Discourse Annotation Guidelines

Yuping Zhou

March 2, 2012

1 Introduction

Discourse annotation aims to depict the structure of a piece of text. This structure is made up of relationships among clauses, sentences, paragraphs etc., so what we focus on in discourse annotation is to find and characterize these relations.

Our approach to discourse relations is said to be “lexically grounded”. In this paradigm, prototypical discourse relations are mediated by discourse connectives, for example, 因为, 所以, 虽然, 但是, 而且 etc. In other words, the discourse connectives are the “lexicalized” representation, and hence the nucleus, of the underlying relation. They are taken to be predicates that take two arguments, much like a verbal predicate like “hit” that takes the hitter and the hittee as its arguments. In the sentence in (1), 因而 is the discourse connective representing the causation relation, and it takes two clauses as its arguments.


This prototype does not cover all cases in that some discourse relations do not have a lexicalized representation, for example,


Although there is no explicit connective between the two sentences, the contrast relation between them is obviously present. This is the so-called “implicit” discourse relation. To cover both classes of discourse relation in our annotation, we use the notion of “sense” of the relation to unify them: In (1), the sense of the explicit relation expressed by 因而 is “causation” and in (2), it’s “contrast”. With this notion of “sense”, the relation between the two sentences in (2), expressed without the help of a connective, can be unified with a relation expressed with the help of an explicit connective like 而, as in the following example.


Whether a discourse relation is expressed explicitly or implicitly, i.e. the type of relation, is only one of the aspects we need to specify in annotation; in the current scheme, we will annotate the following four aspects of a discourse relation:

1. relation type
2. discourse connective/DR-signaling punctuation
3. sense of the discourse relation
4. arguments
The rest of the document is organized as follows. We first describe how to systematically identify all the discourse relations in a piece of text in Section 2, then we expand on the four aspects of discourse relation in the following three sections: relation type (and connectives/AltLex’s) (in Section 3), relation sense (in Section 4), and arguments (in Section 5). In Section 6, we focus on how textspans should be determined in the case of punctuations, connectives/AltLex’s, and arguments. All above sections together define the general annotation framework, then in Section 7, we address areas that receive special treatment and in Section ??, we deal with technical issues in annotation. In most of the sections, the last subsection addresses very specific, potentially disagreement-causing issues pertinent to the concepts introduced in the section, so they may not be essential material in initial reading of the guidelines but will prove extremely useful in annotation.

2 Identification of discourse relations

2.1 General procedure

In our annotation, we use punctuations (pause marks (顿号), commas, periods, colons, semi-colons, ellipses, exclamation points, and question marks) as potential indicators for discourse relations. In other words, the general annotation procedure is to scan the text for punctuations, judge whether there is a discourse relation when a punctuation is encountered, and if there is, characterize the relation and if not, keep on scanning.

One consequence of this procedure is that explicit discourse relations (i.e. those mediated by explicit connectives) that do not include a punctuation in their textspans will be skipped. This does not seem to be a serious problem for two reasons: Such cases are quite rare in Chinese text, and they can be readily identified through other means (due to the presence of connectives) if we decide later that they should be included.

2.2 Is there a discourse relation?

The major problem we need to address with the aforementioned procedure in place is how to judge whether there is a discourse relation. The difficulty mainly lies in the case of comma when discourse connectives are absent. Wanting a satisfactory generalization at present, let us see a few examples in which the underlined comma does not mediate a discourse relation.

(4) a. 以联合国开发计划署和工发组织为代表的国际社会，一直对图们江地区国际合作开发计划给予积极的支持和有效的促进。

    b. 经济界人士称，中国应采取积极稳妥措施解决经济运行中的突出问题，特别要注意及时清理不良资产，加强金融风险防范。

In (4a), the chunk preceding the underlined comma is a complex noun phrase that identifies an entity and it serves as the subject of the rest of the sentence. In (4b), the chunk following the underlined comma is the complement of what comes before it, i.e. the chunk preceding the comma does not express a complete thought without the chunk following it. In both cases, the relation between the comma-separated chunks is not discourse in nature, but rather between a run-of-the-mill verbal predicate (“给予” in (4a) and “称” in (4b)) and its argument.

This is not to say that each argument of a discourse relation must express a complete thought. If we consider the underlined comma in the following example, although the chunk following it, strictly speaking, does not express a complete thought due to lack of the subject, the comma should be considered to mediate a discourse relation.

(5) 经济界人士称，[中国应采取积极稳妥措施解决经济运行中的突出问题]Arg1 [特别要注意及时清理不良资产，加强金融风险防范]Arg2。
The difference between (5) and the examples in (4) is that the missing part of the otherwise complete thought is only a portion of a comma-delimited chunk, instead of the entirety of the chunk: In (5), the missing subject in Arg2, “中国”, is only a portion of Arg1, whereas in both cases in (3), one chunk needs the entirety of the other chunk for the thought to be complete.

