THYME Clinical Cross-document Coreference Annotation Guidelines

Developed by Kristin Wright-Bettner and Martha Palmer with input from Tim O'Gorman, Piet de Groen, Guergana Savova, Timothy Miller, Steve Bethard, and Dana Green.

Synthesizes and expands on the THYME Annotation Guidelines, Clinical Coreference Annotation Guidelines, and Richer Event Description (RED) Annotation Guidelines.

I Introduction

Aim and process

The THYME (Temporal History of Your Medical Events) cross-document coreference project is interested in capturing the coreference and structural relations that exist between events and entities in different but related clinical documents. It utilizes and expands on annotations done by two other projects on the THYME corpus: the THYME 1 Temporal Relations Annotation project[1] (referred to below as *T1Temp*) and the THYME 1 Clinical Coreference Annotation project[2] (referred to below as *T1Coref*).[3] These may be referred to collectively as THYME 1 or T1. The current cross-document project, which includes a modified version of the T1 annotations, is referred to as THYME 2 or T2.[4]

[1] THYME 1 temporal guidelines: http://clear.colorado.edu/compsem/documents/THYME_guidelines.pdf

[2] THYME 1 coreference guidelines: http://clear.colorado.edu/compsem/documents/coreference_guidelines.pdf

[3] In Wright-Bettner et al., 2019, T1Temp was referred to as *THYME 1*; T1Coref was referred to as *Clinical Coreference*.

[4] In Lin and Wright-Bettner et al., 2020, THYME 2 was referred to as *THYME+*.

The THYME corpus consists of de-identified medical records – specifically, doctors' notes about cancer patients. (While the entire THYME corpus consists of both brain cancer and colon cancer notes, only the colon cancer notes are annotated by THYME 2.) T1Temp focused on creating a timeline for each patient note, so this project annotated these records for events that are relevant to the patient's clinical timeline (MRI, surgery, etc.), temporal expressions (such as the date of surgery), and temporal relations between the events (showing whether the surgery came before or after the MRI, for example). T1Coref was concerned with coreference relations – that is, whether two different mentions in a note refer to the same real-life event or entity – and structural relations, such as whether one entity makes up part of another entity. Both coreference and structural relations are referred to below collectively as "coreference relations" in order to distinguish them from temporal relations.

The annotated cancer documents consist of sets of three notes for a given patient, two clinical notes and one pathology note. All previous annotations were done on one document at a time – that is, all the temporal and coreference links only relate markables within a single document. In the current project, we create cross-document coreference links between the preannotated markables for each set of three notes; for example, we identify when a tumor in clinical note A is the same as a tumor in clinical note B, which therefore enables us to track its status over time.

The starting-point material for THYME 2, then, consists of annotations from T1Temp and T1Coref that have been merged. While these two projects annotated the same set of documents, they were done separately and with different purposes. As a result, there are naturally some inconsistencies in the merged data. The current project is therefore done in two major annotation stages, the first of which focuses on synchronizing the data from the two previous projects, as well as adding a new temporal relation (CONTAINS-SUBEVENT), which cross-document experiments have shown to be valuable. The actual cross-document linking is accomplished in the second stage. Finally, the entire dataset goes through a post-processing pass, consisting of both automatic and manual tasks. A key component of this pass is the addition of a second new temporal relation, NOTED-ON, which is discussed in depth in Appendix C.

Therefore, the first annotation pass occurs within-document; the second pass is cross-document. There are also two adjudication stages, in which a third annotator compares two sets of annotation choices and resolves disagreements. Adjudication takes place: a) after the within-document annotation pass (comparing the original merged gold data from the previous two projects and a single annotator's adjustments to that data), and b) after the cross-document pass (comparing two different annotators' versions of the same set of three notes). The main pipeline stages are therefore:

1) Within-document annotation

2) Within-document adjudication

3) Cross-document annotation

4) Cross-document adjudication

5) Post-processing

Additionally, one-third of the files were not originally annotated by T1Coref, so a medically-trained annotator annotates all physical anatomy and property relations in separate within-document and cross-document passes. This pipeline is also discussed in Appendix C.

In summary, this project is necessarily complex: Two different annotation schemas are being merged and synthesized to support a brand-new task (cross-document linking), which itself has exposed the need for changes to the original merged schema. For an at-a-glance overview of all three schemas (T1Temp, T1Coref, THYME 2) and the major changes made, please see the following tables and comments. These are repeated in Appendix D.

Table 1: **THYME 1 temporal relations schema (T1Temp)**

T1Temp markables	**T1Temp relations**	
- EVENTs (*any conceptual event, regardless of POS*)[1] - TIMEX3s - DOCTIMEs - SECTIONTIMEs	*TLINKs*	*ALINKs*
	- BEFORE - OVERLAP - CONTAINS - BEGINS-ON - ENDS-ON	- CONTINUES - INITIATES - REINITIATES - TERMINATES

***Table 2*:** **THYME 1 coreference relations schema (T1Coref)**

T1Coref markables	**T1Coref relations**
- MARKABLEs (*any non-singleton noun, noun modifier, pronoun, or nominalized verb; may be an event, entity, or temporal expression*)[1]	- IDENTITY - APPOSITIVE - SET-SUBSET - WHOLE-PART

***Table 3*: THYME 2 temporal and coreference relation annotation**

Markables		
EVENTs Entities (*including singleton entities*[1, 2]) TIMEX3s SECTIONTIMEs DOCTIMEs		
Single-file relations		
TLINKs	*ALINKs*	*Coreference links*
BEFORE OVERLAP NOTED-ON[2, 3] CONTAINS CONTAINS-SUBEVENT[2, 4] BEGINS-ON ENDS-ON	CONTINUES INITIATES REINITIATES TERMINATES	IDENTICAL APPOSITIVE SET-SUBSET WHOLE-PART[5] CONTAINS-SUBEVENT[2, 4]
Cross-doc relations		
IDENTICAL SET-SUBSET CONTAINS-SUBEVENT WHOLE-PART[5]		

[1] Note that T1Coref's markables were determined by POS and whether there was a coreferential within-document mention; T1Temp's markables were determined by eventiveness, regardless of POS or coreference. Singleton entities were therefore unmarked by either original project and were added new in THYME 2. (All entities are called MARKABLEs in THYME 2's output due to T1Coref's labeling practices. See Terms: entities, markables, and MARKABLEs.)

[2] These terms indicate **new annotation categories** added in THYME 2.

[3] NOTED-ON is a subtype of OVERLAP. See discussion of NOTED-ON under Manual post-processing tasks in Appendix C.

[4] CONTAINS-SUBEVENT is a subtype of CONTAINS. It appears under both TLINKs and Coreference links because it represents both temporal information (temporal containment) and partial-identity information (event-subevent). While other TLINKs have distance restrictions, CON-SUB does not. It was restricted to four semantic categories: cancer treatment events; cancer events; medication usage events; and chronic disease events. See Adding CONTAINS-SUBEVENT links.

[5] WHOLE-PART application dramatically changed between THYME 1 and THYME 2. T1Coref usage subsumed compositional, locative, and subevent relations; THYME 2 usage restricts to compositional whole-part relations between entities only. THYME 2 cross-document application restricts this even further; it was used only for non-anatomy WHOLE-PART

relations. See THYME 2 Anatomy Linking guidelines, Cross-Document Anatomy Pass, for discussion.

Acknowledgements

This project is based on previous work by THYME temporal relations annotation (THYME Annotation Guidelines, developed by Will Styler, Guergana Savova, Martha Palmer, James Pustejovsky, Tim O'Gorman, and Piet C. de Groen, based on the ISO TimeML temporal relations specification) and THYME coreference annotation (Clinical Coreference Annotation Guidelines, developed by Arrick Lanfranchi, Kevin Crooks, and Mariah Hamang with excerpts from ODIE guidelines and modified for SHARPn/THYME). It also draws on many concepts from the RED project (Richer Event Description (RED) Annotation Guidelines, developed by Will Styler, Kevin Crooks, Mariah Hamang, and Tim O'Gorman as a synthesis of the THYME-TimeML guidelines, the Stanford Event coreference guidelines and the CMU Event coreference guidelines, with substantial input from Eduard Hovy and Teruko Mitamura at Carnegie Mellon University, Mariah Hamang, Kristin Wright-Bettner, and Rei Ikuta at the University of Colorado, Boulder, and discussion at the NAACL Events Workshop). As a result, some sections of these guidelines are taken from the guidelines for these other three projects, sometimes conceptually, **sometimes word-for-word**.

II The preannotated THYME 1 temporal and coreference data

As stated above, the material for this project comes preannotated – i.e., many annotations have already been created by the previous two projects. In order to successfully accomplish both stages of the current project, you first have to know what you're looking at.

1 Markables

Markable types

The documents will appear with primarily three types of markables annotated: events, entities, and temporal expressions. ("Markables" are words in the text that we care about annotating; as we'll see, not all words merit an annotation.)

EVENTs

We define an event as any occurrence, action, process or event state which deserves a place upon a timeline. Events can range from procedures to diseases to diagnoses to patient complaints and states. You'll note that event choice is based on semantics – what a term means, i.e., what it's "pointing" to in the real world (or a possible world) – and not on part of speech. So you'll see events that are verbs, nouns, or even adjectives:

(1) The patient is [stable].

(2) The patient has [nausea] and peripheral [edema]

(3) Patient [reports] severe back [pain].

As you can see, copulas and semantically light verbs like "has" in (2) were not marked, because they don't add any semantic information. This includes supporting verbs as in the following:

(4) She has been experiencing neck [pain] since {July}.

(5) We will perform a [hemicolectomy].

In the above examples, if we've already captured [pain] and [hemicolectomy] and put them on a timeline, marking "experiencing" or "perform" doesn't buy us any additional information – if pain exists, it's being experienced; if a hemicolectomy occurs, it's being performed – and the span occupied on the timeline would be exactly the same.

We also do not mark numbers as EVENTs in this schema. It's tempting to do so when they reflect clinically significant values:

- Grade **3** of 4
- Patient's 30-day post-operative mortality is **28**%
- Carotid pulses are **4**/4.

But these are being captured by other projects.

The one exception to this is we do permit marking numbers in reference to medications dosages in the medications section only; this is discussed more below in Medications sections.

Entities

An entity is a participant, location, organization, or other referential thing that might be tracked in the discourse. Broadly speaking, entities may be thought of as things that exist while events are things that happen. You can locate an event on a timeline (the patient's [surgery] on {May 24th, 2010}), while entities can't usefully or intuitively be linked to a timeline – you couldn't put, say, Boulder County Hospital on a timeline (apart from the date it came into existence and the date it ceases to exist – in which case, its inception and its disbandment are the actual events that would be associated with the timeline). In clinical notes, entities are usually mentions of people (doctors, patients, patient family members), anatomical sites, and locations like hospitals or organizations like medical departments:

(6) [I] have answered all questions from [Mrs. Carson] and [[her] family].

(7) [Liver] is questionable for metastasis.

(8) [She] will return to [Dumont General Hospital] for follow-up in one months' time.

We also mark pronouns that refer to both entities and events. We do not mark relative pronouns, however (who, which, etc.).

Temporal expressions, DOCTIME, and SECTIONTIME

Temporal expressions (TIMEX3s) are definitive references to time, such as dates and times. They are discussed more in below in TIMEX3s.

DOCTIME and SECTIONTIME may be thought of as "special" temporal expressions that are annotated similarly but are each their own category. Every note has a single DOCTIME annotation, the start_date at the top of each note:

(9) meta rev_date="11/02/2012" start_date="[10/26/2012]" rev="0002"

The bracketed date (the start_date) identifies "document time," the time the note was created.

Sometimes a specific section has an explicit, different timestamp from the rest of the document:

(10) [start section id="20104"]

Advil 1 tablet by mouth daily.

Vitamin D 1 capsule by mouth two times a day.

These are the patient's medications as of [Thursday, April 9, 2012 at 10:31 AM].

[end section id="20104]

Here, [Thursday, April 9, 2012 at 10:31 AM] would be annotated as a SECTIONTIME.

While temporal expressions are linked to EVENTs in the note to show their temporal relationship, DOCTIMEs are never linked to anything. SECTIONTIMEs should be linked only in rare, specific cases (see SECTIONTIME).

More on distinguishing EVENTs from entities: terms, spans, and implicit EVENTs

Terms: entities, markables, and MARKABLEs

While events are labeled with an EVENT category in our schema, entities are called MARKABLEs (due to the labeling practices of T1Coref).

Because of this, in the rest of the guidelines, lower-case “markable” refers to anything that is referential that is or should be marked in the document – this includes EVENTs, entities, TIMEX3s, DOCTIMEs and SECTIONTIMEs. Upper-case “MARKABLE” refers specifically to the actual category in the online annotation tool, which corresponds to entities.

In the examples in these guidelines, temporal expressions are distinguished by curly brackets { }, while EVENTs and MARKABLEs are enclosed by square brackets [ ]. When it’s necessary to distinguish between the two, EVENTs are marked with an *E* and MARKABLEs with an *M*. Typically, only the markables relevant to the current discussion are marked, rather than the examples being fully annotated.

Spans

You may have noted in the examples above under EVENTs and Entities that the spans of the markables are different – that is, the amount of text that’s been selected is different for EVENTs than it is for entities.

This is because, in T1Temp, EVENTs (E) were marked with minimal-span annotation – that is, only the headword was included in the span. T1Coref, however, used maximal-span annotation for its MARKABLEs (M), including the entire syntactic phrase in the span.

Headedness is the idea that there is a word in each phrase which essentially defines and represents that phrase. An initial test for this is that the "head" can generally stand in for the entire phrase, and mean roughly the same thing with roughly the same grammatical properties. Thus in a noun phrase like "my insatiable need for more donuts," the thing that really represents the notion is "need."

Another way to think about this is to consider what is being modified by the words in this section of the sentence. With "my insatiable need for more donuts," both "my" and "insatiable" give us more information about the nature of the need, and "for more donuts" clarifies more about what the need is for. Chances are, if you look at the whole clause, you'll find that much of the information is pointing to or modifying a single word, and that will usually be your head.

So, given this phrase, minimal-span annotation looks like this: my insatiable [need] for more donuts. Maximal-span annotation looks like this: [my insatiable need for more donuts]. As you can see, the entire syntactic phrase in the maximal-span example includes all modifying information to the head noun, including determiners (my), prepositional phrases (for more donuts), relative clauses, premodifiers (insatiable) and postmodifiers. However, both the minimal-span and maximal-span annotations point to the same real-life event – the need.

With this in mind, observe the span differences for MARKABLEs versus EVENTs in the clinical example below (and note that in the case of possessives, the apostrophe and "s" are included in the span):

> (11) [Patient's]$_M$ recent [scans]$_E$ unfortunately show [progression]$_E$ of [disease]$_E$ in [the rectum]$_M$ and [the right lobe of [the liver]$_M$]$_M$.
>
> **EVENTs**:
>
> - [scans]
> - [progression]
> - [disease]
>
> **MARKABLEs** (i.e., entities):
>
> - [Patient's]
> - [the rectum]
> - [the right lobe of the liver]
> - [the liver]

Because MARKABLEs are marked with maximal span, it's frequently the case that their spans overlap the spans of other markables:

> (12) I will contact [Dr. Warren who performed right [hemicolectomy]$_E$ on [the patient]$_M$ {last week}$_T$]$_M$.

First of all, note that for the long MARKABLE [Dr. Warren who performed right hemicolectomy on the patient last week], the head is "Dr. Warren." The real-life referent the whole phrase is pointing to is this physician. "Dr. Warren" could stand in for the whole phrase, and it would still make sense: "I will contact Dr. Warren."

The relative clause “who performed right hemicolectomy on the patient last week” is modifying “Dr. Warren,” so it’s also included in the span. But that clause itself contains other, different referential things – an EVENT, [hemicolectomy]; another MARKABLE, [the patient]; and a temporal expression, {last week}. The fact that the spans overlap is not problematic.

Finally, you’ll observe that some MARKABLEs have “disjoint spans,” when the phrase in which they’re found is interrupted by other material:

(13) No sign of disease in [[the abdomen] or pelvis].

The span for pelvis is: [the...pelvis].

Implicit EVENTs

One complication to the above distinction between events and entities (see EVENTS and Entities) is that we mark several categories of terms as EVENTs that at first glance look like entities. You’ll quickly notice this if you take a look at the EVENTs that show up in the preannotated data. This is because many entity mentions may be understood as implying events. For example:

(14) Continue Gold Bond medicated [powder].

At face value, [powder] is a physical object, and it apparently meets all the qualifications we just described for an entity. However, when we read this sentence, we know that it means that the patient is going to continue applying or using Gold Bond powder. The patient’s use of a medication is certainly linkable to a timeline, and of course, has obvious clinical relevance. Therefore, we’ve chosen to mark many of these kinds of terms as EVENTs.

It takes some practice to develop this sense, so following are several categories of clinical terms that we consistently mark as EVENTs, and several categories of things that should never be interpreted as eventive (i.e, they’re always entities). Neither are comprehensive, but they do encompass many of the terms you’ll run across.

Note there is no actual feature indicating if an EVENT is implicit or not. You do not actually need to identify an EVENT as being implicit or not for this project; the point here is simply to mitigate confusion for when you encounter EVENTs that look like entities.

Things that are always EVENTs:

a) Reports/records/etc. (referring to the initial act of reporting):

- We have requested the outside [record].
- Please scan the patient’s most recent colonoscopy [report].

("Form," however, as in "The patient signed the form," should be an entity.)

b) Medications and medical devices (referring to the patient's having, ingesting, or using them):

- [Catheter] inserted yesterday.
- Continue Gold Bond medicated [powder]
- I have advised her to take [Advil] 100-mg once a day.

c) Tumors and other abnormal growths/tissue (referring to the patient's having them):

- Scan shows [cyst] in liver.
- Biopsy showed grade 1 (of 4) [adenocarcinoma].
- [Tumor] is invasive.

This includes all references to abnormal tissue, including parts of the tumor and tissue removed from the tumor:

- [part] of the [tumor]...

d) Labs (referring to the act of testing):

- We need to see improvement in [creatinine] prior to starting this medication.
- [Hemoglobin] was low.
- Patient's [WBC] is stable.

e) Physical objects that imply the patient's ingesting or use of them:

- Alcohol intake: Three [beers] a day
- Nutrition: No [fruits]$_{\text{NEG}}$ or [vegetables]$_{\text{NEG}}$.
- Patient quit [tobacco] last year.

