather for the truth or validity of the proposition described by the argument. The difference between the two is again one of directionality, with "Claim" annotated when Arg1 presents the evidence for the truth of Arg2 (Ex. 27), and "Justification" annotated when Arg2 presents the evidence for the truth of Arg1 (Ex. 28).

(25)   *A little over a decade ago, the primacy of T cells in the pathogenesis of autoimmune disease such as RA was undisputed* <u>because</u> **they are the largest cell population infiltrating the synovium**. (Cause:Reason)

(26)   *In addition, the B-1 cell compartment is absent in Xid-mice.* <u>Consequently</u>, **the levels of serum IgM and IgG3 are low**. (Cause:Result)

(27)   *Consistent with this notion, PB T cells from patients with allergic diseases significantly express high levels of SOCS3 transcripts, and the SOCS3 expression correlates well with serum IgE levels and disease pathology [45].* **Higher SOCS1 expression with lower SOCS3 expression in PB CD4+ T cells from RA patients, compared with healthy controls, is** <u>therefore</u> **probably consistent with their systemic bias towards a Th1 phenotype, as has previously been demonstrated [46-49]**. (Cause:Claim)

(28)   *This direct effect also may be limited in active RA patients*, <u>because</u> **their PB CD4+ T cells showed a defective IL-10 inhibition of CD28-costimulated production of both IFN-γ and IL-2**. (Cause:Justification)

<u>Condition</u>

The sense type "Condition" is used to describe all subtypes of conditional relations. In addition to causal influence, "Condition" allows some basic inferences about the semantic contribution of the arguments. Specifically, the situation in Arg2 is taken to be the condition and the situation described in Arg1 is taken to be the consequence, i.e., the situation that holds when the condition is true. The subtype "Hypothetical" (Ex. 29) is labelled when if Arg2 holds true, Arg1 is caused to hold at some instant in all possible futures. However, Arg1 can be true in the future independently of Arg2. The subtype "Factual" (Ex. 30) is a special case of the subtype "Hypothetical", and applies when the connective indicates that Arg2 is a situation that has either been presented as a fact in the prior discourse or is believed by somebody other than the speaker/writer. The subtype "NonFactual" applies when the connective indicates that Arg2 describes a condition that either does not hold at present or did not hold in the past. Arg1 then describes what would also hold if Arg2 were true. There were no occurrences of the Non-Factual conditionals in the corpus.

(29)   <u>If</u> *the nuclei over the foreskin fibroblasts averaged 50 or more grains per nucleus*, <u>then</u> **the rest of the experimental slides were developed**. (Condition:Hypothetical)

(30) <u>If</u> **Btk is a direct target of BOB.1/OBF.1**, *the phenotype of mice deficient in either BOB.1/OBF.1 or Btk should be related.* (Condition:Factual)

The main difference between "Cause" and "Condition" has to do with how truth values are calculated. In the case of "Cause", unless the connective and its arguments are embedded in a matrix that alters their truth value, the situations denoted by Arg1 and Arg2 and the causal relation between them are all taken to hold. In the case of "Condition", however, the truth value of the arguments cannot be determined independently of the connective.

Purpose

The sense type "Purpose" is labelled for relations in which one argument presents a situation and the other argument presents an action, and the engagement of the action enables the situation to occur. The subtypes "Goal" and "Enablement" capture directionality effects: "Goal" applies when Arg1 presents the action that enables the situation in Arg2 to obtain (Ex. 31-32), whereas "Enablement" applied when Arg2 presents action that enables the situation in Arg1 to obtain (Ex. 33.

(31) <u>In order to</u> **define the proteins capable of binding to CCR3 exon 1**, *the exon 1 sequence was analyzed using the publicly available TFSEARCH engine.* (Purpose:Goal)

(32) *We therefore applied the functional unscheduled DNA synthesis (UDS) assay for NER capacity to multiple samples of normal tissue from this patient*, <u>to</u> **determine whether haploinsufficiency for BRCA1 was a mechanism of NER deficiency**. (Purpose:Goal)

(33) *The important role of MZ B cells in this process was established* <u>by</u> **monitoring IgM-IC deposition onto FDCs in mice that lacked MZ B cells**, (Purpose:Enablement)

Temporal

The sense type "Temporal" is used when the connective indicates that the events described in the arguments are related temporally. The subtypes reflect the ordering of the arguments. "Precedence" is used when the connective indicates that the Arg1 event precedes the Arg2 event (Ex. 34), "Succession" applies when the Arg1 event follows the Arg2 event (Ex. 35), and "Synchronous" applies when the Arg1 and Arg2 events overlap (Ex. 36) – without specifying the form of overlap, i.e., whether the two situations started and ended at the same time, whether one was temporally embedded in the other, or whether the two crossed.