It is also worth pointing out that not all arguments in a discourse relation have to be verbal projections. For example, the Arg1’s in the following examples can be said to have a nominal characteristic.


b. 而由于 [台湾当局的阻挠]Arg1， [直到一九九零年刘春曙、王耀华访台以后，两岸文化才走上双向交流的轨道]Arg2。

The commonality these noun phrases share with verbal projections in (5) is that they all denote some event, i.e. something that will/should/have happen(ed) or occurr(ed) or hold/held true.

2.3 Trouble-shooting

2.3.1 Abstraction of event-expressing phrases

Occasionally, some seemingly event/proposition-expressing phrases actually do not express events/propositions due to abstraction. Consider the underline phrases below.

(7) “家庭出身”在中国逐渐为人淡忘始于中共十一届三中全会召开后不久的一九七九年一月，当时中央做出决定，所谓“黑五类”的子女今后在入学 、招工 、参军 、分配工作 等方面 主要看其本人表现，不得歧视。

At first brush, the underlined phrases may seem to express events, but they have been abstracted to become concepts—the tell-tale sign is the word in boldface “等方面”. So the pause marks in these cases should not be marked as relation-mediating punctuations.

3 Relation type

Once a discourse relation is identified, its type needs to be determined. In Section 1, we roughly divided all relations into two classes: explicit and implicit; here, we further divide explicit relations into two: labeled with “Explicit” and “AltLex”. So we end up with three type labels: “Explicit”, “AltLex” and “Implicit”, the meaning of which is explained in detail below.

3.1 Explicit

The label “Explicit” applies to a subclass of explicit relations that are mediated by “connectives”, i.e. a closed group of conventionalized words/expressions whose sole function is to connect or relate two clauses or sentences or other similar structures, for example, 因为 、所以 、虽然 、但是 、如果 、就 、或者 、而且 、随着 、以后 、等. A prototypical connective should satisfy the following criteria:

1. When the connective is deleted, the resulting two clauses/sentences can stand alone to express (almost) the same meaning as before, except that the relation between them is lost;
2. Intuitively, its object is the whole clause/sentence it is attached to, rather than a proper part of it. For example, the object of 至於 in the following example is the noun phrase “需要订购产品者（价格介於一千元至十万元之间）”, a proper part of the clause, rather than the whole clause itself, so it shouldn’t be marked as a connective. (As to 则 in the same argument, it should be marked as a connective.)
Here are a few other examples.

(9a)–(9c) are all alternative ways to express the “Causation” relation, a relation normally expressed by the connective/semospheric relation involved in it. As the name implies, an AltLex is a word/expresion that functions as a connective but is not one for various reasons. For example,

1. It is not a conventionalized expression and/or contains material specific to the context it appears in, for instance, demonstratives like “这个/样”, “如此” (as in “正因为这个原因,...”);
2. Its connective function is derived from its sentence-component-meaning. A prominent example is “使”, a prototypical verb expressing causation. Since not all causal relations are discourse relations (e.g. 去年夏天，10月8日发生的事件，使...), “使” should be marked “AltLex” when the causal relation happens to be a discourse relation (e.g. [货币贬值]Arg1[使]AltLex[墨西哥出口竞争力增加]Arg2).

Here are a few other examples.


(9a)–(9c) are all alternative ways to express the “Causation” relation, a relation normally expressed by the connectives (because)所以. (It needs to be pointed out that not all uses of these constructions/lexical items should be considered “AltLex”; to tell them apart, the same criteria explained in Section 2 should be employed.)

3.3 Implicit
The label “Implicit” is used when a discourse relation is not mediated by any word. For example, the relation between the two sentences in the following example should be classified as “Implicit”.


3.4 Trouble-shooting

3.4.1 Judgment on Connective/AltLex
To judge whether some word (combination) is a connective/AltLex, the criteria listed in Sections 3.1 and 3.2 should be applied. Since this is often a confusing task, here we list some more examples.

- 仍然：It implies some previous state/situation persisting, which makes it feel like a connective. However, it often indicates relation between a clause/sentence and an implicit one sharing the same/similar predicate, not between the two clauses/sentences in the text, so it is not a connective but a component of the clause/sentence it appears in. In the following two sentences, 仍然 provides connection between the old Arabic irrigation system remaining now and it being there sometime before now in (11a), and between the basic situation holding now and sometime before now in (11b), not between the clauses/sentences delimited by square brackets. So neither should be marked as a connective and both relations should be classified as “Implicit”.