(Note this last category often shows up in frequency-type phrases like the first example – three [beers] a day. As an aside, also note that in quantifier phrases like this one, we mark the thing being quantified – beers, in this case – and not the quantifying term.)

Things that are never EVENTs:

a) People

- [Patient] was transfused with two units of blood.
- [Dr. Jennings] will see [her] tomorrow.
- Family history significant for [a sister with brain cancer].

One context that might initially seem counter to this is equational clause constructions (*X is Y*) in which the "Y" element of the clause tells us something about the patient that has clinical significance, such as: *Patient is a [smoker]*. You'll often find that [smoker] in this case has been marked as an EVENT, in an effort to capture the implicit, medically-relevant event that the patient smokes. This is fine.

b) Places

- Patient presents to [Arnold Health Center] for resection of recurrent tumor.
- He will see his primary oncologist when he returns to [Florida].

c) Departments/organizations

- We need [Radiation Oncology] to review the scans.
- Patient scheduled for consult with [Surgery].

d) Functions and properties

These are typically functions or properties of body parts or regions. For example:

- [Extraocular movements]$_M$
- [muscle strength]$_M$
- [heart rate]$_M$

The same is true for cancer properties as well as body-part properties:

- We have discussed [the stage of the cancer]$_M$ and [the grade of the cancer]$_M$.

But, note that the values of these properties are EVENTs, as they are states that may change. For example:

- [The stage of the cancer]$_M$ is [pT3M0NX]$_E$.
- [Extraocular movements]$_M$ [intact]$_E$.
- [[Regular]$_E$ heart rate]$_M$.

This is discussed in more detail in Properties and their states.

e) Body parts, parts of body parts, and tissue removed from body parts

- [Multiple (62) lymph nodes] are identified within [the perirectal fat].
- Concern for cancer in [the rectosigmoid colon].
- Mr. Smith complains of severe pain in [the left lower quadrant].
- [The distal resection margin] is positive for tumor.

Margins are the tissue surrounding an abnormality that has been excised.

- ...[a segment of [colon]].

This is just part of the colon. (Note “colon” is marked as well – there are two different referents here, the colon and the part of the colon under discussion.)

- It is difficult to make this decision without knowing [the exact source of the patient’s December 2011 biopsy].

The [...source...] is talking about where in the body the tissue was removed from for examination. Because it's talking about a location in the body, albeit an unknown one, it's an entity.

- ...[tissue block].

This is just tissue that's been removed for examination.

- [Representative sections] are submitted.

These phrases frequently show up in path notes. Sections are parts of the physical specimen.

Our concern is always with semantics, i.e., what real-life thing is being pointed to by a given term. So for the immediately preceding example, whether it's called "a piece of tissue," "a representative section," or "part of the specimen" doesn't matter – if it refers to tissue that's part of an anatomical site or removed from an anatomical site, then it's an entity.

Finally, note that we treat normally-occurring body tissue references as entities and abnormal tissue references as events (see *Things we always consider to be EVENTs* above). This is because the presence of abnormal tissue in the body (tumors, cysts, etc.) is medically significant, and we want to be able to link the presence of such tissue to the timeline. But saying that normal tissue is present for a patient is like saying they have an arm – it doesn't contribute much to the clinical timeline.

There are a couple cases where this distinction runs us into trouble, however, and that's most noticeable with the term "specimen," which is somewhat unique in that it can consist of nearly anything – normal tissue, abnormal tissue, often both. We've therefore decided to always treat specimen references as entities, and you should heavily lean toward understanding other tissue mentions as entities as well. For tissue to be marked as an EVENT, it must be explicit in the text that it's only abnormal, like the [part] of the [tumor] example above. So for the following, "specimen" is an entity (i.e., MARKABLE):

- Received fresh labeled hepatic mass is [a 2.2 x 0.3 x 1.7 cm left hepatectomy specimen].

The same is true for "margins" and "sections" (examples above).

A key takeaway from this section is that something abnormal in the body is always an (implicit) EVENT (medical devices, tumors, etc.). Normally occurring things in the body are *nearly* always entities (fat, lymph nodes, etc.), but [blood] and [stool] are frequent exceptions to this as they often refer implicitly to clinically significant events, such as the passing of blood or stool by the patient, or the transfusion of blood as medical treatment:

- Mr. Robinson reports {eight} [stools]$_E$ {daily}.
- Patient presents with daily [blood]$_E$ with bowel [movements]$_E$.
- We will transfuse Ms. Wilson with three units of [blood]$_E$.

EVENT features

Each EVENT has been annotated for a variety of more specific information about the event, such as when it happened, whether it's a hypothetical occurrence or a real one, and so forth. You'll be able to view these properties when you click on the event in the program. Understanding them is important for the decisions you'll be making in all stages of this project, so these properties are discussed individually below. (MARKABLEs, however, are featureless.)

DocTimeRel

DocTimeRel is short for "Document Creation Time Relation" and represents the temporal relation between the EVENT in question and the time when the medical record in which it occurs was created (the "document time," or DOCTIME). For the purposes of this schema, we are assuming that writing of the record itself is functionally equivalent to the time of the patient's visit to the physician. So, anything considered true during the visit will be considered true at the time the visit was documented by the physician.

Our schema includes four potential relations between the event and DOCTIME: BEFORE, AFTER, OVERLAP, and the combined relation BEFORE/OVERLAP.

BEFORE

BEFORE is used when the event ended before the patient was seen (and thus, before the document itself was written). The bracketed events below would be marked as "BEFORE" (all other EVENTs and TIMEX3s are unmarked):

(15) We will order another CT to compare with his prior 9-16-03 [study].

(16) She had experienced no [dizziness] until the [start] of chemotherapy.

(17) The patient had had no [fever] before her [surgery] last week.

OVERLAP

OVERLAP is used for events or states which are happening or true at the time that the patient was seen and thus, we presume, when the document was written:

(18) The patient is [alert], [cooperative], and appears to be in no acute [distress].

(19) Moderate sized retention [cyst] or [polyp] in the right maxillary antrum again [noted].

(20) I have [discussed] the treatment plan with the patient and [answered] [questions].

Note that frequently events represented by past-tense verbs are still marked as OVERLAP because they occurred at the time of patient visit. These are often discussion events, as in (20).

AFTER

AFTER is used where the event is planned or understood to begin following the document time:

(21) [Levaquin] 750 mg p.o.q. day will [restart] today.

(22) The patient will [return] tomorrow for [labs] and [exam].

BEFORE-OVERLAP

BEFORE-OVERLAP is used when an event started before the exam or patient visit (i.e., DOCTIME) and continues through to the present.

(23) The patient has [felt] quite well and his appetite has been [good].

(24) She has not [seen] a cardiologist.

(25) She has had no [fever].

BEFORE-OVERLAP requires explicit linguistic evidence in the text that an EVENT started before DOCTIME and continues through it. Even if you as a human can infer that a given EVENT – say, [cancer] – must've started before the document time, an EVENT shouldn't be marked as BEFORE-OVERLAP unless there's evidence in the sentence itself. This evidence may include use of the English present perfect tense (but not always!), and modifying lexical items like "still," "continues," etc., that inherently indicate a prior initiation and continuation of an event.

Type

Some EVENTs don't actually represent clinical events, but instead provide aspectual information about other EVENTs (whether they're starting, stopping, or continuing), or evidential information (attributing what we know to how we know it). To differentiate these EVENTs from the traditional clinical EVENTs which occur on a timeline, we use the "Type" marker. It has three values: "N/A," "Aspectual," and "Evidential." "N/A" is the default value, and represents the vast majority of EVENTs in the schema, and unless explicitly mentioned otherwise, all EVENTs used in examples in the guidelines.

ASPECTUAL

The EVENT type ASPECTUAL is used to indicate an event whose function is to code the aspect of another event, like "continues" or "restart." Every EVENT of Type ASPECTUAL must later participate in an ALINK (discussed in ALINK annotation).

(26) His anterior chest rash has not [reoccurred].
(27) We're going to [hold] her heparin until after her surgery.
(28) The patient will [continue] treatment.

These represent a relatively closed class, and you'll find the same words marked as aspectual EVENTs over and over again. This is expected, and should not be cause for concern.

EVIDENTIAL

The other EVENT type is EVIDENTIAL. EVIDENTIAL EVENTS provide information about how doctors came to identify and learn about other events. EVIDENTIAL EVENTS are also a relatively closed class, generally verbs of showing, demonstration, evidence, confirmation, or revelation. In the clinical domain, these are very commonly associated with tests, imaging, and human observation.

In short, an EVENT is EVIDENTIAL if it serves as the link between a source of knowledge or observation and a piece of knowledge gained from it.

(29) Her CT-scan [showed] a small mass in the right colon.

(30) Subsequent bloodwork [revealed] severe anemia, and the patient was admitted.

(31) The patient [reported] severe back pain throughout the month of May.

It is worth reiterating that one does not mark the test, reporter, or perceiver as EVIDENTIAL; instead, the verb of perception, reporting, revelation, or indication is the EVIDENTIAL EVENT.

Polarity

Polarity in this schema is relatively straightforward, and there are only two possible types: positive (POS) and negative (NEG). POS is the default polarity.

POS

The most commonly used polarity value is POS. This is used for an EVENT that did, in fact, occur (or is occurring, or will occur, to the best of our knowledge).

(32) The patient has [hepatosplenomegaly].

(33) PO [changes] right pterional [craniotomy]

(34) The patient will [continue] [treatment].

NEG

The opposite of POS, as you might guess, is NEG, which is used to indicate when the event didn't take place:

(35) Otherwise, he has not had any [nausea], [vomiting], [diarrhea], chest [pain], [shortness] of breath, or [fever].

(36) No [evidence] for new suprasellar [mass].

Note that NEG does not imply any sort of atemporality, and negated EVENTs can still be temporally CONTAINed or otherwise temporally related. This is discussed further in CONTAINS.

Don't worry about double negation or phrase-level negation. In the phrase "She denies vomiting or nausea," [denies] is POS (as the denial did happen) and [nausea] and [vomiting] are both NEG because they didn't happen. Each EVENT should be considered on its own (rather than as part of a greater denial), and if the EVENT didn't happen, it's NEG, no matter what phrasing may have preceded it.

Finally, two specific situations are worth discussing here:

a) Clinical-sense "negative"

We should highlight that NEG means "did not happen" or "not true," rather than "negative" in the medical test result sense (usually meaning "shows no signs of cancer"):

(37) Her [colonoscopy] was [negative].

The colonoscopy here is still polarity POS (as it did in fact happen), and "negative" is simply telling us that no cancer was found. (So, note that [negative] here would also have a POS polarity! This is saying, "It's the case that no cancer was found.")

An actual polarity NEG colonoscopy would be something like:

(38) We were unable to perform a [colonoscopy] due to bad prep.

b) Reporting the absence of things

Consider the following example:

(39) No pain reported.

This sentence may be interpreted in a couple different ways:

a) The patient reports no pain

(pain = NEG, ACTUAL – pain doesn't exist)

Or:

b) The patient doesn't report pain

(pain = POS, HYPOTHETICAL – there might be pain but we don't know)

In order to make our annotations consistent, we've decided to follow the interpretation in (a) – pain should be NEG, ACTUAL. Furthermore, the evidential EVENT – reported – should be POS, ACTUAL, EVIDENTIAL.

The reasoning here is that while this sentence would be ambiguous out of context, we're reading it in the context of the doctor presenting relevant information he/she has learned upon meeting with and questioning the patient. It's highly unlikely the doctor would actually report a hypothetical pain event, unless it's explicitly stated and tied to some other related event, such as "Patient is comatose and therefore we don't know if she's experiencing pain." What they are going to report on are things they've explicitly discussed with the patient.

In summary, if the patient is the source of an evidential event about clinical symptoms, and if it's ambiguous as to whether the evidential event or the symptom event is the one being negated, follow the interpretation that the symptom is NEG and the evidential event is POS.

Degree

This feature is used in order to express an incomplete degree of an EVENT. In practice, degree is used as a companion to polarity, as a way of allowing us to say that something is "mostly" or "a little bit" true, rather than forcing us to call every EVENT 100 percent positive or negative. This allows greater nuance in our representation of EVENTs than POS or NEG generally allows.

Our three different degrees are N/A, MOST and LITTLE. N/A is the default value, used where there is no need to mark either of the other two degrees on the EVENT. These are used when there has been "a little" of an event, or a large (but not complete) change:

(40) There is a small amount of bright T1 [signal]$_{\text{LITTLE}}$

(41) Abdominal tenderness has nearly [disappeared]$_{\text{MOST}}$

(42) She feels slightly [weak]$_{\text{LITTLE}}$ but has resumed most of her normal activities.

Contextual Modality

Our schema has four contextual modalities: ACTUAL, HEDGED, HYPOTHETICAL and GENERIC.

ACTUAL

The first is ACTUAL, which is used most of the time and is the default option. The majority of EVENTs are ACTUAL, having already happened or being scheduled (without hedging) to happen.

(43) The patient’s new [tumor] is 3.5 cm from the epiglottis.

(44) The patient did not [report] [nausea].

(45) His anterior chest rash has not [reoccurred].

Note that ACTUAL is about whether it is a claim in the "real world," and so NEG events are usually ACTUAL as well.

HEDGED

EVENTS are marked as hedged when the doctor mentions a given EVENT with any sort of hedging, which can be lexical ("seems," "likely," "suspicious," "possible," "consistent with") or phrasal ("I suspect that...," "It would seem likely that"). These EVENTs are strongly implied, but, for safety, liability, or due to lack of comprehensive evidence, are not stated as fact by the doctor. As such, it’s very important that these hedged diagnoses and findings be included in the timeline, but be marked so that they can be easily differentiated from hard and fast diagnoses.

(46) Ultrasound findings were felt to be consistent with a T3, N1 rectal [tumor]

(47) An approximately 3-cm nodular region of intermediate T2 signal involving the body of the corpus callosum is suspicious for residual or recurrent [tumor] but appears unchanged from the patient’s prior examination.

(48) She has a rash not inconsistent with [measles].

(49) The patient may have undergone a mild [stroke].

Note that a doctor providing or commenting on evidence for a given finding does not qualify as hedging, so the following EVENTs would not be marked HEDGED:

(50) She [denies]$_{\text{ACTUAL}}$ [vomiting]$_{\text{ACTUAL, NEG}}$

(51) There is no [evidence]$_{\text{ACTUAL, NEG}}$ of [MS]$_{\text{ACTUAL, NEG}}$

But further active hedging can push this over into HEDGED:

(52) There is no concrete [diagnosis]$_{\text{ACTUAL, NEG}}$ of [MS]$_{\text{HEDGED, POS}}$, but given her [symptoms]$_{\text{ACTUAL, POS}}$, [it]$_{\text{HEDGED, POS}}$ seems extremely likely.

HYPOTHETICAL

The third modality is HYPOTHETICAL. This is useful when annotating diagnoses, theories, or other medically relevant but hypothetical events. Hypothetical EVENTs will often follow "if" statements ("If X happens, then we'll use Y to treat Z") or other sorts of conditionals ("Depending on the patient's response, we might treat A with B or with C").

(53) I've warned the patient that this new medication may cause peripheral [edema].

(54) We suspect either [achalasia] or [pseudoachalasia] here.

(55) If she experiences a [fever], we'll [treat] [it] on an outpatient basis.

It's worth noting that an EVENT occurring in the future does not imply that the EVENT is HYPOTHETICAL (although most hypothetical EVENTs will be AFTER DOCTIME). Although it's true that there's always a degree of uncertainty with anything happening in the future, HYPOTHETICAL marks explicit uncertainty in the text and should not be used just to indicate this future-uncertainty. AFTER, ACTUAL EVENTs can be thought of as future EVENTs for which we are given no reason by the author to think they might not occur. For instance:

(56) We may [recommend] to [resume] the [Cipro] and [Flagyl] and obtain a [CT] of the chest, abdomen and pelvis.

In (56), all the bracketed EVENTs are HYPOTHETICAL and have a DocTimeRel of AFTER.

(57) The patient's [myringotomy] will take place on Friday.

Here, [myringotomy] has a DocTimeRel of AFTER, but is ACTUAL as the text gives us no reason to believe it might not happen.

In addition, polarity and modality of EVENTs don't interact, even though one might expect them to. It's true, at least from a real-world point of view, that a HYPOTHETICAL EVENT, by definition, hasn't happened. However, polarity shouldn't change on that basis alone. There are POS-polarity HYPOTHETICAL EVENTs ("She might develop a rash") as well as NEG-polarity ones ("If this medication works, he will have no soreness").

Although many human languages feature such an interaction, in our schema modality does not interact with polarity or DocTimeRel. Future EVENTs are assumed to be actual unless stated otherwise, and it's possible (even common) to have an ACTUAL negated EVENT. Future or negated do not automatically mean "hypothetical."

GENERIC

GENERIC is our fourth contextual modality. It's used for EVENTs which are only mentioned in a general sense and do not appear on the patient's timeline of treatment. These usually occur when the doctor is justifying decisions, rationalizing a change in care, or simply covering his or her back, and often you'll get several sentences of "general discussion" which could have as easily come from a textbook as a medical record.

All of the below EVENTs would be marked GENERIC under our schema:

> (58) Adjuvant [chemotherapy] following [surgery] is generally recommended in situations similar to this.
>
> (59) I explained that BRAF [mutations] have no predictive value with regard to cetuximab [sensitivity].
>
> (60) In other patients without significant [comorbidity] that can [tolerate] adjuvant [chemotherapy], there is a [benefit] to systemic adjuvant [chemotherapy].

Although HYPOTHETICAL and GENERIC may seem similar, remember that most HYPOTHETICAL EVENTs still refer to the specific patient's care, but depend on some eventuality occurring, whereas GENERIC EVENTs could just as easily appear in any patient's note or a journal paper.

Finally, if an EVENT is GENERIC, DocTimeRel will always be OVERLAP as stated truths are, presumably, true at Document Time.