(34) *For proliferation assays, 1 x 105 purified total or CD4+ spleen T cells per well of 96-well round-bottom plates were incubated with 1 x 104 DC per well and 1 µg/ml ConA (Sigma) for 72 hours.* <u>Then</u> **1 µCi/well 3H-thymidine (Amersham-Buchler, Braunschweig, Germany) was added**. (Temporal:Precedence)

(35)   After bacterial amplification of the cloned PCR fragments by standard procedures, at least three clones from each sample were sequenced with an automated sequencer (ABI Prism 377, Applied Bio-systems, Foster City, USA) as recommended by the manufacturer. (Temporal:Succession)

(36)   *In this system, breast epithelial cells initially retain cell-to-cell contact* <u>while</u> **they proliferate**, (Temporal:Synchronous)

Concession

The sense type "Concession" applies when the connective indicates that one of the arguments describes a situation A which **causes** C, while the other asserts (or implies) ¬C. Two "Concession" subtypes are defined in terms of the argument creating an expectation and the one denying it. Specifically, when Arg2 creates an expectation that Arg1 denies, it is tagged as "expectation" (Exs. 39-40). When Arg1 creates an expectation that Arg2 denies, it is tagged as "contra-expectation" (Exs. 37-38).

(37)   *A reduction in breast cancer mortality has been observed in recent years that has been partially attributed to the widespread adoption of screening mammography [1].* **Traditional screening mammography**, <u>however</u>, **fails to detect 15% of incident cancers [2]**. (Concession:Contra-expectation)

(38)   *A reduction in breast cancer mortality has been observed in recent years that has been partially attributed to the widespread adoption of screening mammography [1].* **Traditional screening mammography**, <u>however</u>, **fails to detect 15% of incident cancers [2]**. (Concession:Contra-Expectation)

(39)   <u>Despite</u> **her status as a BRCA1 mutation carrier, and her mammographically dense breast tissue**, *we did not find increased cell proliferation or deficient differentiation potential in breast epithelial cells from this patient which might have contributed to her cancer susceptibility.* (Concession:Expectation)

(40)   <u>Although</u> **the NER values of these two samples from the same patient are similar**, *they are not close enough to distinguish themselves as coming from the same individual (P = 0.16).* (Concession:Expectation)

Contrast

The sense type "Contrast" is used when the relation is intended to indicate that the values for some shared property in Arg1 and Arg1 are in opposition to each other. These oppositions need not be at opposite ends of a graded scale and are instead highly context-dependent, relying on the intention of the writer. While Ex. (41) shows strict opposition with the contrast between "little" and "high levels" of TNF-alpha secretion, the opposition in Ex. (42) between "being constitutively

active" and "being expressed only during lytic replication" seems less strict. There are no subtypes for this relation sense.

(41) <u>Whereas</u> **stimulation of mixed lymphocyte reactions either by CD11c+ or by CD123+ PB DCs resulted in little TNF-α production**, *stimulation by either of these DCs from RA SF resulted in high levels of TNF-α secretion (Fig. 5).* (Contrast)

(42) *A receptor that binds several CXC and CC chemokines, vGCR appears to be constitutively active [33],* <u>although</u> **some studies indicate that it is expressed only during lytic replication [34].** (Contrast)

Similarity

The "Similarity" sense type is much like "Contrast" in that there is some shared property of Arg1 and Arg2 and the writer intends to convey that the compared values are similar to each other rather than in opposition. Ex. (43) illustrates the use of the connective *similarly* to express this relation.

(43) *The BMP-6 effects could be reversed by addition of the extracellular inhibitor Noggin (Figure 1D).* <u>Similarly,</u> **a combination of the soluble BMP receptors BMP-RIB-Fc and BMP-RII-Fc also neutralized the effects of BMP-6 (data not shown).** (Similarity).

Alternative

The sense type "Alternative" applies when the connective indicates that its two arguments denote alternative situations. It has three subtypes: "conjunctive" , "disjunctive" and "chosen alternative".

The "conjunctive" subtype (Ex. 44) is used when the connective indicates that both alternatives hold or are possible.

(44) The fact that the requirements for individual proteins, such as p50, could change with age suggests *that compensatory mechanisms exist which can overcome cellular defects,* <u>and/or</u> **that more than one pathway to generate these cells may exist**. (Alternative:Conjunctive)

The "disjunctive" subtype (Ex. 45) is used when two situations are evoked in the discourse but only one of them holds.