3.2 AltLex
The term “AltLex” is short for “Alternative Lexicalization (of a discourse connective)” and we use it to refer to both the relation type and the word (combination) involved in it. As the name implies, an AltLex is a word/expression that functions as a connective but is not one for various reasons. For example,
(11) a. [巴伦西亚是西班牙著名的农业区]。[至今仍保留着古老的阿拉伯式灌溉系统]。
   b. [在即将过去的1998年，俄罗斯和美国在中亚地区继续明争暗斗]。[基本态势仍然 是“美俄互守”]。

- 把／将: They may sometimes seem like a connective, as in (12) below:

(12) 两国高层领导的近期会晤又将中苏关系进一步提升。

将 may seem like a connective/AltLex relating “两国高层领导的近期会晤” and “中苏关系进一步提升”，but actually, it is part of the sentence because the object of 将 is intuitively “中苏关系” , not “中苏关系进一步提升”.

3.4.2 Connective vs. AltLex

Once some word (combination) is determined to function like a connective, the next step is to determine if it is a connective or an AltLex. This judgment may also be hard in certain uncommon cases, so here we offer some heuristics:

- Consider the meaning of the word (combination) in isolation. If the meaning(s) one usually associates with it is like that of a connective and this meaning fits the meaning in context, then mark it as a connective; if the connective-like meaning is derived from its normal meaning, mark it as an AltLex. For example,


- Connectives are usually compactly lexicalized items (e.g. 反倒, 像是) whereas AltLex’s are more compositional than lexicalized (e.g. 更重要的是).

(14) [基本态势仍然是“美俄互守”]Arg1。[所不同的是]AltLex：[俄变“消极防守”为“积极防守”]Arg2。

For a list of hard-to-classify connectives and AltLex’s we have seen in data, see Appendix I.

3.4.3 Explicit vs. AltLex

Even when the call on connective vs. AltLex has been made, the call on Explicit vs. AltLex may still be hard in occasional “mixed”-type cases, usually involving a connective and an AltLex. When such a situation arises, one should first try deleting one of them to see if the sentence is impaired in some way. If one element is more important than the other, the more important element should determine the relation type; if both elements are equally (un)important, label the relation as “Explicit”. Consider the underlined items in the following two examples below:
(15) 证券分析员说，[由于]_{Conn}一九九六年香港地产大幅升值，[令]_{AltLex}深港两地地产价格差距拉大，随着回归临近，深圳地产开始回升，而在深港联运方面也有较大发展空间，这些因素为“深业控股”提供了良好的盈利环境。

Normally, “由于” is considered a discourse connective but “令” is not; in fact, they normally do not pair up but the pairing in the above sentence sounds OK. The above case falls into the second category: either word, without the other, can express the exact same meaning with the same degree of fluency, so it should be classified as “Explicit”. In contrast, the example below falls into the first category.

(16) 联合国工发组织近期对四川省江地区工业项目现状进行了充分调查，并对该地区未来工业发展项目提出建议。[他们还对有关国家提出的工业项目邀请专家进行评估]_{Arg1}，[评估之后]_{AltLex} [再]_{Conn} [把这些项目输入工发组织的网络进行国际招商]_{Arg2}。

The phrase “评估之后” is an AltLex and “再” is a connective; in the deletion test, “评估之后” fares better alone than “再” in the context, so the the relation type is “AltLex”.

Regardless of what the relation type is or which category the case falls into, the textspans of the two elements should be labeled as a connective and an AltLex respectively (see Section 6 for details).

4 Sense of the discourse relation

4.1 Division of senses

In the current scheme, we distinguish ten senses.¹

   或（者／是）

2. Causation: relation between cause and effect.
   （因为）。。。 （所以）
   因而／故而／因此／故       从而
   由于

3. Conditional: relation between a supposed condition and a supposed result.
   如果／假如／如／若。。。 就／那么／则       不管／无论。。。 都
   即使。。。 也       除非
   只要。。。 就

4. Conjunction: relation between two equal-status statements serving a common theme.
   并且       既。。。 又
   还       （同时）
   也       （而）
   以及       （首先）。。。 （其次／再者）
   此外／另外       一方面。。。 （另）一方面

5. Contrast: relation between two statements, co-occurrence of which seems contradictory, counter-intuitive, out of the ordinary etc.

¹Two senses were dropped from the initial line-up: “Restatement” and “Progression” due to scarcity and proneness to be mixed up with “Expansion” and “Conjunction”.

6
Several caveats are in order. First, not all uses of the above-listed lexical items are discourse-related. For example, 

4.2.1 General issues regarding sense 

4.2 Trouble-shooting 

10. NoRel: The label “NoRel” applies when no other sense label applies. As one might reasonably assume, it does not occur often in deliberately written text. And when it does, it usually involves a paragraph break. 

4.2 Trouble-shooting 

4.2.1 General issues regarding sense 

Several caveats are in order. First, not all uses of the above-listed lexical items are discourse-related. For example,
4.2.2 Issues with certain senses

- **Expansion:** The term should be construed as expansion of the whole clause/sentence (mainly the predicate), not expansion of one of the noun phrases within. For example, the relation between the two arguments in the following example should not be classified as “Expansion”.