Aspect

Aspect is used to express aspectual ideas about the events which are not coded explicitly with aspectual EVENTs and ALINKs. We have three values for contextual aspect in the schema: N/A, INTERMITTENT and NOVEL. Please note that this is unrelated to grammatical aspect, and these two aspects give information about the temporal relations in the document, not about the grammatical forms used to express them.

N/A

N/A is the most common value (and our default value) for contextual aspect, and simply represents that a given EVENT is neither NOVEL nor INTERMITTENT. If neither of the other contextual aspects seems to fit for a given EVENT, leave the contextual aspect field blank (which will then be auto-filled with N/A).

NOVEL

NOVEL indicates, well, novelty, and is associated with predicate adjectives like "new."

(61) The patient’s new [tumor] is 3.5-cm from the epiglottis.

(62) The newest [MRI] revealed a previously undiscovered mass.

(63) She’s experienced unusual soreness around her new [stitches].

INTERMITTENT

INTERMITTENT is used in situations where there may be a series of smaller events within a single EVENT, rather than a single, constant event. Those events are usually marked with words like "intermittently" or "occasionally.” These indicate that, for instance, the patient has had vomiting since a certain time, but he/she has not been vomiting 24/7 since that point.

It’s important to note that we are only marking INTERMITTENT when there is an explicit mention of intermittency in the sentence. Even if you happen to know that a given disorder or symptom often manifests intermittently, if it’s not stated explicitly as doing so, you should not mark it as such.

If you are unsure about the contextual aspect of a given EVENT, mark it as N/A.

(64) He reports occasional bright red [bleeding] from the rectum.

(65) Patient complains of intermittent chest [pain].

(66) She’s had intense [headaches] on and off since her last visit.

Permanence and Maxspan

These properties will appear with an autofilled value and should be ignored.

TIMEX3s

TIMEX3s, or temporal expressions, are definitive references to time throughout the document or section. Examples of these might be phrases like "today," tomorrow," "24 hours ago," and "early March." Specific dates and times are annotated as TIMEX3 objects as well.

Our approach to marking TIMEX3s is identical to that used in ISO-TimeML, and these guidelines are heavily based on the standard established in Pustejovsky et al., 2010.

As with MARKABLEs, we use maximal-span annotation for TIMEX3s. Syntactically speaking, this means all TIMEX3 annotations should correspond to a:

- Noun phrase ("this weekend," "tomorrow," "yesterday," "Tuesday," "Last May," "May 16th," "6/9/1985")
- Adjective phrase ("two-hour-long," "half-hour" as in "a half-hour trip," "preoperative," "post-partum")
- Adverbial phrase ("lately," "recently," "shortly," "hourly," "intraoperatively")

Importantly, this means that any prepositions which precede (or in some cases, follow) a temporal expression are to left unmarked, even when they seem to provide additional context for interpreting the TIMEX3. For example:

(67) During {the month of July}, she will come visit.

(68) From {May 1st} to {the 3rd}, she will refrain from eating solid food.

(69) After {tomorrow}, she should be OK to walk with crutches.

(70) He'll come in for a follow up {two days} after his surgery, and again {the next week}.

These prepositions (referred to as SIGNALs in ISO-TimeML), although important in the interpretation of the meaning of the temporal expressions, provide separate temporal information which will be automatically extracted later.

Note that post-expression adverbs (often “ago”) are still captured in our spans, so:

(71) Patient s/p lumpectomy {2 yrs ago}

For the most part, if you have two separate temporal expressions, they’ll be two separate TIMEX3s, but two adjacent TIMEX3s which together specify a single time can be treated as a single span. Look at the contrast in (72):

(72) a. I’ll come by to check on her at {3:30pm Tuesday}.

b. I’ll come by to check on her at {3:30pm} and on {Tuesday}.

In (b), we know that the doctor is referring to two different timepoints, so we mark two TIMEX3s, whereas in (a), the “3:30pm” and “Tuesday” combine to specify a single timepoint.

The only exception to this rule is when the two temporally connected TIMEX3s are syntactically separated, as below:

(73) On {Tuesday}, I'll come by to check on her at {3:30pm}.
(74) She'll come in for another consult {the day} after {tomorrow}.

In these cases, mark two separate TIMEX3s, even though they combine to specify one timepoint. The sole exception to this rule comes with long-form times:

(75) Mr. Mullins arrived at {5 minutes to five}.

But in medical texts, these are extremely rare.

Finally, as one might imagine, TIMEX3s which are separated by a conjunction are to be treated separately as well:

(76) She should be fine for discharge on {Tuesday} or {Wednesday}.

(77) He will come in on {the 1st} and {the 5th} for followups.

We have six different classes of TIMEX3s. Note that in the examples below, if there are multiple TIMEX3s, only those which are of the indicated class will be marked.

DATE

The majority of TIMEX3 annotations you make will be of the class DATE. DATE represents dates. These can be calendar dates (such as {January 4}) and other verbal expressions which can be mapped to calendar dates either concretely (such as {Last week}, {This month}, {next Friday}, or {this time}), or in a more fuzzy sense ({lately}, {the past}).

(78) MRI of the brain without and with gadolinium contrast utilizing tumor followup protocol compared with prior studies of {29, February 2005} and {28, January 2005}.

(79) His anterior chest rash has not reoccurred since the PCN VK was discontinued {24-hours ago}.

(80) At {this time}, we see no reason to discontinue the treatment.

(81) The last cyclosporine level was 373 in {January}. His dose was adjusted downward to 300-mg twice-daily. A cyclosporine level will be repeated on {Friday morning}.

(82) I stated that if there is no other evidence of any disease recurrence, in {approximately one-month's time} we would proceed with approximately six-months worth of adjuvant therapy.

(83) The form was signed and scanned on {December 29, 2009}.

(84) She came in {a couple months ago}.

(85) Mr. Zegler was seen in the Hamilton University Medical Center with Dr. Carr {the middle of December 2009}.

(86) Carotid artery disease, last checked {greater than two years} prior.

(87) After {next week}, we'll see where her pain level is.

(88) {June 6, 2008} through {October 6, 2008}, treated with FOLFOX

As mentioned above, DATE is also used for very generic sorts of TIMEX3s, where you may not be able to point at a specific day, week, or month on a calendar, but can still gesture at the overall timeline. So, for instance, the following expressions would be TIMEX3s of the type DATE:

(89) She has experienced heavy bleeding in {the past}.

(90) She complains that she's felt tired {lately}.

(91) {Recently}, she has had several episodes of syncope.

TIME

TIME is used for specific time points within a day, for instance, {3:00PM} or {23:45}, and once again can be relative (as in example 93):

(92) The patient's MRI is scheduled for {5:30pm}

(93) Following the patient's latest seizure, {20 minutes ago}, we are re-evaluating her medications.

(94) Surgery will need to be completed by {2:45} to have the biopsy to the lab sooner.

Put differently, temporal expressions which give minute-by-minute or hour-by-hour detail are marked as TIME. Day-by-day (or larger) detail is marked with DATE.

DURATION

Sometimes, you'll be given a single temporal expression interpreted as reflecting a span of time, rather than a point. These are things like "for {24 hours}" or "All of February," and these are marked with the class DURATION.

(95) The patient continuously experienced nausea for {nearly two weeks}.

(96) For {the next 12 hours}, we will lower the patient's morphine drip and then we will re-evaluate his pain.

(97) During {the last 12 months}, she’s been nauseous frequently.

(98) In {the last week}, his pain has significantly worsened.

(99) He has been doing this for {five years}.

(100) I stated that if there is no other evidence of any disease recurrence, in approximately one-month’s time would proceed with {approximately six-months} worth of adjuvant therapy.

(101) In {the time} between now and the 15th, she should attend physical therapy whenever possible.

Note that in (101), both {now} and {the 15th} would also be TIMEX3s, but of type DATE. Remember again that more abstract temporal expressions ("lately," "in the past," "in the future"), although representing loosely defined spans of time, are considered DATE rather than DURATION, as they are predominantly only bounded at the start or the end, not both, as above.

Finally, remember that two dates can be used to construct a duration, but, because each represents a single point in time (rather than duration), both will still be labeled DATE, rather than DURATION:

(102) From {May 1st}$_{\text{DATE}}$ to {the 3rd}$_{\text{DATE}}$, she will refrain from eating solid food.

QUANTIFIER

Although it may seem odd at first, expressions like "Twice," "four times," and "18 times in the month of May" are all TIMEX3s. These are annotated with the QUANTIFIER class.

(103) The patient vomited {twice} before the surgery.

(104) We have seen Mr. Lastname {three times} for his ulcerative colitis.

(105) On {two to three incidents} she has had blood in the stools.

Note that QUANTIFIER only applies for number of occurrences of an EVENT, not for quantifying objects, like "She has two eyes" or "She [lost] 5 units of blood."

QUANTIFIERs should not be linked to anything.

PREPOSTEXP

Similarly odd, pre- and post- expressions ("preoperative," "post-exposure," "post- surgery," "prenatal," "pre-prandial") all actually designate specific temporal spans ("The time before the surgery," "The time after exposure") related to an implicit EVENT, and thus are TIMEX3s, marked with the class PREPOSTEXP.

> (106) Patient underwent a partial hemicolectomy in July 2009. {Postoperative} scarring noted during exam.
>
> (107) The patient exhibits {post-exposure} changes.

These will not always begin with "pre-" or "post-". Terms like "intraoperatively" can sometimes be PREPOSTEXP as well:

> (108) {Intraoperatively}, there were no difficulties securing his airway.

And sometimes, you'll have bare expressions which clearly express the PREPOSTEXP meaning, but don't contain the whole expression:

> (109) Pulmonary recommendations for {pre}, {peri} and {postop} were made.

SET

SET is used exclusively in our schema for covering expressions which give both a quantifier and an interval (like "Three times weekly," "monthly," or "1/day") and represent a frequency. This is different from QUANTIFIER ("twice") which only gives a quantifier, and different from DURATION ("all week") which only gives a span.

Even though most sets could be interpreted as a QUANTIFIER and a DATE/TIME juxtaposed, we should mark them as a single span ("{twice a month}" rather than "{twice} {a month}").

> (110) Will administer Lariam {twice daily}.
>
> (111) Patient has checked into the ER {roughly three times a month}.
>
> (112) We will proceed with {weekly} consultations to monitor the patient's condition.
>
> (113) Mirtazapine REMERON 7.5-mg tablet 1 tablet by mouth {every bedtime}.
>
> (114) Simvastatin ZOCOR 20-mg tablet 1 tablet by mouth {one-time daily}.

TIMEX3s of type SET should always be TLINKed to EVENTs using the TLINKs of the type OVERLAP.

2 Relations

You've now been introduced to all the different markables that you'll see in the data: EVENTs, MARKABLEs (entities), TIMEX3s, SECTIONTIMEs, and DOCTIME.

These markables have also already been linked to each other with a variety of relations, which fall under two categories: **temporal relations** (when an EVENT or TIMEX3 occurs relative to another one in time) and **coreference relations** (whether one markable refers to the same thing as another markable, or is somehow part of it or an instance of it). Some of the following language is instructive, which makes it sound like you'll be creating these from scratch; that's not the case, but knowing some of the guiding rules that went into creating these relations initially will be helpful for the current task.

General rules for these relations:

- Temporal relations are always only used for EVENTs and/or TIMEX3. MARKABLEs are never involved in temporal relations.
- Coreference relations are used for both EVENTs and MARKABLEs (not TIMEX3s), except for the WHOLE-PART relation, which is only for MARKABLEs.
- MARKABLEs and EVENTs should never be in the same link, no matter the type.

Temporal relations

By "temporal relations," we're referring to a relatively limited set of timeline relations between EVENTs and EVENTs, or between EVENTs and TIMEX3s, referred to as TLINKs ("Temporal Links"). These relations exist for one reason: to allow us to arrange and order the various EVENTs and TIMEX3s accurately on a timeline.

Although these temporal links are annotated in a more interactive fashion, we will display them in this document using the following format:

(115) [EVENT1] RELATION [EVENT2]

where RELATION is BEFORE, CONTAINS, OVERLAP, BEGINS-ON, or ENDS-ON, as determined by the type of relation being stated. These types are described in detail in TLINK annotation.

To give a realistic example, imagine the following sentence:

(116) The patient was to follow-up with oncology this month.

The EVENT [follow]-up (note that only the head is marked for phrasal verbs) is clearly temporally related to the TIMEX3 {this month}, because [follow]-up will occur during {this month}, giving you more specificity than the follow-up just occurring after document time. So here you would create a TLINK annotation, insert [follow] into the "target" slot, {this month} into the "source" slot, and select CONTAINS as the relation. By doing so, {this month} is established as a narrative container anchor, which may contain additional EVENTs later on in the note. So, the finished annotation would look like:

(117) The patient was to follow-up with oncology this month.

a. {this month} CONTAINS [follow-up]

These temporal links (along with the rarer "aspectual link" discussed in ALINK annotation) are eventually used to reconstruct the patient's timeline.

But before we discuss the specifics of these temporal links, we need to discuss our strategy for annotation, and a crucial metaphor which, once understood, will yield a higher quality of annotation, while saving us a great deal of time and effort.

Narrative containers

Temporal structure of a narrative is often hierarchical. If, on a particular Tuesday, you go on a series of errands, one of which is going to the store, and during that visit you get milk, you could say that that Tuesday temporally contains the errands; that they temporally contain your trip to the store; and that the trip to the store temporally contains the event of getting milk.

The importance of that is that if someone were to know that an event happened before Tuesday, they then inherently know that it also happened before all of Tuesday's events.

While we build narrative containers using this temporal relation CONTAINS, the idea of "Narrative containers" is not simply the idea of that relation, but this goal of using these increasingly large temporal "buckets" to cleanly capture the temporal essence of a passage in an informative manner (as detailed in Pustejovsky and Stubbs 2011 and Styler et al., 2014).

A narrative container can be thought of as a temporal bucket into which an EVENT or series of EVENTs may fall. These narrative containers are often represented (or "anchored") by dates or other temporal expressions, but may also be anchored by a reference to an EVENT capable of containing another EVENT – a surgery might contain an incision, or a war may contain battles. By focusing on placing events in progressively larger temporal buckets, you eliminate the need for most relationships between individual events – **if container A is BEFORE container B, we know that all events inside A are before the events in B, without needing to make those annotations manually**. In the example above, if you crashed your car before Tuesday (therefore forcing you to do all of Tuesday's errands by bike), you wouldn't have to say that you crashed your car BEFORE getting milk, and BEFORE the trip to the store, and BEFORE running errands; we can infer all those temporal relations simply by knowing that you crashed your car BEFORE Tuesday, the narrative container in this case.

As such, rather than marking every possible temporal relation (TLINK) between each EVENT, we instead try to link as many EVENTs as possible to a narrative container, and then

link those containers so that the contained EVENTs can be linked by inference.
Here's a clinical example:

(118) {December 28th}: The patient experienced a [stroke] at {approximately 9:30am}, during her [surgery].

a. {December 28th} CONTAINS [surgery]

(i) [surgery] CONTAINS [stroke]

(ii) [surgery] CONTAINS {approximately 9:30am}

aa. {approximately 9:30 am} CONTAINS [stroke]

Here, we have an overarching container, {December 28th}, which CONTAINS [surgery]. The surgery then CONTAINS both [stroke] and {approximately 9:30am}. Then, to complete the annotation, we'd mark that {approximately 9:30am} CONTAINS [stroke]. Note there is no need to say {December 28th} CONTAINS [stroke], because it is clearly contained within it by virtue of its container being contained within it.

Finally, while TIMEX3s and EVENTs can be temporal containers, we also have another kind of temporal container: the DocTimeRel itself. {December 28th} above is therefore a temporal "bucket" that is within a larger bucket of BEFORE – the set of all events happening before the document creation time. Importantly, we do not need to create TLINKs between EVENTs that have different DocTimeRel, as those links can all be easily inferred by our knowledge of the ordering of the containers.

Choosing the anchors of narrative containers

When creating narrative containers as discussed above, you will need to choose either a TIMEX3 or an EVENT to be the "anchor" of the container, the temporal span which all of the other EVENTs fall within (or begin/end on). Choosing which temporal span to be the anchor of a given narrative container can be difficult, so here are a few ground rules to help make these decisions easier and more consistent:

- The majority of TIMEX3 annotations will be narrative container anchors.
- If you have a choice between using an EVENT or a TIMEX3 as the narrative container anchor, you should pick the TIMEX3.
- If you use an EVENT as a narrative container anchor, try to TLINK it to a few other container anchors to avoid it being stranded. (Note this will not always be possible.)
- If stuck between two possible containers, use the one with the larger temporal span.

Ordering within narrative containers

Because of the difficulty of capturing detail within a given narrative container, not all relations between EVENTs will be captured. If a patient undergoes a preoperative [evaluation], and that [evaluation] CONTAINS a [CT] and an [x-ray], but the text doesn't specify the order of the two tests, it's fine to leave the two unlinked (and in fact you should leave them unlinked, rather than guessing). We don't have to worry about relations that aren't present in the note; we're just concerned with capturing the relations that are present. (Naturally, if it is clear which test came first, then you should create a BEFORE link between the two.)

So, although it may seem like some of the narrative containers that you define may be a loose bucket of temporal EVENTs, that's not actually a problem, as the greater timeline of the patient's care is more important than the fine structure within a given narrative container, and as you'll find out, there are still plenty of TLINK annotations to be made. Of course, if a relation is explicitly stated, it should always be marked.

TLINK annotation

TLINKs, as previously mentioned, are relations you can mark between EVENTs and EVENTs, or between EVENTs and TIMEX3s, to show the temporal relations present within the document and to clearly define the bounds of the narrative containers at work beyond what DocTimeRel will naturally give us.

Before we talk about exactly when to use these TLINKs, we'll first discuss the different types of TLINKs used in this schema.

There are five different temporal relations used in this schema: BEFORE, OVERLAP, BEGINS-ON, ENDS-ON and CONTAINS. There is no default relation type for TLINKs.

BEFORE

BEFORE is fairly straightforward; it simply orders two events in time.