(45) <u>Either</u> *purified total T cells were directly used for proliferation assays,* <u>or</u> **CD4+ T cells were first negatively purified by adding 5 μg/1x107 cells OX8 mAb for 30 min on ice and, after washing, 75μl Dynabeads-M450 coupled to goat anti-mouse IgG (Dynal, Hamburg, Germany)**. (Alternative:Disjunctive)

The "chosen alternative" subtype (Ex. 46) is used when the connective indicates that two alternatives are evoked in the discourse but only one is taken.

(46)   *In the present study, we were not able to detect BMP-6-induced changes in the phosphory-*
       lation status of STAT3 or p38 in human peripheral B cells. <u>Instead</u>, **a rapid and marked**
       phosphorylation of Smad1/5/8 was revealed. (Alternative:Chosen alternative)

Instantiation

The sense "Instantiation" (Ex. 47) is used when the connective indicates that Arg1 evokes a set
and Arg2 describes it in further detail. It may be a set of events, a set of reasons, or a generic
set of events, behaviors, attitudes, etc. Typical connectives often tagged as "Instantiation" are *for*
example, *for instance* and *specifically*. There are no subtypes for this sense.

(47)   *A complex interaction between cells in inflamed RA joints might produce a variety of*
       proinflammatory cytokines and chemokines, which also activated other cells in the joints.
       <u>For example</u>, **IL-17 stimulates rheumatoid synoviocytes to secrete several cy-**
       tokines such as IL-6, IL-8 and tumor necrosis factor-stimulated gene 6 as well
       as prostaglandin E2 in vitro [12,28,29]. (Instantiation)

Restatement

A relation is marked as "Restatement" when it indicates that the situation described by Arg2 re-
states the situation described by Arg1. The subtypes "specification", "generalization", and "equiv-
alence" further specify the ways in which Arg2 restates Arg1.

"Specification" applies when Arg2 describes the situation described in Arg1 in more detail, as in
Ex. (48).

(48)   *These CD123+ cells are similar in appearance to those previously demonstrated as CD123+*
       pDCs in human tonsil, <u>in that</u> **they are smaller than CD11c+ myeloid DCs, with**
       shorter dendritic processes, and cell clusters gave the appearance of locally
       proliferating cells (Fig. 1g) [5,36]. (Restatement:Specification)

"Generalization" (Ex. 49) applies when the connective indicates that Arg2 summarizes Arg1, or in
some cases expresses a conclusion based on Arg1.

(49)   *PCR amplification, cloning and sequencing of the bisulfite-treated DNA showed a specific*
       methylation pattern of the analyzed 62 CpG sites in all cell lines (Figure 4 and Table 1).
       <u>In general</u>, **the methylational status ranged from one cell line with a nearly non-**
       methylated IRF-4 promoter (SD-1, IRF-4-positive) to a completely methylated
       IRF-4 promoter in CML-T1 (IRF-4-negative). (Restatement:Generalization)

"Equivalence" applies when the connective indicates that Arg1 and Arg2 describe the same situation
from different perspectives. There were no annotations found for the equivalence sense.

Conjunction

The sense "Conjunction" is used when Arg1 and Arg2 are members of a list, defined in the prior discourse, explicitly or implicitly. In Ex. (50) the topic for the list is defined explicitly. No subtypes are defined for this sense.

(50)   Studies with transgenic mice overexpressing a constitutively activated form of TGF-β1 [6], as well as function-ablating mutant TGF-β signaling receptors, were confirmatory. *The ectopic expression of TGF-β1 resulted in a significant decrease in lateral branching*, <u>and</u> **mutant TGF-β receptors expressed in the stroma increased lateral branch infilling [7].** (Conjunction)

Exception

The type "Exception" applies when the connective indicates that Arg2 specifies an exception to the generalization specified by Arg1, as in Ex. (51). In other words, Arg1 is false because Arg2 is true, but if Arg2 were false, Arg1 would be true. No subtypes are defined for this sense.

(51)   *Similar experiments were performed as those examining c-jun phosphorylation*, <u>except that</u> **we used specific antibodies against CD69 to determine if the SEB reactive cells had been activated**. (Exception)

Reinforcement

The sense type "Reinforcement" (Ex. 52) is used when Arg2 is provided as fact to support claims or effects associated with Arg1. No subtypes are defined for this sense.

(52)   *However, in vivo, we observed much reduced IL-2 secretion, and reduced expansion of Vβ8+CD4+ cells, although this was not as dramatic as the reduction in IL-2 secretion.* <u>Indeed</u>, **it has been observed that SEB induced T cell expansion in IL-2 null mice, indicating that IL-2 is not necessary for expansion of T cells in vivo following SEB exposure**. (Reinforcement)

Continuation

The sense "Continuation" (Ex. 53 applies when Arg1 simply expands the discourse by identifying an entity (concrete or abstract) in Arg1 and saying something about it. Crucially, for this relation, it must be the case that no other discourse relation holds. "Continuations" occur frequently as implicit relations, but it is possible for *and* to be used with the continuation sense as well.