  ![Example](22)

  Although the underlined noun phrase in Arg2 refers to the events listed in Arg1, Arg1 is not an expansion of Arg2; an expansion of Arg2 would be listing what effects each move has accomplished.

  An exception to the rule is when there is a clear intention of listing all the items in a noun phrase. For example,
4.2.3 Issues between certain senses

*Purpose*:

- **Expansion vs. Conjunction**

  For example, in \["Of course, presence of separate time modifications doesn’t exclude the "Temporal" label automatically, emphasized points made in the latter case: one being that what have been done and two being that they were effective. The emphasis heavily shifts to the second point in the following form:

\[a. \text{a.} \quad \text{b.} \quad \text{c.} \quad \text{d.}\]

- **Purpose**: The key to the sense “Purpose” is intention, so if and its derivatives with a verbal element suffixed to it (e.g., to, in, at, for) are treated as the same as by itself.

- **Temporal**: When both arguments have their own clear time modification (as in the example below), resist the temptation of classifying it as “Temporal” immediately (unless there is another clear indication of a temporal relation). The reason is because the temporal relation, in the example below, is not *directly indicated by the author*, but rather *indirectly* through the reader’s (semi-)automatic processing of the two temporal modifications (i.e., 1993 is in the past, hence before “now”).

(23) a. 产品品类多、各家的奖金制度又不同，消费者如果有意购买产品或加入传销行列，[可以] arg1 \[... \] arg2.

b. 如果大家推选我为村委成员，那{你们得答应我三条要求} arg1: [一不许损坏草场，以草当柴，牲畜要实行圈养；二不许乱砍滥伐，造成水土流失；三过去过量宰杀的牲畜数要恢复到牲畜归户时的水平] arg2.

In fact, this is an exception in form only; it can be seen as a shorthand, with “切记” and “你们得答应我” being omitted from the expanded lists respectively. Contrast this with (22): There are two equally emphasized points made in the latter case: one being that what have been done and two being that they were effective. The emphasis heavily shifts to the second point in the following form:

(24) 以下举措取得了显著成效: 一、国务院三建委在北京召开了三峡库区工程移民和经济开发方面的情况介绍会...

So the sense for (22) is “Conjunction” and that for (24) is “Expansion”.

- **Purpose**: The key to the sense “Purpose” is intention, so if and its derivatives with a verbal element suffixed to it (e.g., to, in, at, for) are treated as the same as by itself.

- **Temporal**: When both arguments have their own clear time modification (as in the example below), resist the temptation of classifying it as “Temporal” immediately (unless there is another clear indication of a temporal relation). The reason is because the temporal relation, in the example below, is not *directly indicated by the author*, but rather *indirectly* through the reader’s (semi-)automatic processing of the two temporal modifications (i.e., 1993 is in the past, hence before “now”).


Of course, presence of separate time modifications doesn’t exclude the “Temporal” label automatically, for example, in “[一九九三年，......], arg1; 接着, [在一九九五年......], arg2”, the underlined word clearly indicates temporal relation. And there are other ways of accentuating the temporal relation in the presence of temporal modifications.

**4.2.3 Issues between certain senses**

**Expansion vs. Conjunction**

The “Expansion” sense covers relations in which one argument is an elaboration or restatement of the other whereas the “Conjunction” sense covers relations between two equal-status statements serving a common theme. So their differences can come from two aspects:

- Level of abstractness/concreteness: the two statements involved in a “Conjunction” relation should be at the same level of abstractness/concreteness; “Expansion” is more complicated: when it involves an elaboration, the two arguments should be at different levels, whereas when it involves a restatement, the two arguments could be at the same level.

(26) a. Expansion-elaboration:


9
b. Conjunction:
[这就要求[中葡双方加强合作]_{Arg1} , [加快工作进度]_{Arg2} , 也[要求澳门社会各界和广大民众的积极支持和自觉参与]_{Arg2} , 才能够顺利实现澳门的平稳过渡、政权交接，为澳门特别行政区的美好未来打下一个牢固的基础。]

- Same/different point(s) being made: the two arguments in “Expansion” should make/contribute to the same point, whereas those in “Conjunction” should make two different points (albeit serving a common theme).

(27) a. Expansion-elaboration:
[农业生产是今年经济增长中的一个亮点]_{Arg1} , [尽管遭受了严重的洪涝灾害，今年全年粮食总产将达到四点九亿吨以上，与去年基本持平，连续四年获得丰收]_{Arg2} .

b. Conjunction:
[当时，中国的电话普及率只有百分之零点三八]_{Arg1} , [被称为“没有电话的国家”]_{Arg2} .

• Conjunction vs. EntRel

The distinction is related to the bigger context: if the meaning of the two arguments is more or less related to the theme of the article/paragraph, even if the internal relation between the arguments is not very close, the relation should still be “Conjunction”; in contrast, if the two arguments are about the same entity which plays a role in the theme of the article/paragraph, but the meaning of the arguments is otherwise not pertinent to the theme, then it’s “EntRel”.