(119) She [vomited] shortly before [surgery].

a. [vomited] BEFORE [surgery]

(120) His anterior chest [rash] has not [reoccurred] since the [PCN] VK was [discontinued] {24-hours ago}.

a. [discontinued] BEFORE [reoccurred]$_{\text{NEG}}$

(b. [rash] ENDS-ON [discontinued] – this link discussed below)

(c. {24-hours ago} CONTAINS [discontinued] – this link discussed below)

([PCN] will be linked to [discontinued], using an ALINK of the type TERMINATES, as described in TERMINATES.)

When annotating, remember that "X occurred after Y" can be expressed by saying "Y occurred before X":

(121) She was [seen] by Dr. Jones in cardiology following the stent [placement].

a. [placement] BEFORE [seen]

(122) He had a [neckache] after [surgery].

a. [surgery] BEFORE [neckache]

CONTAINS

CONTAINS signals that the EVENT is completely contained within the temporal span of the EVENT or TIMEX3 it's related to. In other words, the contained event occurs entirely within the temporal bounds of the event it's contained within.

CONTAINS is a very specific relation implying complete containment within a narrative container, and if you annotate that X CONTAINS Y, it's assumed that there's also an OVERLAP relation between the two. You should only use CONTAINS when you're sure that the nature of the overlap is one of complete containment.

(123) {March 2005} - Patient underwent [appendectomy]

a. {March 2005} CONTAINS [appendectomy]

(124) [Levaquin] 750 mg p.o. q. day will [restart] {today}

a. {today} CONTAINS [restart]

(125) [Comparison] is made with prior MRI head [examination] without and with gadolinium from {10-23- 03}.

a. {10-23-03} CONTAINS [examination]

(126) An ENT performed the [myringotomy] during {Friday}'s [surgery].

a. [surgery] CONTAINS [myringotomy]

b. {Friday} CONTAINS [surgery]

In addition, we have made one specific regulation involving the use of CONTAINS[5]: All test results or observations are to be linked to the test which generated them (their narrative

[5] Note: This annotation rule will be changed in a manual post-processing step by the current project, such that all *test-CONTAINS-results* TLINKs became *results-NOTED-ON-test*. See Appendix C.

container) using a CONTAINS relation, assuming there is actually temporal containment. This is not always intuitive, because as humans with some inferencing ability, we realize that the tumor likely existed before the CT scan which revealed it, but from a machine-learning perspective, it's important to have that consistency. So:

(127) [Colonoscopy] ({January 7, 2010}): Diminutive [polyps] of the rectosigmoid, [removed].

a. {January 7, 2010} CONTAINS [Colonoscopy]

b. [Colonoscopy] CONTAINS [polyps]

c. [Colonoscopy] CONTAINS [removed]

There are many relations which seem like a sort of semantic containment (things like part/whole, cause/effect, disorder/symptoms). However, the CONTAINS relation should only be used when there exists strict temporal containment (the temporal span of the container fully encompasses those of the contained).

Note that we can still temporally link negated EVENTs to other EVENTs or TIMEX3s, via CONTAINS or other relations:

(128) Ms. Patton [recovered] from her surgery without any $[\text{complications}]_{\text{NEG}}$.

a. [recovered] CONTAINS $[\text{complications}]_{\text{NEG}}$

On the one hand, it might seem odd to say that a non-EVENT is temporally contained by another EVENT – the complications never did occur at all, so how could they be temporally bound by the recovery? On the other hand, when an author suggests that there is some kind of significance or relevance between a negated EVENT and another EVENT, we want to capture that relation.

Pragmatically speaking, we don't temporally enjoin negated activities unless relevant or exceptional. If one asks a traveler about where they are and they reply "Not in Washington," the implication is that they should be in Washington, but due to whatever unpleasant or unexpected situation, they're not. Thus, the current state of not being in Washington is mentionable, exceptional, and specific, whereas in the rest of their life up until that point, they hadn't felt the need to bring up the fact that they're not in Washington.

Similarly, if a note says "no complications during recovery," or something to that effect, we argue that it represents an exceptional (and thus, mentionable) state of not having complications during the mentioned period, and that exceptional state itself is temporally contained by the recovery.

OVERLAP

OVERLAP is a single temporal relation that encompasses all the different notions of two things happening at the same time, but is less specific than CONTAINS. This can refer to two simultaneous events, an EVENT that occurs during another, larger EVENT or time reference (but where containment is not entirely sure), or any other sense in which two events are occurring in the same timeframe:

(129) The patient had some rectal [itching] and mild [pain] {today}, mostly {this morning}.

a. {today} CONTAINS [itching]

b. {today} CONTAINS [pain]

c. {this morning} OVERLAP [itching]

d. {this morning} OVERLAP [pain]

(130) He does have a history of peri-rectal [abscess] with his last round of [chemotherapy].

a. [chemotherapy] OVERLAP [abscess]

In short, OVERLAP is meant for situations where two events overlap in some way, but where you're not sure (or don't have enough information to tell) whether there is containment.

OVERLAP is also used for linking TIMEX3s of type SET with other EVENTs:

(131) We'll keep her on rate-control [medications] 100 mg {twice daily}

a. {twice daily} OVERLAP [medications

OVERLAP provides relatively little information for the actual processing of text, especially compared to CONTAINS. If you're reasonably sure that the relation is one of containment, you should use CONTAINS instead, and often, one can represent a potential OVERLAP more specifically using narrative containers and a bit of additional thought.

BEGINS-ON

BEGINS-ON signals that the EVENT begins on the EVENT or TIMEX3 it's related to. This type of TLINK will only occur with EVENTs which have a non-trivial temporal span. Relations with punctual EVENTs will usually be marked with BEFORE instead.

(132) She has had Abdominal [Cramping] since {January}.

a. [Cramping] BEGINS-ON {January})

(133) He reports intermittent chest [pain] since his prior [MI].

a. [pain] BEGINS-ON [MI]

ENDS-ON

ENDS-ON signals that the EVENT ends on the EVENT or TIMEX3 it's related to. As with BEGINS-ON, this type of TLINK will only occur with EVENTs which have a non-trivial temporal span. Relations with punctual EVENTs will usually be marked with BEFORE instead.

(134) She has had no [bleeding] since her [stitches] were [removed].

a. [bleeding]$_{NEG}$ ENDS-ON [removed]

b. [stitches] ENDS-ON [removed]

Note that ENDS-ON can be used in concert with BEGINS-ON to mark a duration:

(135) She was on [chemo] from {March} through {July}.

a. [chemo] BEGINS-ON {March}

b. [chemo] ENDS-ON {July}

When to TLINK

TLINKs themselves are relatively straightforward, and in fact, the more difficult part of annotating TLINKs is to know when to stop. Without any constraint, one could see making TLINKs between every EVENT in the document, which leads to exponential growth of TLINKs and a tangle of relations which nobody, let alone a machine, would like to unpack.

So, to constrain this process a bit, we have developed a few rules to govern TLINKing:

- **TLINK only when it captures more information than just marking DocTimeRel**

Because the DocTimeRel attribute of EVENTs expresses the relation of the event to the time the document or section was written, you will never need to TLINK to the DOCTIME annotations. Marking an EVENT as "after" in the DocTimeRel field gives us the same information as making an BEFORE TLINK between the EVENT and DOCTIME, so you need not explicitly mark that. Similarly, if one EVENT has a DocTimeRel of OVERLAP and another has a DocTimeRel of AFTER, there's no need to make a TLINK between those two EVENTs.

- **TLINK all EVENTs to their narrative container, if possible**

As previously discussed, most EVENTs will fall into a narrative container of some kind. If a given EVENT is in a narrative container (like "August 22nd" or "during her recovery"), you should always TLINK that EVENT to the TIMEX3 or EVENT which represents that narrative container, using the appropriate link. Once again, though, this should only be done if the result will be more informative than just analyzing the DocTimeRels. See the following example sentence and the TLINKs required to annotate it:

(136) {December 19th}: The patient underwent an [EKG] as well as emergency [surgery]. During the [surgery], the patient experienced another [MI], and repeated [tachycardia].

a. {December 19th} CONTAINS [EKG]

b. {December 19th} CONTAINS [surgery]

c. [surgery] CONTAINS [MI]

d. [surgery] CONTAINS [tachycardia]

Not every EVENT will be a part of a detailed narrative container (that is, not every EVENT will be linkable to a more specific narrative container than the one already identified by its DocTimeRel). However, it's vital that you, as an annotator, stop to ask yourself whether each EVENT you examine is a part of a narrative container, and whether the TLINKs you created are sufficient to mark that membership.

- **Try to only link EVENTs and TIMEX3s within the same sentence**

In a perfect world, nearly all TLINKs would occur across two EVENTs or TIMEX3s in the same sentence. That said, often you need to link to an EVENT or TIMEX3 in a previous sentence to put an EVENT in the proper narrative container. In these situations, you may do so, but you should double-check to ensure that there's no other way of going about it, and remember that coreference annotation will be done to link pronouns and subsequent mentions, so linking an EVENT to a subsequent reference to the narrative container is acceptable as well. That said, because of the nature of the notes, TLINKs should never link items in different sections.

(CONTAINS-SUBEVENT TLINKs, which we are adding in the current task, are an exception to this. They're discussed below, in Adding CONTAINS-SUBEVENT links.)

- **ACTUAL or HEDGED EVENTs should never be linked to HYPOTHETICAL or GENERIC EVENTs, and vice versa**

This may seem quite specific, but HYPOTHETICAL and GENERIC EVENTs aren't really on the same timeline as other EVENTs dealing with the patient's actual care. Because of this, it can be quite tricky to link them to the overall timeline, and we need to take care to avoid having non-real EVENTs showing up on patient timelines. To avoid the complications and potential broken relations when non-real EVENTs are pruned from timelines, ACTUAL or HEDGED EVENTs should never be linked to HYPOTHETICAL or GENERIC EVENTs, and vice versa. In this way, "real" EVENTs are never linked to non-real ones.

(137) Adjuvant [chemotherapy] following her upcoming [surgery] would generally be recommended, but given her poor [health], this is not an option.

Here, because the [chemotherapy] is HYPOTHETICAL, it cannot be TLINKed to [surgery], even given the explicit mention.

- **Don’t TLINK TIMEX3s to one another**

Although there is certainly a temporal relation between, say, "January 15th, 2009" and "March 2013," part of the post-processing of these annotations is the normalization of these temporal expressions, which will allow us to order these expressions on a timeline. Although you will certainly still TLINK EVENTs and TIMEX3s, the TIMEX3s in the document can be temporally ordered without the annotator’s help.

ALINK annotation

The final sort of temporal annotation performed on the data is the ALINK (aspectual link) annotation. ALINKs are created between an aspectual EVENT and a non-aspectual EVENT. Any EVENT previously marked with the class ASPECTUAL will be ALINKed to another, non-aspectual event, and you will never make an ALINK which includes a TIMEX3 or two non-aspectual EVENTS. ALINKs are much less common in the text than TLINK annotations.

Unlike TLINKs, no matter the circumstance, an ALINK should never cross a sentence boundary. As with TLINKs, they are created interactively, but have the basic form:

(138) *EVENT1* [aspectual relation] *EVENT2*

Following are the four ALINK subtypes:

CONTINUES

CONTINUES is used when an aspectual event shows the continuation of another event:

(139) We will [continue] to [monitor] LFTs carefully along with his weight.

a. [continue] CONTINUES [monitor]

(140) The patient will [remain] on [dialysis] until her condition [changes].

a. [remain] CONTINUES [dialysis]

b. [remain] ENDS-ON [changes]

(141) She is not interested in pursuing chemotherapy at this time but is interested in [continued] [surveillance].

a. [continued] CONTINUES [surveillance]

INITIATES

INITIATES is used when an aspectual event indicates the start or initiation of another event:

(142) Patient will [begin] a high-fiber [diet] upon [release].

a. [begin] INITIATES [diet]

b. [release] BEFORE [begin]

(143) We will [start] Ms. Miller on a normal saline [infusion] at 75 an hour for a total of 1 L.

a. [start] INITIATES [infusion]

REINITIATES

REINITIATES is used when an aspectual event indicates that another event will be restarted or reinitiated:

(144) [Levaquin] 750 mg p.o. q. day will [restart] {today}.

a. [restart] REINITIATES [Levaquin]

b. {today} CONTAINS [restart]

TERMINATES

TERMINATES is used when an aspectual event indicates the ending of another event:

(145) We will [hold] her [heparin] until after the [surgery].

a. [hold] TERMINATES [heparin]

b. [hold] ENDS-ON [surgery]

(146) Patient [nausea] was successfully [stopped] by 1-mg [Ativan] p.r.n.

a. [stopped] TERMINATES [nausea]

(147) His anterior chest [rash] has not [reoccurred] since the [PCN] VK was [discontinued] {24-hours ago}.

a. [discontinued] TERMINATES [PCN]

b. [discontinued] BEFORE [reoccurred]$_{\text{NEG}}$

c. [rash] ENDS-ON [discontinued]

d. {24-hours ago} CONTAINS [discontinued]

Coreference relations

Just as you'll see temporal relations already present in the preannotated data, you'll also find coreference relations marked. Coreference relations can be thought of in two categories – relations that show that two things are the same (IDENTICAL and APPOSITIVE) and relations that show that they have a structural relationship (WHOLE-PART, SET-SUBSET). Coreference relations may occur between EVENTs and EVENTs or between entities and entities, with the exception of WHOLE-PART, which may only occur with entities. TIMEX3s should never participate in coreference relations. All four relations are discussed below.

Importantly, while TLINKs may only occur within-section (except CONTAINS-SUBEVENT links), coreference links can and should occur across the entire document.

These guidelines for coreference annotation are based on the original Clinical Coreference guidelines and the RED (Richer Event Description) guidelines, which in turn were adapted from the OntoNotes v. 7.0 and ODIE (2010) guidelines. Additional concepts and information were synthesized from Poesio 2004, Poesio and Traum 1997, and Savova et al. 2011.

IDENTICAL

Two markables have an IDENTICAL relation if they refer to the same discourse referent. The IDENTICAL relation has several important semantic characteristics (following the MUC-7 specifications):

- The relation is symmetrical, i.e., non-directional: If A is IDENT to B, then B is IDENT to A. This is different from relations like WHOLE-PART and SET-SUBSET, which are directional by definition.
- It is also transitive: If A is IDENT to B and B is IDENT to C, then A is IDENT to C. The canonical example of this, and perhaps the bulk of IDENT annotation work, will be the marking of pronominal antecedents:

  (148) [Mr. Smith]$_{M1}$ complained of a headache. [He]$_{M1}$ also had a sore throat.

Yet in many instances we will be linking together lexical nouns (or verbs, or other parts of speech) that refer to the same thing. Different terms that still refer to the same thing belong in the same chain – remember we are dealing with semantics, i.e., what things mean. So, when considering IDENTICAL relations, ask yourself, "Does this mention point to the same actual event or entity as this other mention?" Consider the following:

(149) Colonoscopy on September 12, 2013 showed a [tumor] consistent with [adenocarcinoma].

Often (though not always), the language “consistent with” introduces a hedged diagnosis. Note that for the phrase, “[tumor] consistent with [adenocarcinoma],” [tumor] and [adenocarcinoma] should be IDENT. This is because **the referent for both terms is the same**[6], even though the terms differ in how specific they are – both point to the presence of the abnormal growth in the colon.

Consider the non-clinical example:

(150) [The White House] has been undergoing remodeling. Contractors say [the building] is showing signs of age.

Clearly [the White House] and [the building] are referring to the same structure, even though one term is more general and the other is more specific.

Finally, as with temporal links, we permit linking ACTUAL and HEDGED EVENTs to each other with coreference links where appropriate, but neither of these should be linked to HYPOTHETICAL or GENERIC EVENTs.

APPOSITIVE

Two entities have an APPOSITIVE relation if two NPs having the same semantic meaning occur adjacent to one another, separated only by punctuation – almost always a comma, colon, dash, or parenthesis.

While actual linguistic analysis may consider many other things to be "apposition," our definition absolutely requires the presence of punctuation separating the two parts of an apposition construction. This is purely for the sake of consistency.

(151) I met [the President], [Barack Obama].

a. [the President] APPOSITIVE [Barack Obama]

(152) [My friend who works there], [John Smith], makes bread.

a. [My friend who works there] APPOSITIVE [John Smith]

APPOSITIVE is a directed relation with a HEAD and ATTRIBUTE. Unlike IDENT chains, which may contain many different referents in the same relationship, APPOSITIVE is a two-term relation with only two possible mentions, and the mentions have different roles, that of the HEAD and ATTRIBUTE.

[6] Following T1Coref’s interpretation of such terms: *We assume that cancer refers to the disease event and the carcinoma refers to the physical manifestation thereof* (T1Coref guidelines, p. 12)

If one of the two mentions is a proper name, it is automatically the HEAD. Otherwise, the leftmost term must be the HEAD and the right mention the ATTRIBUTE:

(153) I met [the President]$_{\text{ATTRIBUTE}}$, [Barack Obama]$_{\text{HEAD}}$

(154) If Mr. Sampson recovers well from surgery, we will proceed with [chemotherapy]$_{\text{HEAD}}$ ([FOLFOX]$_{\text{ATTRIBUTE}}$).

We will only consider the HEAD of an APPOSITIVE relation as eligible to participate in other relations besides the APPOSITIVE relation. Even if one of the mentions seems more contextually appropriate, you must use the HEAD. For example:

(155) Patient consulted with [three doctors]: [a radiation oncologist]$_{\text{ATTRIBUTE}}$ ([Dr. Carson]$_{\text{HEAD}}$), [a surgeon]$_{\text{ATTRIBUTE}}$ ([Dr. Mitchell]$_{\text{HEAD}}$), and [her primary oncologist]$_{\text{ATTRIBUTE}}$ ([Dr. Ashley]$_{\text{HEAD}}$), and a treatment plan was jointly decided upon.

a. [a radiation oncologist] APPOSITIVE [Dr. Carson]

b. [a surgeon] APPOSITIVE [Dr. Mitchell]

c. [her primary oncologist] APPOSITIVE [Dr. Ashley]

d. [three doctors]$_{\text{SET}}$ / [Dr. Carson]$_{\text{SUBSET}}$, [Dr. Mitchell]$_{\text{SUBSET}}$, [Dr. Ashley]$_{\text{SUBSET}}$

The SET-SUBSET relation is discussed further in the next section. The key point to recognize here is that the ATTRIBUTEs of the APPOSITIVE relations are not permitted to be part of that link.