(53)   *Members of the transforming growth factor (TGF-beta) superfamily play central roles in controlling cellular proliferation, differentiation, migration and apoptosis [1].* <u>Implicit=NONE</u> **These cytokines can be divided into three subgroups: TGF-beta, the activins/inhibins,**

and the bone morphogenetic proteins (BMPs), of which the latter constitute the largest family. (Continuation)

<u>Circumstance</u>

In the sense "Circumstance", one argument provides the circumstances under which the situation in the other argument was obtained. No causal relation is implied here. In the biomedical articles, this relation was introduced specifically to capture the circumstantial relation between an experimental set-up and the observations and results obtained therefrom. Two subtypes capture directionality. In "Backward Circumstance", Arg1 describes the circumstance and Arg2 describes the resulting situation (Ex. 54). In "Forward Circumstance" (Ex. 55), Arg2 describes the circumstance and Arg1 describes the resulting situation.

(54)   *In order to test the 1.6-kb CCR3 promoter element (that includes exon 1) for in vivo function, this region was used to generate transgenic mice that expressed a reporter protein.* <u>Implicit=NONE</u> **Strong transgene expression was achieved, with the pattern of expression suggesting a broad acting promoter**. (Circumstance:Backward circumstance)

(55)   *They are equally effective in suppressing AA* <u>upon</u> **treatment of established disease [11,12] (and authors own unpublished observations)**. (Circumstance:Forward circumstance)

<u>Background</u>

The sense "Background" is used when one argument provides information that allows for a more well-specified interpretation of the other argument. THe two subtypes "Forward Background" and "Backward Background" capture directionality of this relation. In "Backward Background" (Ex. 56, Arg1 provides the background information for Arg2, while in "Forward Background", Arg2 provides the background information for Arg1 (Ex. 57). Connectives cannot be inserted for this relation, and a dummy NONE is inserted instead. No further subtypes are specified.

(56)   *Bone morphogenetic proteins (BMPs) belong to the TGF-β superfamily and are secreted proteins with pleiotropic roles in many different cell types. A potential role of BMP-6 in the immune system has been implied by various studies of malignant and rheumatoid diseases.* <u>Implicit=NONE</u> **In the present study, we explored the role of BMP-6 in normal human peripheral blood B cells**. (Background:Backward background)

(57)   *Significant changes in gene expression were calculated in the form of Z ratios and/or Z test values ([25]), using Z score values in all calculations.* <u>Implicit=NONE</u> **Z ratios constitute a measure of the change in gene expression of a given gene from its baseline value (in this case – time 0), expressed in units of standard deviation from the**

average change of all genes for that comparison. Z ratios are a direct measure of the likelihood that an observed change is an outlier in an otherwise normal distribution and, as such, are independent from underlying intensity values. (Background:Forward background)

5   Annotation Task Design Overview

For the task of annotating discourse relations, each annotator is given an article and instructed to read the article from beginning to end while marking up relations. No pre-defined lists of connectives were provided to annotators, although the connective list from PDTB was provided as an example of what to look for. Annotators were strongly encouraged to identify additional connectives when they were observed. At a high-level, the annotation procedure is encapsulated as follows:

For every new sentence encountered while reading the text:

1. First determine if there is an explicit connective that relates the sentence to the prior context via a discourse relation. If so, mark this explicit connective, its arguments, and its sense. Label the relation type as "Explicit".

2. If there is no explicit connective present to relate the sentence with the prior context, try to insert an implicit connective to express the inferred implicit relation, annotate its sense and mark its arguments. In case the inferred relation is one of the senses of CONTINUATION, BACKGROUND, or CIRCUMSTANCE, no connecive can be inserted, so use the dummy label "NONE" in place of an implicit connective. Label the relation type as "Implicit".

3. If insertion of an implicit connective leads to a redundancy in the expression of the relation, identify and mark the AltLex expression that expresses the relation, annotate its sense, and mark its arguments. Label the relation type as "AltLex".

4. If the sentence does not seem to coherently relate to any sentence in the prior text, label the relation type as "NoRel", mark the current sentence as Arg2 and the previous sentence as Arg1.

References

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

PDTB-Group (2008). The Penn Discourse TreeBank 2.0 Annotation Manual. Technical Report IRCS-08-01, Institute for Research in Cognitive Science, University of Pennsylvania.