For example, the following sentence is taken from an article reporting on Chongqing being considered to become a municipality directly under the Central Government and its purpose is to introduce the city of Chongqing in general terms.

(28) [重庆是中国特大城市之一]_{Arg1} , [具有三千多年的悠久历史]_{Arg2} .

Since Arg2 is not directly related to Chongqing being considered to become a municipality directly under the Central Government, the only connection between the two arguments is the city of Chongqing, an entity, so the relation should be labeled “EntRel”. If the purpose were to explain why Chongqing came up for consideration and Arg2 were about how important a role Chongqing plays in economical, industrial development of the Southwest region, then the sense should be classified as “Conjunction”. Consider the following example for which the sense label “Conjunction” is more appropriate:

(29) [联合国开发组织近期对图们江地区工业项目现状进行了充分调查]_{Arg1} , [并对该地区未来工业发展项目提出建议]_{Arg2} .

Certainly both clauses are about the entity, 联合国开发组织, however, the relationship between them is much closer than that: both describe what the organization has done to help carry out the plan, which is the local theme. So in cases like these, the label “Conjunction” is more appropriate.

It is worth stressing that “Conjunction” (actually, all the senses other than “EntRel” and “NoRel”, for that matter) takes precedence over “EntRel” (and of course “NoRel” too). So if a relation can be classified as one of the first group, ignore “EntRel” (and of course “NoRel”).

• Purpose vs. Causation

The key to “Purpose” is intention (and whether the intention becomes a fact or not is unimportant) whereas the key to “Causation” is facts (and whether one of them was someone’s initial intention or not is unimportant).
addition, the case of “Temporal” is more complex: when the temporal relation is synchronous, i.e. when neither co-occur in the same discourse relation; they just happen to attach to different arguments when they are in use. In argument is attached to.

specified in Table 1, in terms of the prototypical connective or the part of a prototypical parallel connective each argument is attached to.

The two arguments of a discourse relation are ordered; the order is dependent on the sense of the relation, as specified in Table 1, in terms of the prototypical connective or the part of a prototypical parallel connective each argument is attached to.

It should be noted that in the cases of “Expansion”, “Purpose” and “Temporal”, the two connectives never co-occur in the same discourse relation; they just happen to attach to different arguments when they are in use. In addition, the case of “Temporal” is more complex: when the temporal relation is synchronous, i.e. when neither
the “在。。。之后” type nor the “在。。。之前” type is in use, the order of arguments is determined by the text order.

In cases of ambiguity in context, if the two senses dictate different orders for the arguments, the first sense takes precedence.

6 Textspan

6.1 General rules

Textspan should be specified for four types of element: discourse-relation-mediating punctuations, discourse connectives, AltLex expressions, and arguments. In determining the text span, there are two general rules. First, the textspan of any element excludes the flanking punctuation/space as shown in the example below:

(33) 经济界人士称，(中国应采取积极稳妥措施解决经济运行中的突出问题)Arg1 [，]Punc [特别]Conn [要注意及时清理不良资产，加强金融风险防范]Arg2 __

Note that the underlined punctuations (as well as the space between them and the arguments) are outside the arguments they flank; the comma inside Arg2 is OK because it doesn’t flank Arg2.

One minor exception to the rule is when the left quotation mark is included in the Arg span, the right one should also be included (as in (34a), unless it would include a relation-mediating punctuation (as in (34b)).

(34) a. 中国法律规定：具有中国籍的香港、澳门居民，不管是否持有外国护照和旅行证件，只要[未向国籍管理机关申报为外国人]，[均可申请“港澳居民来往内地通行证”]Arg2 __

b. [当时，中国的电话普及率只有百分之零点三八]Arg1 [，] [被称为“没有电话的国家]Arg2 __ 现在，电话的普及率已经达到百分之八点一。

The second rule is that discontinuous textspans of a unit should be specified as one element. For example, all textspans with an underlined label in the following examples are part of a unit with a discontinuous textspan, and they should be united with those sharing the same label in the sentence.


6.2 Separation of arguments from connective/AltLex

When separating the arguments from the connective/AltLex of a relation, the main principle is keeping the textspan of each argument continuous as much as possible and marking up all the textspans. So connectives/AltLex’s at the periphery of the arguments are completely separated whereas embedded connectives/AltLex’s are both inside the textspace of the embedding argument and marked up again as connective/AltLex. Note the difference between the embedded and peripheral connectives/AltLex’s in the following examples.