A specific nominal construction that counts as APPOSITIVE is when there's any sequence of titles or degrees in which apposition is used, as in:

(156) [John Smith]$_{\text{HEAD}}$, [Pathologist]$_{\text{ATTRIBUTE}}$

(157) [Lane J. Carson]$_{\text{HEAD}}$, [M.D]$_{\text{ATTRIBUTE}}$

SET-SUBSET

A SET-SUBSET relationship exists when an entity or event can be thought of as one or several members of a larger group. As with all other coreference relations except WHOLE-PART, SET-SUBSET relations can be between entities or between EVENTS.

(158) Colonoscopy showed multiple [growths]$_{\text{SET}}$ throughout her colon. The endoscopist removed a small [polyp]$_{\text{SUBSET}}$ from the transverse colon and another small [polyp]$_{\text{SUBSET}}$ from the cecum.

Here, the individuated polyps are two of the growths in the patient's colon, so they are members of that group of growths. (Remember we interpret references to abnormal growths as the implicit event of having those growths, so the above mentions are EVENTs with minimal-span marking.)

Note that the term *SUBSET* does not need to be literally another set of things; it can refer to single members as it does here. Of course, it is possible that the SUBSET can itself be a group or another set; if the previous example text included another sentence like, "Patient has many [growths] throughout her body," the relations would be as follows:

- many [growths]$_{\text{SET}}$ - [growths]$_{\text{SUBSET}}$ throughout her colon
- [growths]$_{\text{SET}}$ throughout her colon - [polyp]$_{\text{SUBSET}}$, [polyp]$_{\text{SUBSET}}$

This example demonstrates two other properties of the SET-SUBSET relation:

- Any number of SUBSETs (members) may be included in a SET-SUBSET relation, although only one markable is allowed to fill the SET slot.
- SET-SUBSET relations can (and should) be nested where appropriate; there's no need to link [polyp] to "many [growths]," because that relation can be inferred from the fact that [polyp] is a member of "[growths] throughout her colon," which is in turn a subset of "many [growths].

A couple more examples:

(159) Patient has [two arms]$_{\text{SET}}$. [Her left arm]$_{\text{SUBSET}}$ is scarred.

(160) [Patients]$_{\text{SET1}}$ with [cancer]$_{\text{SET2}}$ are advised to avoid this medication. [Our patient]$_{\text{SUBSET1}}$ has kidney [cancer]$_{\text{SUBSET2}}$ and lung [cancer]$_{\text{SUBSET2}}$, so we will not prescribe it.

The second example demonstrates one final principle of SET-SUBSET relations, which is that we do permit linking GENERIC markables to ACTUAL/HEDGED markables for this relation only, when appropriate. As stated above, for all other coreference relations, like temporal relations, we follow the principle of not linking HYP/GEN to ACT/HEDGED events. SET-SUBSET is the sole exception. So in the above example, [Our patient] (ACTUAL), is a SUBSET of the SET of generic [Patients]. Same goes for the cancer-cancer mentions.

Finally, note that "SET" here represents a group of things, and is therefore an entirely different use of the term than the TIMEX3 subtype "SET" (which represents frequencies). They will always be marked differently in the files, so shouldn't be hard to differentiate.

WHOLE-PART

Two entities exist in a WHOLE-PART relationship if one can be thought of as part of the other, larger entity.

(161) [The hand]$_{\text{PART}}$ was broken when the boulder fell on [his arm]$_{\text{WHOLE}}$.

As with SET-SUBSET, any number of PARTs may be included in a WHOLE-PART relation, although only one entity is allowed to fill the WHOLE slot:

(162) Colonoscopy showed multiple growths throughout [her colon]$_{\text{WHOLE}}$. The endoscopist removed a small polyp from [the transverse colon]$_{\text{PART}}$ and another small polyp from [the cecum]$_{\text{PART}}$.

This relation can only be used for cases of compositionality, i.e., narrow readings. In order for an entity to be PART of another, it must be 100% contained by the WHOLE, and compositionally part of it. Thus, the following does not contain any WHOLE-PART relations:

(163) Mt. Everest lies on the border between China and Nepal.

We do permit some more abstract WHOLE-PART relations, such as between an organization or department and a person who is part of that department:

(164) [[Dr. Castillo's]$_{\text{PART}}$ service]$_{\text{WHOLE}}$

(165) We will discuss with [Dr. Cho]$_{\text{PART}}$ ([Pathology]$_{\text{WHOLE}}$).

(Note this is not an appositive relation! The noun-phrase referents are different, even though the syntax fits the APPOSITIVE criteria.)

Also, note that like SET-SUBSET links, WHOLE-PART relationships should be nested where appropriate (the iris is part of the eye, the eye is part of the head, etc.).

An exception to WHOLE-PART links for these notes is that anatomical sites will not (and should not) be linked to the person to whom they belong.

Finally, remember that while IDENT, SET-MEMBER and APPOSITIVE will also be used to mark EVENTS, **WHOLE-PART is only for entities**. Relationships between EVENTs that you are tempted to mark as WHOLE-PART should probably be CONTAINS- SUBEVENT (see first pass annotation guidelines) or SET-SUBSET instead.

III Converting THYME 1 to THYME 2

Now that you know what you're looking at when you see a preannotated file, we can focus on the task at hand. As described in the introduction, the current project consists of two annotation stages:

1) Data Synching

The preannotated data consist of annotations made by two previous projects that have been automatically merged. This pass resolves discrepancies in the merged data occurring as a result of: a) complications from the automatic merging task; b) conceptual variations between the two prior projects in the treatment of certain linguistic phenomena. In this pass, we'll also complete annotation of medications and allergies sections, and add one new relation (CONTAINS-SUBEVENT), which will streamline cross-document annotation.

2) Cross-document coreference linking

Finally, we'll create coreference links (IDENTICAL, SET-SUBSET, and WHOLE-PART) across all three notes for a given patient, as well as one temporal link (CONTAINS-SUBEVENT).

1 The first annotation stage: Data synching

This annotation stage will take place on a single document at a time (there is no cross-document component to this pass). In this pass, annotators will complete three tasks:

1) Marking entity singletons

2) Creating CONTAINS-SUBEVENT temporal links

3) Reconciling the merged gold data from THYME and Coref

This pass is task-heavy and there's a lot of necessary overlap between the tasks. Take your time. Appendix B contains a checklist of these tasks as well as a summary of the most important annotation rules, and may be used for quick reference once you're thoroughly acquainted with the guidelines.

We are also resolving some project-internal inconsistencies and errors in this pass. Because these tend to overlap with the synching adjustments, and because it's not necessary to distinguish between the two sets of issues, fixing these errors will be rolled into the third task, i.e., reconciling the merged data.

Marking singleton entities

The T1Temp project annotated EVENTs but not entities. The T1Coref project annotated only those EVENTs, entities, and temporal expressions that participated in a coreference relation with something else in the document. Therefore, the notes include some entities that are not yet marked – singleton entities, that is, entities that don't have a coreference relation.

Because the next phase of this project is cross-document coreference, we now want to mark these entities. Even though they are not part of a within-document link, they may participate in a cross-document coreference link.

We'll use the featureless MARKABLE to annotate these new singleton entities. Entities should fit your traditional understanding of entities – non-eventive markables such as people, places, and anatomical sites; referential things that you could not easily put on a timeline as something that "happened." Refer back as needed to Markables.

(166) Pathologist identified extension of the tumor into the pericolonic fat.

In this example, both "Pathologist" and "the pericolonic fat" represent singleton entities that should be marked during this pass.

Naturally, you don't need to check for coreference relations for these markables, as we can assume that was done by T1Coref and none were found – which is why they're currently unmarked.

Entities that have been de-identified to nonsense strings rather than to real names or terms don't need to be marked at all:

(167) When completed, results will be available in DBCZLH.

You should mark entities that don't exist and entities that used to exist:

(168) Patient has [no left arm] as a result of a farming accident several years ago.

However, we don't mark metaphorical entities, like "truck" in the following:

(169) Patient said fatigue made her feel like she had been hit by a truck.

A word about addresses: You do not have to mark any entities in stand-alone addresses, but if an address-related entity plays a role in a sentence, it should be marked. Therefore, (170) has no markable entities, but (171) does, as shown:

(170) Mattson General Hospital
58776
Florida, Georgia

(171) Patient will return to [Houston] for adjuvant therapy..

And a word on premodifying entities (like "Dr. Castillo's" in "Dr. Castillo's service" or "liver" in "liver lesions"):

- We don't need to mark these. The T1Coref project only marked premodifying entities if they had an IDENTICAL head noun reference elsewhere in the document. For consistency's sake, we're sticking to that rule and therefore not marking singleton premodifying entities. (Note this means singleton premodifying entities are handled differently in this project than singleton premodifying *EVENTs*, which were marked by T1Temp.)

- You'll notice that if a premodifying entity has an identical head noun mention elsewhere, but it's part of the name of a disease, it's still unmarked. For example, "colon" in "colon cancer" isn't marked even if "colon" appears elsewhere as the head of a phrase. This is because they're not really referential to the body parts; they're just identifying the kind of disease. This becomes more obvious with something like "lung cancer." If a note discusses a "patient with lung cancer and" refers to "the disease in her right lung," what is "lung" in "lung cancer" referring to? Is it a singular reference to both her lungs? Should we assume it's referring specifically to her right lung? The confusion here seems to be due to the fact that "lung" in this context isn't really pointing to either lung or both lungs; it's just telling us what kind of cancer she has.
- You'll also find that T1Coref didn't mark adjectival forms of entities. In the noun phrase "resected rectal adenocarcinoma," for example, "rectal" will be left unmarked, even if there's a bare noun mention of [rectum] elsewhere in the document.

Finally, it is rare but possible that you may also find events that were erroneously unmarked.

If these events are clearly important and relevant to the patient's clinical timeline, you may mark them as EVENTs, adjust the features as necessary, and check for any coreference relations:

(172) Patient had three [surgeries] last year. One was to remove her appendix.

Here "One" refers to one of the surgeries the patient had, but it was left unannotated. However, we want to mark referential pronouns in this project. To fix this error, mark [one] as an EVENT with DocTimeRel BEFORE, and create a SET-SUBSET relation between it and [surgeries].

Make sure you also avoid marking EVENTs for the sections we intentionally leave unannotated (see Appendix B). When in doubt as to whether an EVENT has clinical relevance or not, default to marking it.

Adding CONTAINS-SUBEVENT links

We're adding one brand-new relation in this pass: CONTAINS-SUBEVENT. We treat the event-subevent relation as a type of temporal link, i.e., a TLINK, because anything that is truly a subevent will necessarily be temporally contained within the greater whole, in the same way that a part of an object is spatially contained (in most cases) within the whole of the object. But, as you'll see, it's frequently lumped in with the coreference links in regard to how we use it, because it carries structural information as well as temporal.

While a CONTAINS relation indicates only temporal containment, CONTAINS-SUBEVENT additionally says there's some kind of meaningful, hierarchical connection between two events. More specifically, a subevent can be thought of as "part of the script" of the larger event; something that makes up part of the essential structure of the larger event. In short, if an EVENT is a typical subpart of the larger event's type, it's a SUBEVENT.

Compare the following two examples (only relevant events annotated):

> (173) Patient taken for [surgery] on January 13, 2014. The surgeon [removed] the mass and [repaired] abdominal hernia as well.
>
> (174) During Friday's [surgery], the patient's heart rate [spiked] and she [fell] off the table.

In (173), [removed] and [repaired] are subevents of the larger [surgery] event, as these are typical subparts of surgery.

In (174), however, the events [spiked] and [fell] are not in a subevent relation with [surgery]. A patient falling off the table is certainly not part of the structure of a surgery. The patient's heart rate spiking is similarly unrelated – this is a finer line of distinction, because it's possible the spike may have been caused by the surgery, but note this is a causal relationship (which we do not annotate), not a subevent relation. A heart rate spike is in no way "part of the script" of a resection. Therefore, [surgery] would CONTAIN [spiked] and [fell], since there's clearly still temporal containment, but there's no more meaningful relation here for us to mark.

Consider a non-clinical example: If you went to the grocery store and ran into an old friend, the encounter with your friend would be caused by the grocery store trip, but it's not part of the errand – whereas buying milk would be.

Note that our definition says *typical*, not *necessary*. Surgeries may not necessarily include a hernia repair, for example, but it's certainly part of the structure of the surgical procedure in a way that the spiking and falling events are not.

We will be annotating CONTAINS-SUBEVENT links for the following four categories:

1) Cancer events

2) Cancer treatment events, including:

> **a)** Surgeries
>
> **b)** Chemotherapy/radiation

3) Chronic disease (episodes)

4) Medications

All other event-subevent relationships in the texts should be ignored.

CON-SUB and cancer events

By “cancer EVENTs,” we mean terms like [cancer], [tumor], [adenocarcinoma], [mass], [metastases], etc. – any term that refers either to the overall cancer or to some part of the cancer.

We distinguish between terms that point to the overall disease event (“cancer,” “disease”) and terms that point to physical manifestation of that disease (“tumor,” “metastasis,” “carcinoma,” etc.). (Appendix B contains lists of these terms.) This is because a cancer may consist of multiple manifestations, so we don’t want to say it’s IDENTICAL to any individual one.

Rather, we consider a tumor event (a physical-manifestation event) to be a SUBEVENT of the whole cancer:

(175) Patient has colon [cancer]. CT scan last week showed [carcinoma] in the sigmoid as well as a single liver [metastasis].

a. [cancer] CONTAINS-SUBEVENT [carcinoma]

b. [cancer] CONTAINS-SUBEVENT [metastasis]

We understand [carcinoma] and similar diagnostic terms as referring to the presence of the tumor in the sigmoid; [metastasis] refers to the presence of a different, secondary growth in the liver. Therefore, the two mentions are different SUBEVENTs of the colon cancer. While histologically they’re the same, the two terms point to the presence of two different masses (which also have different temporal natures).

However, keep the following in mind:

- Words are used variably! The terms “disease” and “cancer” may be modified such that they refer to physical manifestations of the cancer rather than the entire disease:

  (176) We will attempt to remove the metastatic [disease] in the pelvis as well as resect the primary [tumor] for Ms. Smith’s colon [cancer].

  Here, [disease] is modified as being in the pelvis, so it may (and should) be linked as IDENTICAL to specific mentions of the metastases or growths in the pelvis elsewhere in the note, rather than to terms referring to the entire cancer event. So the links here are:

  a. [cancer] CONTAINS-SUBEVENT [disease]
  b. [cancer] CONTAINS-SUBEVENT [tumor]

However, **cancer and disease terms must be explicitly modified in the text to be treated as such**. When they are not explicitly modified, and for borderline cases, assume that general terms like *cancer* and *disease* should be containing events and therefore in CON-SUB relations with tumor terms.

- Be aware that not all masses, lesions, nodules, or tumors are cancerous. Benign masses should not be SUBEVENTs of the cancer. Context should help clarify. There must be evidence in the note that a mass is cancerous in order to link it to the cancer. Sometimes the doctor doesn't know yet if a mass is cancerous or not:

  (177) Liver [mass] is indeterminate.

  The mass here is "indeterminate," meaning the doctor doesn't know whether it's a metastatic growth stemming from the patient's colorectal cancer or not. We don't want to link what's explicitly unknown, so this should not be a subevent of [cancer].

Finally, **except for parts of tumors, all cancer subevents should be linked directly to the cancer**, rather than to each other via nesting:

(178) Patient has been seen here in the past for rectal [cancer]. His primary [tumor] was resected as well as three liver [metastases].

a. [cancer] CONTAINS-SUBEVENT [recurrence]

b. [cancer] CONTAINS-SUBEVENT [metastases]

This is because the finer-grained temporal structure of cancer subevents is often ambiguous or simply unknown – for example, in (178), we know the [tumor] began before the [metastases], but that's all we know. We can't say whether the presence of the tumor temporally contained the existence of the metastases or whether it was resected while the metastases were still present, and so forth.

Moreover, linking all subevents directly to the cancer isn't incorrect; it sacrifices some granularity, but buys us greater consistency in our annotations.

The one exception is that explicit references to parts of tumors may be linked as SUBEVENTs to the tumor itself rather than the overall cancer. This is simply because it's usually more intuitive to do it this way, which again buys us greater consistency. So:

(179) We are only able to remove [part] of the [tumor].

a. [tumor] CONTAINS-SUBEVENT [part]

Distinguishing cancer subevents from other cancer-related events

We're taking a fairly restricted view of subevents, where SUBEVENTs of cancer should include tumors and things the cancer "does," like [involve] the lymph nodes or [invade] surrounding tissue, but not symptoms or attributes of the cancer. Recognize that there are other semantic relations that are present but not annotated by this project, such as causative and attributive relations. Symptoms, for example, we understand as being caused by the cancer; they are not subevents.

Consider the following examples:

(180) Patient has experienced abdominal [pain], significant [bloating], and rectal [bleeding] over the last three months.

Even if the doctor were to clearly state that [pain], [bloating] and [bleeding] are related to the patient's cancer, these aren't subevents of the cancer – they're symptoms caused by the cancer. Therefore, there's no link to be made between these EVENTs and the cancer.

(181) The [thickening] of the right colon seen on CT scan is consistent with patient's colon [cancer].

The [thickening] again is a symptom of the colon cancer and is not a subevent.

(182) Pathology demonstrates the carcinoma is moderately [differentiated]. Stage is [pT3N0MX].

These two events are attributes of the cancer, not subevents.

(183) Patient comes for treatment of rectal [cancer]. The [involvement] of the lymph nodes by the [carcinoma] is concerning.

The fact that the cancer involves the lymph nodes is part of the "structure" of the cancer itself, so [involvement] is a SUBEVENT of [cancer] (as is [carcinoma]). Remember that it would be linked directly to the overall cancer event, not to the carcinoma, so:

- [cancer] CONTAINS-SUBEVENT [involvement]
- [cancer] CONTAINS-SUBEVENT [carcinoma]

Metachronous and synchronous cancers

Finally, one important note on cancer events: Watch out for cancers that are described as being either "metachronous" or "synchronous."