When dealing with connectives/AltLex’s of a previous relation and arguments of the current relation, the connectives/AltLex’s at the periphery of the argument should be excluded and those embedded should simply be unmarked. For example,

(37) a. 证劵分析员说。由于[一九九六年香港地产大幅升值，深港两地地产价格差距拉大]Arg1，[随着回归临近，深圳地产开始回升]Arg2，而在深港联运方面也有较大发展空间，这些因素为“深业控股”提供了良好的盈利环境。


All the underlined connectives/AltLex’s are not pertinent to the discourse relations currently under consideration. In (37a), “由于” is in the periphery of Arg1, so it is excluded from the textspan of Arg1 of the current relation; “令” is embedded so it is included in Arg1, unmarked. In (37b), “再” is also embedded, therefore it is unmarked and included as part of Arg1. When we consider the discourse relation mediated by “再”, it should be separated from the argument it is embedded in, as shown below:


Similarly, “评估之后” and “再” are included in the Arg2 of this relation, unmarked.
6.3 Connective/AltLex

The textspan of connectives/AltLex’s is generally quite straightforward, but two things are worth mentioning. First, connectives/AltLex’s embedded inside an argument often escape annotators’ attention, so here we list some to enhance awareness: 又, 还, 也, 再, 仍, 却, 则, 就 (partnering up with “只要/如果...” etc.), 自 (partnering up with “...以来/前/后/时” etc.)

Second, connectives/AltLex’s sometimes have their own modifiers, they should be in a continuous textspan with (part of) the connective/AltLex they modify as much as possible. The underlined parts in the following examples are modifiers and they are included in a continuous textspan with (part of) the modified they are adjacent to.


However, two cautions with respect to this rule:

- **Modals:** Modals are modifiers of the predicate, not of the connective/AltLex. For example,


- **Connective vs. AltLex:** The distinction between connectives and AltLex’s should be maintained, i.e. connectives should be separately marked with AltLex’s even when they are adjacent to each other. For example,


6.4 Arguments

The textspan of an argument can range from part of a comma-delimited chunk to several sentences, all depending on the meaning.

(42) a. Part of a comma-delimited chunk:


b. One or more comma-delimited chunks:


c. One sentence:

Among these possibilities, four situations often trip up annotators.

- **Boundary embedded in a comma-delimited chunk**
The left boundary of Arg1 and the right boundary of Arg2 in (42a) above exemplify such situations. In addition to noun-phrase-modifying clauses, such a situation often arises in conjoined complements of predicates like 要求, 命令, 请(求), 发现, 注意到, 看到, 听说 etc., as well as conjoined complements of modals like 必(须/然/将), 需(要), 得(要), 能(够), 可(以), 要, 会, 应(该) etc. Some examples are presented below:

(43) a. 在中美关系发展的关键时刻，江泽民主席提出“增进了解、扩大共识、发展合作、共创未来”的主张，呼吁 [中美双方从长远观点和战略高度审视和处理双边关系]Arg1, [把握两国人民的根本利益]Arg2，着眼于一个繁荣、稳定、和平的世界带人２１世纪，从而为冷战后中美两国关系的发展指明了方向。

b. 中国取得一九九八年抗洪抢险决定性胜利后，开始启动国有林区天然林资源保护工程，下决心 [从根本上遏制中国生态环境恶化的趋势]Arg1, [治理水土流失]Arg2，增加森林植被，保证大江大河的安全。

c. 白皮书由此得出结论认为，使用原产地统计的同时，必须 [充分认识它的局限性]Arg1, [透过表面数字，认真分析各方在贸易中获得的实际利益]Arg2。

d. 首先 [保障群众的生活供养]Arg1, [采取综合措施]Arg2，稳定群众情绪，保持社会安定。

- **Comma-delimited predicate modifiers and subjects**
Comma-delimited predicate modifiers and subjects should be part of the first argument, even when they are shared by both arguments.

(44) a. Subject
[妥善而尽快地解决这些问题，直接关系到一九九九年十二月二十日中国对澳门恢复行使主权的顺利与否，和澳门特别行政区成立后的正常运作]Arg1, [直接关系到在澳门落实“一国两制”方针，实现“澳人治澳”、高度自治的根本问题]Arg2。

b. Temporal modifier
[五年前，中国的数字电话程控交换机市场为多种品牌的外国产品垄断]Arg1, [我们还担心中国能不能研制成功数字程控交换机]Arg2。

c. Scope-defining modifier
[公社会调查的二百一十家传销公司中，有九十九家是以购买「创业资料袋」（费用介於五十元到一干八百元之间）为加入条件]Arg1, [至於需要订购产品者（价值介於一千元至十万元之间）]Arg2, [则]Conn 有四十八家]Arg2。

d. Locational and temporal modifiers
其中青仙戏首次入岛亮相，演出六场，[从台北的大剧院到台南的圣母庙大殿，场场爆满]Arg1, [即使在台风将至的晚上，观众依然在电幕细雨中看完演出]Arg2。

e. Aspect-defining modifier
[在公务员本地化方面，一九九三年，澳督韦奇立向各政务司下达了一份有关公务员本地化计划的备忘录，作为本地化工作的一项方针]Arg1, [一九九六年，政府各部门百分之七十以上的处长、组长实现了本地化]Arg2, [一九九七年十一月，澳葡政府任命了历史上第一位华人司长；澳葡政府还提出了在一九九八年底前基本实现公务员本地化的目标。]
A special kind of temporal modifier, those introduced by temporal discourse connectives/AltLex’s, should be treated with caution. In cases like the following example, because all the semi-colon-delimited material is in Arg2 of the temporal relation (as shown in (45a)), the temporal connective-introduced modifier should not be included in Arg1 of the relation mediated by the first semi-colon (as shown in (45b)).