Metachronous colon cancers are different, new cancers from any previous cancer the patient had. They are therefore not IDENTICAL to any prior cancer. Synchronous colon cancers refer to multiple, different cancers that the patient has at the same time. They are not related to each other.

(184) Ms. Lang has metachronous rectal [cancer]$_1$. She was treated here in the past for rectal [cancer]$_2$ with successful resection of the [primary].

a. ~~[cancer]$_1$ IDENT [cancer]$_2$~~

b. [cancer]$_2$ CONTAINS-SUBEVENT [primary]

The two [cancer] EVENTs are not IDENT – the term “metachronous” is explicitly identifying this cancer event as being new and different from the prior rectal cancer. The reference to the primary tumor is clearly a subevent of the original cancer and not the new cancer.

CON-SUB and cancer treatment events

By “cancer treatment EVENTs,” we mean all EVENTs that represent treatment of the cancer, including surgery, chemotherapy, and radiation EVENTs. Cancer treatment events can be subdivided into two major categories: surgical procedures and chemotherapy/radiation administration. Both types will be discussed below, but first, a word on the term “treatment” itself:

General terms like [treated] and [treatment] are hard to handle consistently because sometimes the treatment apparently consists of one event; sometimes it consists of a variety of events; and oftentimes it's ambiguous.

Because of this variation in use, [treatment] should sometimes be IDENTICAL to the specific form of treatment mentioned, and it should sometimes be CONTAINS-SUBEVENT with the specific form. Make the best choice you can based on the way the author presents it in the text.

(185) Patient was [treated] with [surgery] at that time.

a. [treated] IDENT [surgery]

The wording here suggests that the entire treatment in view was the surgery, so an IDENTICAL relation is appropriate.

(186) Ms. Garcia’s colon cancer was successfully [treated] with three months of neoadjuvant [FOLFOX] and then [resection].

a. [treated] CONTAINS-SUBEVENT [FOLFOX]

b. [treated] CONTAINS-SUBEVENT [resection]

In situations like this one where there are clearly multiple different types of treatments, the general treatment term itself should not be linked as IDENT to either; rather, both treatment types are SUBEVENTs of the overall treatment.

A word on the phrase “treatment effect”:

(187) Mass demonstrates [treatment] [effect].

In this phrase, “treatment” always refers to chemotherapy and/or radiation, not surgery and should be linked accordingly.

Finally, while we don't do nested subevents for cancer events, we do nest treatment events as appropriate, as you'll see below.

Surgical procedures

Doctors sometimes use specific terms like colectomy and resection in reference to an entire surgical procedure. Other times, they're used to refer to specific subparts of the procedure, that is, the actual event of removing the colon/tumor/etc.

This makes consistent coreference linking tricky, because the difference in use is often ambiguous. Therefore, generally speaking, **specific procedural terms like [resection] should be SUBEVENTs of general terms like [surgery]**, rather than IDENT. (Appendix B contains lists of "general" terms and "specific" terms.) For example:

(188) October 15th, 2015 – Dr. Martin performed low-anterior resection. Ms. Hollister's recovery from surgery has been without complication.

a. [surgery] CONTAINS-SUBEVENT [resection]

(189) PLAN: Low-anterior resection and gallbladder removal. Patient is scheduled for surgery on October 15th, 2015.

a. No coreference relation.

b. [surgery] CONTAINS-SUBEVENT [resection], [removal]

The second example clarifies why we're not linking general terms to specific terms – here it's clear that the [surgery] consisted of at least two subprocedures, the [resection] and the gallbladder [removal]. So it would be incorrect to say that the [surgery] is IDENTICAL to either one.

You'll note this is very similar to our distinction between "cancer" and "tumor" events above. Just as with those events, it's possible that the doctor may explicitly modify a "general" surgery term such that it clearly refers to some subprocedure. In that case, you may and should link it as IDENT to the subprocedure:

(190) On September 18, 2012, patient underwent surgical [management] of her cancer including [removal] of the primary tumor and liver [surgery].

a. [management] CONTAINS-SUBEVENT [removal]
b. [management] CONTAINS-SUBEVENT [surgery]

Here [surgery] is clearly referring to a part of the overall procedure. As with cancer events, though, a general term must be explicitly modified in order for you to treat it as a subevent instead of a containing event. Default to the general-specific distinction for borderline cases. This will greatly ease cross-document linking.

Subprocedures can themselves contain SUBEVENTs. Note the nesting that occurs as a result:

(191) Low-anterior [resection] performed, including five lymph nodes [removed].

Patient's recovery from the [procedure] has been without complication.

a. [procedure] CONTAINS-SUBEVENT [resection]

b. [resection] CONTAINS-SUBEVENT [removed]

Finally, note that colonoscopies and endoscopies are diagnostic procedures, not surgical procedures. We don't create subevent relations for these, nor for other tests like MRIs, CT scans, etc. We also don't make subevent links for procedures that are performed for something other than cancer treatment. For example, you don't have to create subevent links for an appendectomy. However, if a non-cancer-related procedure, such as a hernia repair or skin tag removal, is incidentally part of the cancer-treatment surgery, include it as a SUBEVENT. Put differently: Only create SUBEVENT links for procedures that treat cancer at least in some way; but, for those procedures, create links for *all* their subevent relations present in the text. When in doubt, create a link.

Chemotherapy and radiation

We understand [chemotherapy] and [radiation] EVENTs as referring to the events of the patient being treated with chemotherapy and radiation. While these EVENTs have limited types of subevents (usually references to subparts of a treatment course or dose increases/decreases), they constitute some of the most challenging relationships to annotate due to their complex temporal structure and the variable way that's presented in the texts.

Let's start with some examples:

(192) Recent scans demonstrates good results from the [chemoradiotherapy].

Patient received [capecitabine] for two months with concurrent radiation

[treatments] (total of 5040 [cGy]). She tolerated the [Xeloda] well.

a. [chemoradiotherapy] CONTAINS-SUBEVENT [capecitabine]

b. [chemoradiotherapy] CONTAINS-SUBEVENT [treatments]

c. [capecitabine] IDENT [Xeloda]

d. [treatments] IDENT [cGy]

The term *chemoradiotherapy* refers to the event of treating the patient with chemotherapy and radiation, so the individual mentions of chemo and radiation administration are subevents of this treatment. [capecitabine] is a type of chemotherapy, and is merely a different name for Xeloda; hence the IDENT link between the two. cGy (centigray) is a measurement unit used specifically for radiation, so the administration of the total dose of cGy refers to the same EVENT as the administration of the radiation treatments.

Clearly there's a need for medical knowledge in making some of these decisions. When you're not sure what the right annotation choice is due to lack of clinical knowledge, use the *Needs_Medical_Opinion* feature (discussed below in Reconciling the merged data) or contact your supervisor, depending on the complexity of the issue.

Another example:

(193) After resection, Mr. Hall underwent adjuvant chemoradiation [treatments].

Twelve radiation [fractions] were administered along with [FOLFOX].

[Therapy] ended December 13, 2009. [Chemotherapy] was poorly tolerated.

a. [treatments] IDENT [Therapy]

b. [treatments] CONTAINS-SUBEVENT [fractions]

c. [treatments] CONTAINS-SUBEVENT [FOLFOX]

d. [FOLFOX] IDENT [Chemotherapy]

This example demonstrates a few things:

a) A course of chemotherapy or radiation may be viewed as either one contiguous therapy event or as multiple discrete administration events, and they're variably referred to as both in these notes. You can see this in the fact that the adjuvant treatment course is referred to in plural form early in the paragraph and in singular form later on ([Therapy]). We want to link references to the same treatment course as IDENT regardless of plurality.

b) This contiguous/discrete variability also impacts structural links. When a treatment is referred to as individuated events, the instinct is to use S-SS; when it's discussed as a continuous course, the instinct is for CON-SUB. For the sake of consistency, **you should always use CONTAINS-SUBEVENT to link larger therapy events to smaller ones**, whether they are referred to as groups of separate events or one ongoing event.

c) Relatedly, chemotherapy events, like many other kinds of events, may be referred to by general or specific terms. As always in this project, we're interested in representing the real-life events and relationships that are present, regardless of what word is used. In this example, the patient receiving FOLFOX is the same event as the patient receiving chemotherapy – it's just in one case the doctor uses the specific term to refer to that event (FOLFOX) and in the other case he/she uses the general term (chemotherapy).

To expand on (c): If a specific type of chemotherapy is the *only* type of chemotherapy a patient receives, it should be IDENTICAL to the more general chemotherapy term. If the patient receives multiple types of chemo, they should be SUBEVENTs of [chemotherapy]. Compare:

(194) Patient received [chemotherapy]...Patient received [FOLFOX].

a. [chemotherapy] IDENT [FOLFOX]

(195) Patient received [chemotherapy]...Patient received [FOLFOX] and [Avastin].

a. [chemotherapy] CONTAINS-SUBEVENT [FOLFOX], [Avastin]

In (194), the [chemotherapy] event *is* the [FOLFOX] event; in (195), the [chemotherapy] event consists of both the FOLFOX and Avastin events.

Finally, these notes frequently discuss “cycles of chemotherapy,” which follow all of the above guidelines, but warrant more explanation here. The term “cycles” (and similar ones, e.g. “fractions”) should be annotated both as QUANTIFIERs (a type of TIMEX3) and as EVENTs. The discussion here is regarding their treatment as EVENTs. All examples are simplified for the sake of clarity, and the discussion assumes there's not counter information in the rest of the document.

a) The total number of [cycles] that a patient receives should be IDENTICAL to mentions of the overall course of treatment.

Say you have:

(196) Patient received [chemotherapy] ... Patient received a total of 12 [cycles].

[chemotherapy] and [cycles] are IDENT. The administration of the 12 cycles refers to the same real life event has the administration of the chemotherapy. (This can happen within the same syntactic phrase, so they'd still be IDENT in the following: "Patient received a total of 12 [cycles] of [chemotherapy].")

b) For “quant of X” phrases, such as “four cycles of chemotherapy,” assume that “X” refers to the entire course of treatment as it's presented in the text by the doctor (taking dates and other linguistic cues into account).

This is a bit more nebulous, but compare the following examples (IDENT links are both shown by subscripts and listed separately for clarity):

(197) 1. August 2012 through September 2012: Received four [cycles]$_1$ of [chemotherapy]$_1$.

2. September 28, 2012 through October 3, 2012: Patient out of the country

on vacation.

3. October 2012, through November 2012: Six [cycles]$_2$ of [chemotherapy]$_2$ administered.

a. [cycles]$_1$ IDENT [chemotherapy]$_1$

b. [cycles]$_2$ IDENT [chemotherapy]$_2$

No relation between the “1” EVENTs and the “2” EVENTs.

(198) Patient will receive one additional [cycle]$_2$ of [FOLFOX]$_1$. Scans show good results from the first four [cycles]$_3$ of [chemotherapy]$_1$.

a. [FOLFOX]$_1$ IDENT [chemotherapy]$_1$
b. [FOLFOX]$_1$ CONTAINS-SUBEVENT one additional [cycle]$_2$
c. [FOLFOX]$_1$ CONTAINS-SUBEVENT the first four [cycles]$_3$

In both examples, we are understanding [chemotherapy] as referring to the whole course of treatment as presented by the author:

In (197), the “1” and “2” chemotherapy references are understood to be separate EVENTs because of the way they’re explicitly related to different timeframes by the authoring doctor. (If there was a reference elsewhere in the note to the patient’s entire adjuvant treatment, this would have a CONTAINS-SUBEVENT relation with both sets of chemotherapy references.)

In (198), we have the linguistic cue "additional" letting us know that this final cycle is still considered to be part of the same course of treatment. So [FOLFOX] in the first sentence is IDENT to [chemotherapy] in the second sentence, because they're both referring to the same course of treatment.

c) Follow the practice of always using CONTAINS-SUBEVENT to nest cycles of chemotherapy events.

(199) Mr. Connor has received 12 [cycles] of [Xeloda]. Dose [increase] after fifth [cycle].

a. 12 [cycles] IDENT [Xeloda]
b. 12 [cycles] CONTAINS-SUBEVENT [increase]
c. 12 [cycles] CONTAINS-SUBEVENT fifth [cycle]

This is simply reinforcing the earlier point that CONTAINS-SUBEVENT should always be used to represent structural relationships for chemo and radiation events, rather than SET-SUBSET. You can clearly see why here since there are both singular and plural EVENTs in the overall chemotherapy event chain. If [Xeloda] rather than 12 [cycles] had been chosen for the link with fifth [cycle], then a CONTAINS-SUBEVENT link would have been the intuitive relation.

CON-SUB and chronic disease events

We create CONTAINS-SUBEVENT links between chronic disease events and episodes or flare-ups of that disease, where references to the chronic disease are the containing EVENTs and references to specific flare-ups are the SUBEVENTs.

(200) Main complaint: Chronic atrial [fibrillation]$_1$. Mr. Jones was brought to the ER two months ago while in atrial [fibrillation]$_2$. Last week another [episode] occurred.

a. [fibrillation]$_1$ CONTAINS-SUBEVENT [fibrillation]$_2$
b. [fibrillation]$_1$ CONTAINS-SUBEVENT [episode]

Observing the preannotated temporal links will come in handy for making these subevent linking decisions:

c. {two months ago} CONTAINS [fibrillation]$_2$

d. {Last week} CONTAINS [episode]

[fibrillation]$_1$ has a DocTimeRel of OVERLAP, meaning it is present at the time the document was written (which is clearly after the two EVENTs contained by the temporal expressions shown above). In other words, what we have here is a reference to the patient's ongoing chronic disease (fibrillation$_1$), and references to two specific episodes of that disease. We don't want an IDENTICAL link between these three EVENTs, because it's illogical to say that an EVENT that's temporally contained by one date is the same event as an EVENT that's temporally contained by another date – and that both those EVENTs are the same as another EVENT that OVERLAPs DOCTIME, yet another date!

We specifically only do SUBEVENT linking for episodic subevents of the chronic disease; you don't have to worry about what "counts" as other subevents for chronic diseases. So be on the lookout for this, but it'll likely not be very common.

CON-SUB and medications events

CONTAINS-SUBEVENT links for medications events are discussed below in Medications sections.

CONTAINS-SUBEVENT general guidelines

Keep the following guidelines in mind as you create CONTAINS-SUBEVENT links:

- **CONTAINS-SUBEVENT links may only be used for EVENTs.**

WHOLE-PART can be thought of as the equivalent link for entities. Additionally, TIMEX3s should never be involved in CONTAINS-SUBEVENT links.

- **We are not annotating HYPOTHETICAL or GENERIC SUBEVENT relations.**

In fact, we aren't annotating any relations for HYPOTHETICAL or GENERIC EVENTS in this project at all (discussed more below in Coreference links and modality).

As with all other links, you may link ACTUAL and HEDGED EVENTs to each other with CONTAINS-SUBEVENT as appropriate, but if an EVENT is HYPOTHETICAL or GENERIC, you should not create any link for it.

- **CONTAINS-SUBEVENT should be used across the whole document.**

Typically we don't permit TLINKs to cross sections of a medical note. CONTAINS-SUBEVENT is the one exception to this since it conveys more than just temporal information. Like coreference links, it can and should be used to relate EVENTs wherever appropriate, regardless of whether the EVENTs appear in the same section or different sections.

- **You only have to make one CONTAINS-SUBEVENT link for the same two EVENTs.**

Say you have multiple mentions of the same colon cancer and multiple mentions of the same tumor that's part of that cancer. You only have to create one CONTAINS-SUBEVENT link between one of the cancer references and one of the tumor references. The CON-SUB relation can then be inferred for the other mentions because they are linked as IDENT.

- **Don't worry about CONTAINS links that are already present.**

When two EVENTs that are in an event-subevent relation are already linked via a CONTAINS TLINK, simply leave the CONTAINS link and create a new CONTAINS-SUBEVENT TLINK.

- **Negated EVENTs may be SUBEVENTs.**

Recall that when an author suggests that there is some kind of significance or relevance between a negated EVENT and another EVENT, we want to capture that relation:

(201) Patient has colon [cancer] without [disease]$_{\text{NEG}}$ in the lymph nodes.

a. [cancer] CONTAINS-SUBEVENT [disease]$_{\text{NEG}}$

This admittedly pushes the limits of the part of our definition that says a SUBEVENT is "a typical subpart of the larger event's type" – how can we say that not having disease in the lymph nodes is a typical part of colon cancer? However, the author is still positing a meaningful, structural connection between these two EVENTs that we want to capture – the non-presence of lymph node disease spread is a clinically important part of the structure of the patient's cancer. So we will continue to follow the policy of marking relations the author suggests are important for negated EVENTs.

- **Use CONTAINS-SUBEVENT for therapy and medications nesting; use SET-SUBSET for labs and tumor nesting**

As discussed above, doctors frequently use count nouns and mass nouns interchangeably to refer to the same event/set of events. So they might refer to a patient's course of chemotherapy at one point as "neoadjuvant therapy" and elsewhere as "neoadjuvant treatments." In the first case, the chemo is conceptualized as an ongoing course of treatment made up of subevents; in the second case, it's conceptualized as a set of temporally discrete events, made up of member events. But, we certainly want to link these two terms as IDENT (therapy and treatments), so the question then is whether to use CON-SUB or S-SS when linking to a "smaller" event. Because we have to be consistent, and because it's going to be counterintuitive at some point either way, use the following hierarchical relations for the following categories of events:

a) therapy/treatments: CONTAINS-SUBEVENT

- patient finished chemotherapy. After sixth cycle, we re-did her port.
  - [chemotherapy] CON-SUB [cycle]
- patient finished 12 cycles. After sixth cycle, we re-did her port.
  - [cycles] CON-SUB [cycle]

b) medications/treatment: CONTAINS-SUBEVENT

- patient on treatment for diabetes. Taking insulin$_A$ and insulin$_B$
  - [treatment] CON-SUB [insulin]$_A$, [insulin]$_B$
- patient on multiple medications for diabetes. Taking insulin$_A$ and insulin$_B$.
  - [medications] CON-SUB [insulin]$_A$, [insulin]$_B$

c) adenocarcinoma/masses: SET-SUBSET

- Colonoscopy demonstrated adenocarcinoma forming two masses. One mass shows invasive growth into colon wall.
  - [adenocarcinoma] IDENT [masses]
  - [adenocarcinoma] S-SS [mass]

d) labs/labwork: SET-SUBSET

- Labs: WBC…hemoglobin…creatinine
  - [Labs] S-SS [WBC], [hemoglobin], [creatinine]
- Bloodwork: WBC…hemoglobin…creatinine
  - [Bloodwork] S-SS [WBC], [hemoglobin], [creatinine]

Reconciling the merged data

Recall that the goal of this pass is to prepare within-document annotations for cross-document linking. In addition to marking singleton entities and adding CONTAINS-SUBEVENT relations, this means also making sure that the two sets of annotations from the two prior projects "make sense" together; if they don't, cross-document annotation will be much more confusing!

To that end, we've identified several specific categories of problems that we know are present in the data as a result of merging the two groups of annotations. These cases are discussed below, and you should check each document for these issues and make adjustments as necessary. A summary checklist of these issues appears in Appendix B.

However, **it's impossible to anticipate and describe every inconsistency you may come across.** Therefore, in addition to these specific issues, you should also check each markable's coreference links for other obvious errors or inconsistencies and fix them if found. You are also permitted to change any other egregious error you encounter – emphasis on "egregious" (more on this below).

The reason coreference links need to be checked is not necessarily because T1Coref did a poor job; it's because their interpretations of certain phenomena differed from T1Temp's, and they did not have all the EVENT features available to them. Here are a couple examples of the kinds of egregious coreference linking errors you may run into:

(202) Patient previously took [Advil] but is not currently using [Advil].

a. [injections] IDENT [injections]

These two Advil EVENTs were linked as IDENTICAL in the merged preannotated data. However, the first one is BEFORE, POSITIVE, indicating the Advil the patient ingested in the past; the second is OVERLAP, NEGATIVE, indicating the fact that the patient is currently not taking Advil. These are therefore two different medication events, and the IDENTICAL link should be deleted.

(203) We discussed that the risks of this kind of [surgery] are generally minimal.

Patient agreeable for proceeding with [surgery] next week.

a. [surgery] IDENT [surgery]

In this example the first mention of [surgery] points to surgeries generally (it's GENERIC), while the second one points to a specific surgery event the patient will experience (it's ACTUAL). These are simply two different surgery EVENTs, so the IDENTICAL chain should be deleted. (A SET-SUBSET link should be created instead, since Mrs. Smith's surgery is an instance of surgeries as a whole.)

Again, it's impossible to list all the possible kinds of coreference errors that might occur. Base your decisions on what you already know about coreference links (referring to the appropriate sections of the guidelines as needed), and on your own understanding of the semantics of a given document. When in doubt about how to interpret a given EVENT, refer to its features – for example, the fact that one of the [Advil] EVENTs in (202) is *POS* and one is *NEG* is a pretty good indicator they represent different events! (It's sometimes the case that an EVENT's feature may be wrong, in which case you're permitted to change the feature instead of the link, but you should default to assuming that the feature is correct unless egregiously inaccurate.)

Finally, and importantly:

- While you should check every markable's coreference links, **you do NOT need to check the temporal links**. If you happen to discover a completely obvious TLINKing error along the way, you are permitted to fix it, but don't go looking for them. You should also thoroughly review Temporal relations prior to changing a TLINK.
- Relatedly, you are permitted to fix any type of error or inconsistency in this pass – whether a coreference linking error, a temporal linking error, an EVENT feature error, etc. – but in any case **the error must be egregious to warrant fixing it. Don't spend time on borderline cases**. We are only interested in addressing the most obvious errors, and, moreover, there are likely reasons that edge cases were decided one way or another. So, for example, if there's an ambiguous [we] markable in the text and it could be read as SET-SUBSET with the doctor and the patient, or it could be read as SET-SUBSET with the doctor and the rest of the medical team, don't change what's there and don't spend a lot of time thinking about it.
- Note that adding CONTAINS-SUBEVENT links will "force" other errors. For example, you'll often find a specific procedural term like [resection] in an IDENTICAL chain with a general one like [surgery]. In addition to creating the CONTAINS-SUBEVENT link, you should also delete the IDENTICAL link, since it doesn't make sense to say that [surgery] is both the containing event for the resection and that it's the same event as the resection! In other words, you can and should change anything you need to in order to bring the annotations into alignment with our current rules, without contradiction.
- Many relations require some degree of medical knowledge. For this reason, all links come with a "Needs_Medical_Opinion" feature, which appears as a property of the relation and defaults to "False." You may change the flag to "True" when you're not sure of the accuracy of a relation due to lack of clinical knowledge. (Note that this is different from questions about our annotation style and practices, what the relations mean, etc. You should contact your supervisor with these types of questions rather than using Needs_Med_Opinion for them.)

Following are the specific issues you should look for and adjust as appropriate, divided into five categories: EVENTs and entities; medications and allergies sections; coreference links; temporal expressions/DOCTIME/SECTIONTIME; and pathology notes.

EVENTs and entities checks

1) Converting eventive MARKABLEs to EVENTs and vice versa

For the following discussion, recall that lower-case "markable" refers to anything that is referential and is or should be marked in the document – this includes EVENTs, entities, TIMEX3s, DOCTIMEs and SECTIONTIMEs. Upper-case "MARKABLE" refers specifically to the actual category in the online annotation tool – these are markables that are realized by the featureless MARKABLE category.

While T1Temp used the schema category "EVENT" to mark actual events, T1Coref used "MARKABLE" for any type of markable – events, entities, and temporal expressions. They didn't distinguish between these categories at all. As you can imagine, these two different approaches lead to some interesting consequences for the merged files!

Because T1Coref permitted marking certain things that T1Temp didn't (and vice versa), and because T1Coref did not distinguish between entities and events, it is sometimes the case in the preannotated data that a MARKABLE is conceptually an event rather than an entity, and may therefore appear in links with other EVENTs.

In this pass, one of our main goals is to resolve this issue. **We will delete all MARKABLEs that in fact refer to events and recreate them as EVENTs**. This means that at the end of this pass our schema categories will consistently reflect our conceptual categories, which will make cross-document linking much easier down the road. Note that entities will still be called MARKABLEs, because that's the term our tool uses, but all MARKABLEs will actually refer to entities after doing this conversion.

Importantly, this also means that at the end of this pass, EVENTs and MARKABLEs should **never be linked to each other**. One simple but extremely helpful visual trick is in order: When you find a coreference chain that has both blue markables (EVENTs) and gray markables (MARKABLEs) in it, that's a great indicator that something needs to be changed – either you've got an incorrect chain, or you've got a MARKABLE that needs to be converted to an EVENT, or an EVENT that needs to be converted to a MARKABLE. At the end of this pass, all links should have only blue in them or only gray in them. (LOCATIVE WP, discussed below, is the sole exception; but this relation will not be part of our project's output.)

Process for MARKABLE conversion

For each MARKABLE in a document, you need to ask yourself whether it refers to an event or an entity. If it's talking about an entity, leave it as a MARKABLE. If it's talking about an event, you'll delete the MARKABLE and create an EVENT instead.

In order to do this, you need a solid grasp on the difference between events and entities. To that end, you should go back to The preannotated THYME 1 temporal and coreference data and review:

- EVENTs
- Entities
- Implicit EVENTs

These sections should guide your decisions about what constitutes an eventive MARKABLE and what does not.

Of course, there are borderline cases between events and entities. If you find a borderline markable not accounted for by the discussion in the sections listed above, you should:

- Let the links it's involved in help guide your choice. If, say, you've got a single MARKABLE in an IDENTICAL chain with a bunch of EVENTs, chances are good it should be an EVENT.
- Lean EVENT-heavy. For entities, you should essentially stick to the categories listed in *Things that are never EVENTs* under Implicit Events and let other EVENTs that "look" like entities remain EVENTs.

Once you've looked at a MARKABLE and decided that it represents an event and not an entity, follow these steps:

a) Create an EVENT for the mention

Let's work with this example:

(204) Patient will undergo [colon cancer]$_M$ surgery next week.

[colon cancer] is a MARKABLE that should be an EVENT. Before deleting the MARKABLE, first create an EVENT:

(205) Patient will undergo [colon [cancer]$_E$]$_M$ surgery next week.

Note that the span should be different! Remember we use maximal spans for MARKABLEs and minimal spans for EVENTs, so when you create the EVENT, be sure to just mark the headword. (Of course, [surgery] in this example should also be marked as an EVENT.)

b) Select DocTimeRel for the EVENT and adjust other features as appropriate

Now that we're marking these as EVENTs, we have the opportunity to provide more information about the events. Make these choices based what you already know about DocTimeRel, polarity, modality, etc., and refer to these topics in The preannotated THYME 1 temporal and coreference data guidelines if necessary. For our current example, the cancer is currently present, so it should get a DocTimeRel of OVERLAP, and all the other properties should have their default value.

c) For all associated links, replace the MARKABLE with the EVENT

For our current example, the MARKABLE [colon cancer] is currently in an IDENTICAL chain with other mentions of the patient's colon cancer elsewhere in the note. Delete the MARKABLE [colon cancer] from this IDENT chain and replace it with the new [cancer] EVENT you just created.

d) Delete the MARKABLE

The reason for doing this last is so you don't end up forgetting what links it's part of, resulting in either loss of information or a lot of unnecessary additional work.

Most common eventive MARKABLEs

Due to the two previous projects' different markability rules, it's possible for just about any kind of event to show up as a MARKABLE. However, you'll likely run into the following three categories most frequently. Other eventive MARKABLEs should be somewhat rare by comparison.

a) PREPOSTEXPs

We interpret PREPOSTEXPs ("pre- and post- expressions") in two different ways:

i) As temporal expressions referring to "the period of time before X" or "the period of time after X";

ii) As events referring to the actual event within the term.

Example:

(206) Ms. Gray had surgery last month for her rectal cancer. Her {[postoperative]} recovery has been smooth.

a. {postoperative} BEGINS-ON [postoperative]

b. {postoperative} CONTAINS [recovery]

c. [surgery] IDENT [postoperative]

Here, {postoperative} is a TIMEX3 referring to the period of time after the surgery. **[postoperative] is also understood to be an event referring to the surgical procedure itself**. Therefore, it should be an EVENT with all the same feature assignments (DocTimeRel/modality/etc.) as the [procedure] – BEFORE, POS, ACTUAL in this case. It should also therefore be in an IDENTICAL chain with all mentions of the overarching surgery: [surgery] IDENT [postoperative]. Finally, because {postoperative} refers to the time period immediately following surgery, and [postoperative] refers to the surgery itself, {postoperative} BEGINS-ON [postoperative].

The interpretation of PREPOSTEXPs as EVENTs is what we're primarily concerned with in this task, because, as just pointed out, they always refer to an event, so they should always be converted from MARKABLEs to EVENTs, if they appear as MARKABLEs in the data.

Note that PREPOSTEXPs most often refer to surgeries, but may refer to other events as well, such as "hospitalization" in [posthospitalization].

b) Pronouns

Pronominal events will often appear as MARKABLEs. As with other MARKABLEs, identify what the pronoun is referring to, and convert it to an EVENT if appropriate. (Naturally, pronouns may refer to entities or events.)

> (207) Patient suffers from [diabetes]. Generally speaking [this] has been poorly-controlled.

[This] refers to the diabetes, so it should be changed from a MARKABLE to an EVENT.

c) Events in medications and allergies sections

Discussed below in Annotating medications and allergies sections.

Converting EVENTs to MARKABLEs

We're also converting EVENTs to MARKABLEs when entities have mistakenly been marked as EVENTs. Refer again to Implicit EVENTs.

EVENT-to-MARKABLE conversion should be fairly rare; the only category you're likely to come across with some consistency is for tissue references like *specimen*, *margin*, and *section*, which were sometimes marked as EVENTs in THYME, and which we want to change to MARKABLEs. You should follow the same steps outlined above for converting MARKABLEs to EVENTs. (If an EVENT that should be a MARKABLE participates in a TLINK, you may simply delete the TLINK.)

2) Double-tagged events

Recall that T1Temp marked EVENTs with minimal-span annotation, including only the headword in the span. T1Coref used maximal-span annotation for its MARKABLEs, including the entire syntactic phrase in the span. The data has undergone an automated merge, where, for the most part, if a markable for a given file was marked as both an EVENT in T1Temp and a MARKABLE in T1Coref, the MARKABLE has been automatically deleted and the EVENT left.

However, in some cases the automated task was unable to identify when the same markable was in fact annotated by both projects, meaning that at times a markable will appear in a file as both an EVENT and a MARKABLE (i.e., it's been double-tagged). For example:

(208) [Invasive moderately differentiated [adenocarcinoma]$_E$]$_M$ is identified in the sigmoid.

Here, the adenocarcinoma has been marked twice:

- [adenocarcinoma]$_E$
- [Invasive moderately differentiated adenocarcinoma]$_M$

Terms are considered to be double-tagged if the headword for both the MARKABLE and the EVENT is the same. Another (more semantic) way of thinking about it is if the referent being talked about is the same. So, for this example, "adenocarcinoma" has been double-tagged because, while the MARKABLE happens to include all the descriptors of the adenocarcinoma in its span, both markables are ultimately referring to the same real-life thing – the adenocarcinoma. (Review Spans for more discussion on headedness if necessary.)

If the headword for two markables is the same – that is, if the same core event is being referred to – delete the MARKABLE and keep the EVENT.

Contrast the preceding example with:

(209) She has [a mother with heart [disease]$_E$]$_M$

The two markables here do not have the same headword, so these annotations are fine. The MARKABLE [a mother with heart disease] is referring to the mother, and the EVENT [disease] is referring to the heart disease. Remember that because MARKABLEs are annotated with maximal span, there will sometimes be other markables that are contained within that span.

Therefore, what you're looking for is not simply whether spans overlap, but whether headwords of two different markables are the same.

Once you've found a double-tagged event, before deleting the MARKABLE, you should follow essentially the same order of steps discussed above in markable conversion. (Note that this is nearly the same issue; it's just that the EVENT has already been created for you!)

For example, say the above MARKABLE [Invasive moderately differentiated adenocarcinoma] is in an IDENT chain with [mass] in the following:

(210) The sigmoid [mass] is quite large.

To fix this, you should:

1) Delete the MARKABLE *[Invasive moderately differentiated adenocarcinoma]* from the IDENT chain

2) Add the corresponding EVENT *[adenocarcinoma]* in its place

3) Delete the MARKABLE *[Invasive moderately differentiated adenocarcinoma]*

The only time you should keep the MARKABLE and delete the EVENT in the case of double-tagging is if the term actually refers to an entity and has erroneously been marked as an EVENT. In that case, just do the opposite – delete the EVENT and keep the MARKABLE. This'll likely be fairly rare.

Markables we DO double-tag

By and large, we don't want markables double-tagged. There are four exceptions:

1) PREPOSTEXPs. Discussed in Most common eventive MARKABLEs.

2) Cycles. Discussed in Chemotherapy and radiation and Quantifier spans.

3) Physical properties in the Vital Signs section, section ID 20110. Discussed in Vital Signs.

4) Physical properties in other sections when they are the only way we can capture a testing event. Discussed in An exception: When properties refer to tests.

Nothing else should be double-tagged.

Quantifiers and double-tagging

Finally, there's one complicating factor on the topic of double-tagging, and that's the fact that, in the case of quantifier-type constructions, the syntactic headword is not the same:

(211) Endoscopist saw an [area of [thickening]$_E$]$_M$ in the transverse colon.

Here, the syntactic headword for the MARKABLE is "area"; for the EVENT, it's "thickening." But *semantically* the implicit event being referred to by both markables is exactly the same. In this kind of situation, you should delete the MARKABLE and keep the EVENT, just as you do with double-tagged things that have the same syntactic headword. Semantics trumps syntax!

In general, for these quantifier phrases (which usually follow the pattern *NP-of-NP*), if you believe that a single event or entity has incorrectly been marked as two, you may delete one of the mentions. If the markable refers to an event, keep the EVENT headword; if it refers to an entity, keep the full span MARKABLE.

Examples of single events that have been double-tagged:

(212) She has had [a couple of benign [tumors]$_E$]$_M$ removed in the past.

Delete the MARKABLE; keep [tumors].

(213) Patient’s father had [some type of [cancer]$_{E}$]$_{M}$ in his 60s.

Delete the MARKABLE; keep [cancer].

And an example of a single entity that’s been double-tagged and linked:

(214) [a portion of orange-tan [tissue]$_{M}$]$_{M}$

a. [tissue]$_{WHOLE}$ - [a portion of orange-tan tissue]$_{PART}$

You'll often see situations like this one, where two different MARKABLEs have been created and put in a WHOLE-PART relation. However, the phrase here is really pointing to a single entity. If we had "Darn, Santa put a lump of coal in my stocking" (unlikely in clinical data), there's only a single entity being identified by the phrase "lump of coal." There's one lump, and it happens to consist of coal.

So here, you'd delete the WHOLE-PART link, and then you’d delete the [tissue] MARKABLE and keep [a portion of orange-tan tissue]. (For MARKABLEs, we treat the quantifying term as the headword and grab the full span for which it’s the head; for EVENTs, we mark the actual thing being quantified, as you can see above. This is again for consistency with the original projects.)

The trick is recognizing when two entities are in fact present. Syntactically this looks exactly the same (*NP of NP*), so it's up to you to determine what's being represented semantically.

One rule we'll follow is that **any mention of a specific body part should be preserved as its own entity**:

(215) [a small portion of [the bladder]$_{M}$]$_{M}$

(216) [22.0 cm of [sigmoid colon]$_{M}$]$_{M}$

Keep the body parts and the W-P relations here, because the body part mentions can be understood as referring to the patient's actual anatomical sites.

A test you can use in other cases is to see if you can add a definite article (*the*) before the WHOLE and see if it makes sense (again, semantically – it's syntactically feasible to say "a portion of the orange-tan tissue," but it doesn't really make sense to say that unless the text has previously identified a specific section of tan tissue).

Of course, make sure other links that these are involved in make sense with the changes you make. The goal is that permitting this type of change will actually help clarify the coreference links by resolving situations where one entity has been mistakenly treated as two.