Other than situations like this, temporal modifiers introduced by connectives/AltLex’s should be treated as normal.

• One argument embedded in the other

In Section 6.2, we put forth a principle that the textspan of an argument should be continuous as much as possible, but in rare cases, when one argument is interrupted by the other, the textspan of an argument has to be discontinuous. The following are some examples.


The reason we are sure that one argument is embedded in the other is because the underlined phrase is only the subject of Arg2, not the subject of Arg1. Contrast the above two cases with the following example, in which there is (judged to be) no argument-embedding:


The underlined phrase in (47) is the shared subject of both arguments, so in such cases, it is possible to ensure continuity of the argument span and we do so readily.

• Textspan spanning several argument-worthy units

Sometimes it is hard to determine where the argument boundary should lie when a string of argument-worthy units with ambiguous relations to each other are encountered. The first line of defense is to test each simplex unit for argument-hood. For example, given the string “s-unit1, s-unit2, s-unit3” with relations “Causation” and “Expansion” between the two unit pairs, test if the “Causation” relation holds between s-unit1 and s-unit3. If it does, annotate the string as “(s-unit1) Causation ((s-unit2) Expansion (s-unit3))”; else, “(s-unit1) Causation (s-unit2) Expansion (s-unit3)”. More tricky cases will be addressed in detail in Section 6.4.1.

6.4.1 Textspan-determining principles and their interactions

If more clear-cut methods fail, take into account the following four principles when determining the textspan:

1. **Continuity**: the two arguments (along with the punctuation and the connective/AltLex) should form a continuous string of text, EXCLUDING attribution and meta-expressions (see Section 7.1 for details);

2. **Minimality**: everything else being equal, choose the smallest eligible unit as argument;
3. **Symmetry**: everything else being equal, choose argument spans displaying symmetric structure of meaning;

4. **Completeness**: everything else being equal, choose argument spans with a more complete meaning. Two points should be noted:

- Here “complete” means including units describing different aspects of an entity, NOT in the sense of including additional units such that omitted elements of the clause/sentence are explicitly expressed in the argument;
- Semi-colon, period, paragraph break, etc. are different levels of indication of completeness. For example, the right boundary of Arg2 in the following example could be placed in front of the underlined comma, due to the ambiguous scope of “还”， however, considering the whole sentence expresses a more close-knit, complete action, boundaries of Arg2 coincide with the sentence boundaries.

(48) [联合国工发组织近期对图们江地区工业项目现状进行了充分调查，并对该地区未来工业发展方向提出建议] Arg1。 [他们]还 Conn 对有关国家提出的工业项目邀请专家进行评估。 评估之后再把这些项目输入工发组织的网络进行国际招商] Arg2。  

Of course, “everything else” is seldom, if ever, equal. Here we can only offer treatment of some common patterns.

- When there is a string of relations of the same sense AND the exact spans are hard to determine, follow the principle of Minimality;

(49) 中国取得一九九八年抗洪抢险决定性胜利后，开始启动国有林区天然林资源保护工程，下决心[从根本上遏制中国生态环境恶化的趋势] Conjunction1—Arg1，[治理水土流失] Conjunction1—Arg2, Conjunction2—Arg1，[增加森林植被] Conjunction2—Arg2, Conjunction3—Arg1，[保证大江大河的安全] Conjunction3—Arg2。  

- When the pattern “conclusion1, expansion-of-conclusion1; conclusion2 expansion-of-conclusion2……” is present, the proper way to delimit the relation between the two conclusions is “[conclusion1 expansion-of-conclusion1] Arg1 [conclusion2 Arg2 expansion-of-conclusion2……] to ensure continuity. The following is an example.


b. [目前美国已成为中国第二大贸易伙伴和投资国，中国则成为美国第四大贸易伙伴] Conjunction—Arg1。 [中美经贸关系的改善给两国人民带来了巨大的利益] Conjunction—Arg2, Expansion—Arg1，[仅美国向中国出口一项，就给美国提供了 3 0 万个直接的高薪就业机会和上百万个间接就业机会] Expansion—Arg2。  

7 **Special-treatment items**

7.1 **Attribution and meta-expressions**

Attribution reveals sources of information being reported; meta-expressions conveys author’s attitudes, actions, opinions etc. with respect to the information being reported. Under normal circumstances, both types should be excluded from the textspan when intra-sentential relations are considered. The underlined expression below are such cases.

are often hard to annotate, except for the following three situations. This special rule is instituted because without a connective/AltLex, discourse relations across paragraph breaks should be applied; and the second specifically deals with inter-paragraph relations, to which the special rules are applied as “NoRel”.

Except for the above three types of situation, all implicit relations across a paragraph boundary should be annotated as “NoRel”.

Since it may be confusing to mix this special treatment up with normal annotation, we recommend a two-round annotation process: the first round deals with intra-paragraph relations, to which “normal” annotation standards are applied; and the second specifically deals with inter-paragraph relations, to which the special rules are applied.

---

b. 经专家测算，[今后三年上海固定资产投资总额预计为六千亿元至七千亿元]Arg1，[社会消费品零售总额预计年均增长百分之十三至十四]Arg2。外贸出口预计年均增长百分之十至十二，这些速度与国内生产总值年均增长百分之十至十一是相适应的。

c. 综合各方面因素，[今年对外贸易额仍将稳步增长]Arg1，[进口增加]Arg2，出口增长速度有所放慢。

d. 展望虎年，中国的经济列车尽管会有颠簸起伏，但[只要]Conn [调控措施适时、得当]Arg1，相信[会沿着预设的轨道稳健前行]Arg2。

e. 值得注意的是，[目前健康状况较好的老大、老三、老六是一卵三胎的男婴]Arg1，[而]Conn [另外三个两女一男是单卵单胎]Arg2。


g. 事实证明，冷战结束后，[中美两国的共同利益不是减少了，而是增加了]Arg1；[双方合作的领域不是缩小了，而是扩大了]Arg2；[双方对维护世界和平与稳定的责任不是减轻了，而是加重了]Arg2。

h. 毋庸置疑，[中美之间仍存在分歧]Arg1；[中美关系的发展仍面临困难和挑战]Arg2。

When several attribution/meta-expression-containing sentences are being included in an argument, only exclude the one(s) at the periphery and leave the internal ones unmarked.

(52) 据介绍，[近年来广东省对外经贸迅速发展，而包括来料加工、进料加工和外商投资企业从事的加工贸易占了相当大的份额。去年广东省整个加工贸易出口值达四百多亿美元，约占全省出口总值的百分之八十。其中一般贸易的一半、来料加工的一部分和外商投资企业的百分之八十以上属加工贸易出口。目前全省从事加工贸易人数五百多万，仅“三来一补”的企业就达三万多家，从业人员二百多万。近几年台湾厂商也在广东设立七千多家加工企业]Arg2。[加工贸易在广东外经贸发展中占有举足轻重的地位，同时也是粤港澳台经贸合作的重要内容]Arg1。

7.2 Implicit relation across a paragraph break

This special rule is instituted because without a connective/AltLex, discourse relations across paragraph breaks are often hard to annotate, except for the following three situations.

1. The last sentence of the first paragraph (or a proper part of it) is being expanded by the second paragraph, in which case, it should be annotated as “Expansion”.

2. The paragraph break coincides with the break between two clearly-stated points, in which case, it should be annotated as “Conjunction”.

3. There is strong cohesion between the two paragraphs, in which case, it should be annotated as “EntRel”.

For example,

(53) [交通的不便，严重制约了经济的发展]Arg1。

<paragraph break>
[为解决交通运输的“瓶颈”制约作用，广东省八十年代率先在中国引入市场竞争机制，推行“谁投资，谁收益；以路养路，贷款修路”等举措，多方筹集建设资金]Arg2。

Except for the above three types of situation, all implicit relations across a paragraph boundary should be annotated as “NoRel”.

Since it may be confusing to mix this special treatment up with normal annotation, we recommend a two-round annotation process: the first round deals with intra-paragraph relations, to which “normal” annotation standards are applied; and the second specifically deals with inter-paragraph relations, to which the special rules are applied.
7.3 Textspan of “NoRel” and “EntRel”

As pointed out in Section 4, “EntRel” and “NoRel” are somewhat degenerate senses compared to the others, and especially in the case of “NoRel”, it is hard to find arguments for it. So we think it makes sense to restrict the argument textspan in those cases.

- “NoRel”: The text spans of both arguments should strictly coincide with the involved sentences, as defined by the punctuations in text since it usually occurs at the paragraph break.

\begin{quote}
(54) [我希望很快就能公布，但是现在还需要等体检结果，”他说]_{Arg1}。\\
\textless paragraph boundary> \\
[德波尔大名鼎鼎，今年入选荷兰队，并在法国世界杯赛上打入半决赛]_{Arg2}。
\end{quote}

Notice that this rule trumps rules regarding attribution (since “他说” is included in Arg1), and regarding pairing up of quotation marks (since Arg1 contains only the right one).

- “EntRel”: The text spans of both arguments should not go outside the confines of the involved sentence(s).