In any case, if you're really not sure, leave it as previously annotated.

3) Properties and their states

You’ll frequently run into discussions of properties of body parts or areas such as *heart rate*, *muscle strength*, and *diaphragmatic movement*. In general, these properties should be marked only as entities (i.e., MARKABLEs), not EVENTs. It’s always the case that for as long as the heart exists it has some kind of rate, no matter how fast or slow. Similarly, as long as a muscle exists it has some level of strength, no matter how great or small. Therefore, these properties aren’t eventive – they are not things that “happen.” We couldn’t link them to a timeline any more usefully than we could link a heart to a timeline.

The described *states* of these properties, however, are eventive and therefore should be annotated as EVENTs. The condition of a patient’s heart rate, for example, is a clinically relevant, eventive state that, moreover, has temporal significance – it could very well be the case that one document could describe the rate at that point in time as [regular], while a later document notes that it’s [irregular].

Properties and their states were annotated inconsistently in the data, so they should be changed accordingly. The following examples show the way they *should* be marked:

(217) [[Symmetrical]$_E$ diaphragmatic movement]$_M$.

(218) [Extraocular movements]$_M$ [intact]$_E$.

(219) [Heart]$_M$: [[[Regular]$_E$ rate]$_M$ and rhythm]$_M$. [Normal]$_E$ [S1]$_E$ and [S2]$_E$.

(220) Musculoskeletal: [Muscle strength]$_M$ 5/5 throughout with [[normal]$_E$ tone]$_M$.

(221) [Tympanic membranes]$_M$ and [ear canals]$_M$ [clear]$_E$.

(222) [Normal]$_E$ bowel [sounds]$_E$ [heard]$_E$.

A few more things to point out from these examples:

- Recall that part of speech doesn’t matter in determining eventivity. This is always true, but worth pointing out here. In (217), for example, the EVENT is a premodifier – [symmetrical]. (This is different from entities; we only mark premodifying entities if there’s an IDENTICAL head noun mention elsewhere in the note.) In (218), [intact] is the EVENT, though here it’s an adjectival predicate.
- As examples (219) and (222) show, all bodily sounds should be EVENTs, not entities (i.e., they’re not properties). We understand sound EVENTs as referring to the presence of those sounds. So in (222), e.g., [sounds] is an EVENT referring to the sounds' presence or existence; [normal] is an EVENT referring to the state of those sounds; and [heard] is an EVIDENTIAL EVENT referring to how we know those sounds are there.
- Recall that numbers may not be marked as EVENTs. Note in (220) we don’t mark “5.”
- This overlaps with the coreference checks you’ll be doing, but is appropriate to point out here: If the specific body part is mentioned along with its property, they should

be in a WHOLE-PART relation. For example, in (219): [heart]$_W$ - [Regular rate]$_P$, [Regular…rhythm]$_P$.

- Finally, we also mark references to body systems. You'll frequently run into terms like the following, where "musculoskeletal" is referring to the musculoskeletal system, and so forth:
    - Musculoskeletal
    - Neuro
    - Neurologic
    - Respiratory
    - Cardiovascular
    - Lymph
    - GI
    - Integumentary
    - Psychological
    - Mental

These should be marked as MARKABLEs. However, we're not attempting to annotate WHOLE-PART links for these systems, so relations for these MARKABLEs will be rare (you may occasionally come across an IDENTICAL relation for them).

An exception: When properties refer to tests

So far we've said that body properties should be marked and treated as entities. There are two situations in which we will also mark properties as EVENTs. One is in the Vital Signs section (discussed below). The second is as follows:

(223) [Carotid pulses]$_M$ are 4/4.

Here [Carotid pulses] is appropriately understood as a property. But there's also an implicit event present here, and that's the fact that the medical team tested the patient's carotid pulses. This test is a clinically relevant event that we want to be able to link to the timeline and possibly to other EVENTs.

Typically we're able to capture the testing event through either annotating a verb ("we will [check] her carotid pulses") or a result predicate (see most of the preceding examples; e.g., "[symmetrical] diaphragmatic movement"). If we have a test result, we know there was a test, even if one isn't explicitly mentioned.

In (223), though, neither option is available – we don't mark numbers, so we can't mark the "4/4" test result, and there's no markable verb. Because there's no other way we can mark the fact that they tested the patient's carotid pulses – and therefore it would not otherwise show up on the patient's clinical timeline – we also mark the property here as an EVENT, with the understanding that it refers to the event of testing the patient's carotid pulses.

In short, physical properties should always be marked and treated as entities. If there's no verb to mark instead AND there's no result predicate that can be marked, they should also be double-tagged as EVENTs.

More examples:

(224) [JVP]$_M$ is [elevated]$_E$.

Here "JVP" (jugular venous pressure) is not an EVENT because we have a result that's marked, [elevated].

(225) [Ventricular [rate]$_E$]$_M$ 76.

[rate] is an EVENT here as well as an entity because there's no markable verb or result.

(226) The ED [monitored]$_E$ [her heart rate]$_M$.

"rate" isn't an EVENT because we have a markable verb, [monitored].

Finally, if you have to *change* which markable represents a given event based on these guidelines (for example, had [rate] been marked instead of [monitored] in the preceding example), be sure to preserve and transfer all associated temporal links to the new EVENT.

4) Vital Signs

Nearly every clinical note has a Vital Signs section (section ID 20110 in the documents), which looks like the following:

(227) [start section id="20110"]

Date/Time=March 14, 2015:

Weight=X.X kg,
Weight=X.X [lb-av],
Temperature=X.X [degF],
Systolic=X mm[Hg],
Diastolic=X mm[Hg],
Pulse Rate=X/min,
[end section id="20110"]

The properties in these sections were annotated as EVENTs in T1Temp, with the understanding that they referred to the acts of measuring these properties. Properties in this section only may be left as EVENTs. However, they should also be marked as entities (MARKABLEs) and coreferred where appropriate. For example, the entity [Pulse Rate] above would be in an IDENT relation with [Regular rate] in *Heart: Regular rate and rhythm* elsewhere in the note.

A few more notes on Vital Signs sections:

- The dates listed (*March 14, 2015*) should be marked as SECTIONTIMEs, not TIMEX3s. These do not need to be linked to anything in the section, because we determine DocTimeRel based on the relation it has to SECTIONTIME, so the relation to SECTIONTIME is already understood. (The same is true for DOCTIME.) The measuring EVENTs in this section should be OVERLAP.
- We do not mark numerical test results, so the measured values (represented by *X*s in this example) should be left unmarked, with the exception of the pulse rate value. This should be marked as a TIMEX3 of type SET, as it's a frequency – {X/min}.
- If the words *Date/Time* are marked as either MARKABLEs or TIMEX3s, these should be deleted.

5) Non-annotated sections

The following sections were annotated by T1Coref but not by T1Temp:

- 20116 - Advance Directives
- 20104 - Current Medications
- 20105 - Allergies
- 20138 - Patient Education
- 20148 - Patient Consent
- 20123 - Patient Diet and Nutrition

As a result, some MARKABLEs and coreference links are present for these sections, but there are no temporal links, TIMEX3s, singleton entities, SECTIONTIMEs, or EVENTs. For all of these sections, except meds and allergies (20104 and 20105), you should delete all MARKABLEs, after first deleting them from any associated links.

Note that sometimes the above sections don't appear in their own numbered section of the note, but are tacked onto the end of another section. They should still be identifiable by header, and you should still treat them the same way. This might look like the following:

Annotation of medications and allergies section is discussed below.

6) Events of location

Another scenario which you'll find marked inconsistently is when a thing is located somewhere remarkable. You should make changes as need be to make this type of context align with the following guidelines:

a) When an *EVENT's* existence or location is predicated, don't mark the adjective or preposition; the EVENT itself is all we need.

(228) The [mass] is located in the distal transverse colon.

- *located* is not an EVENT.

(229) [Ostomy] is present.

- *present* is not an EVENT.

(230) A small cyst is noted within the right kidney.

- *within* is not an EVENT.

b) When an *entity's* location is predicated, mark the adjective if present. If not, mark the preposition. Do not mark the entity as an EVENT.

(231) [25 lymph nodes]$_M$ are [present]$_E$ within the mesenteric fat

(232) [25 lymph nodes]$_M$ are [within]$_E$ the mesenteric fat.

(233) Patient has noticed [blood]$_M$ [in]$_E$ [her stools]$_M$ over the last eight weeks.

Recall from Implicit EVENTs that something abnormal in the body is always an (implicit) EVENT (medical devices, tumors, etc.); normally occurring things in the body are nearly always entities (blood, stool, fat, lymph nodes, etc.), even if their location in the body is eventive.

7) Conjoined NP and list NP MARKABLEs

Sometimes you'll encounter MARKABLEs consisting of conjoined noun phrases:

(234) Scan of **[abdomen and pelvis]** was clear.

We understand these as referring to a whole, connected region. They should be kept, and you should also add WHOLE-PART relations where necessary between the conjoined noun phrase and its parts – i.e., [abdomen and pelvis] should be in a WHOLE-PART relation with [abdomen] and [pelvis].

Additionally:

- Don't go looking for other PARTs for the conjoined NP region besides the actual members of the conjoined phrase. For the example above, [abdomen and pelvis] is likely already in a WHOLE-PART relation with one or more other PARTs. Simply add [abdomen] and [pelvis] as PARTs for that link and stop there.
- **Don't create any new conjoined NP MARKABLEs.**
- If you encounter a conjoined noun phrase that represents two body parts that are actually disjoined – e.g., [hands and feet] – delete it and any associated relations.
- Delete all "list" NP MARKABLEs, i.e., a single MARKABLE that consists of a list of referents like: [the cecum, the terminal ileum, and the descending colon]. Remember to delete it from associated links first.

8) Test results

In early THYME, test results were not marked as EVENTs unless they referred to a specific diagnosis (e.g., *scan showed a [tumor] in the colon*). We now want to capture all test results. If you find unmarked test results, mark them as EVENTs, as shown:

(235) X-ray was [unremarkable].

(236) MRI was [normal].

(237) Margins are [negative].

However, we do not mark the terms *negative* and *positive* if the phrase also contains the actual thing that has been confirmed or ruled out:

(238) CT [scan] was negative for [metastases].

Here [scan] and [metastases]$_{NEG}$ are the only EVENTs. We don't need to mark *negative* itself because this information is captured by the negative polarity on [metastases].

Finally, recall that numerical test results should not be marked.

9) Spans

You may find some markables that include extra spaces or punctuation in their spans. These should be deleted: [Heart], not [Heart:]

Possessive MARKABLEs should include the apostrophe and following "s" in their span, however: [Mr. Smith's], not [Mr. Smith]'s.

Annotating medications and allergies sections

Medications sections (clinical section ID 20104) and allergies sections (clinical section ID 20105) were annotated by T1Coref but not by T1Temp. We need to complete annotation for these sections.

Importantly, for both sections, we're NOT annotating:

- TIMEX3s
- TLINKs (besides CONTAINS-SUBEVENT)
- ALINKs
- Formulaic/copied-and-pasted headers and final sentences, e.g.:
  - *Indication:*
  - *Instructions:*
  - *Medication:*
  - *Non-medication / Food:*

- *Radiology:*
- *Allergies above current as of [Tuesday, November 6, 2011 at 11:53 AM]*
- *These are the patient's medications as of [Thursday, July 8, 2014 at 4:34 PM]*

Exceptions:

- [the patient's] should be marked in the preceding sentence.
- SECTIONTIME will already be marked for both sections, as shown above by brackets.

What we *are* annotating are all EVENTs, entities, coreference links, and the CONTAINS-SUBEVENT TLINK, discussed in detail below. Medications and allergies sections necessarily come with a lot of guidelines, due in part to the need to synchronize the inherited annotations from T1Coref with the annotations from both projects in the rest of the note, and due to the fact that they tend to be auto-filled, template-style sections which the doctor can also interrupt at any time with current, actual information.

1) Medications sections

Medications sections look like this:

(239) Ferrous [Sulfate] 100-mg tablet 1 [TABLET] by [mouth] two times a day.

[Advil]100-mg tablet 1 [TABLET] orally as [directed] by [prescriber] as-[needed].

Indication: [headaches].
Instructions: Take one [tablet] at [onset] of [headache]. [Patient] has not [taken] in the last two years.
These are [the patient's] medications as of [Wednesday, December 21, 2013 at 12:34 PM].

This example demonstrates how a medications section should look once it's been fully annotated for all markables. Annotation details are discussed below. EVENTs are shown in blue, MARKABLEs in black, SECTIONTIME in red.

1) Consider the first line above:

Ferrous [Sulfate] 100-mg tablet 1 [TABLET] by [mouth] two times a day.

[Sulfate] is considered to be the event of having the prescription for that *type* of medication. [TABLET] is considered to be the event of having the prescription for that particular *dose* of that medication. They should therefore be marked as OVERLAP, ACTUAL (referring to the event of the patient having the prescription), and should be in a SET-SUBSET relation, where [Sulfate] is the SET and [TABLET] is the SUBSET. This S-SS linking should occur for all these list-form medication events.

Because these were annotated by T1Coref but not by T1Temp, these will frequently appear as MARKABLEs instead of EVENTs when you open a file. You should convert them to EVENTs since we consider all medication references to be eventive (Implicit EVENTs), and since we're wanting to align our schema categories with our conceptual categories (Converting eventive MARKABLEs to EVENTs). (A script has been run to automatically convert some of the most frequent terms to EVENTs. You should still briefly check these for accuracy.)

So, what you'll actually find in the preannotated data will look more like this:

[Ferrous sulfate 100-mg tablet]$_S$ *[1 TABLET by mouth two times a day]*$_{SS}$

As with all other MARKABLE conversions, you'll follow the same process of creating the EVENT; substituting it for the MARKABLE in all associated links; and then deleting the MARKABLE. (Sometimes you'll find the data have a W-P link between the type event and the dosage event; we want them all to be S-SS.)

However, remember we're going from a maximal-span MARKABLE to a minimal-span (i.e., headword) EVENT. Because syntax for these medication phrases is often hard to parse, the following discussion should guide your decisions about headword selection:

a) When the specific name of the medication is present, choose that as the headword (rather than *tablet*, *capsule*, etc.):

- [Omeprazole] 20-mg capsule enteric-coated
- [Synthroid] 125 mcg tablet

b) Follow this practice even when the medication name could be construed as a more straightforward premodifier:

- [Ferosul] tablet
- [Multivitamin] tablet

c) When the medication name consists of more than one word, follow normal headword-finding practices of selecting the rightmost term, unless there's a true postmodifier as in the "Vitamin D-3" example:

- Flaxseed [Oil] 1,000 mg capsule
- Metoprolol [Tartrate] 100-mg tablet
- [Vitamin] D-3 1000 unit tablet

d) Sometimes the brand name is also provided in brackets or parentheses. Ignore all bracketed or parenthetical terms (i.e., don't mark them as EVENTs, and therefore don't worry about creating APPOSITIVE relations for them):

- [Nitroglycerin] (NITROQUICK) 0.4 mg tablet sublingual
- Metoprolol [Tartrate] (LOPRESSOR) 100-mg tablet

e) When the medication name is not present (as is the case for the dosage events), choose the true syntactic head:

- 1 [TABLET] by mouth two times a day
- 1 [capful] by mouth as-needed

f) Once in a while the head will be elided, or a standardized unit of measurement will be used. This is the only situation in which we're permitting marking numbers and units of measurements as EVENTs:

- [1] by mouth two times a day
- 0.8 [mL] every-other-week

You'll also find other EVENTs and entities that aren't marked at all. Recall that that's because T1Coref only annotated markables that had a coreferential relation with something else in the note. We now want to fully annotate the medications section for all EVENTs and entities.

2) Medications EVENTs that are not in this list format are typically considered to be EVENTs of the patient actually taking (or not taking) that medication. You should put these usage events in CONTAINS-SUBEVENT relations with the prescription-type event. So if there happens to be a mention of the patient taking ferrous sulfate elsewhere in the note, it should be in a CON-SUB link with [Sulfate] in the medications section, where the prescription event [Sulfate] is the containing EVENT, and the usage event is the SUBEVENT. The prescription EVENT can be thought of as an "umbrella" event that contains all the usage EVENTs.

3) When are usage EVENTs the same and when are they different?

Say you find the following elsewhere in the note:

> (240) Patient was on ferrous [sulfate] but we have put the [iron] on hold until after surgery. Patient will restart [iron] postoperatively.

First, note this demonstrates the common practice in these notes of using different words to refer to the same type of medication. Second, note there are three different sulfate EVENTs here, in order:

- [sulfate]$_{\text{BEFORE, POS, ACTUAL}}$ – the past event of the patient taking sulfate
- [iron]$_{\text{OVERLAP, NEG, ACTUAL}}$ – the current event of the patient not taking sulfate
- [iron]$_{\text{AFTER, POS, ACTUAL}}$ – the future event of the patient taking sulfate

These are not IDENTICAL, because they all point to different taking EVENTs on the timeline. Each should remain unlinked to each other, but each should be linked as a SUBEVENT to the Sulfate prescription EVENT in the list in the medications section:

- [Sulfate] 100-mg tablet CONTAINS-SUBEVENT [sulfate]$_{\text{BEFORE, POS, ACTUAL}}$
- [Sulfate] 100-mg tablet CONTAINS-SUBEVENT [iron]$_{\text{OVERLAP, NEG, ACTUAL}}$
- [Sulfate] 100-mg tablet CONTAINS-SUBEVENT [iron]$_{\text{AFTER, POS, ACTUAL}}$

If this is confusing, consider the following non-clinical example:

(241) It [snowed] yesterday, it's not [snowing] today, it's going to [snow] tomorrow.

This is analogous to the sulfate example, but a bit more intuitive – there are three different snowing events here that we would not want to say are the same.

Of course, this doesn't mean that two usage EVENTs are never referring to the same thing. You might find two different references to the sulfate the patient will start taking again after surgery. These should go in an IDENTICAL chain, since they're talking about the same taking-sulfate EVENT.

4) As with other CONTAINS-SUBEVENT links, a negated medications SUBEVENT may still be put in a CON-SUB link. Note that the negated sulfate EVENT above is still a SUBEVENT of the Sulfate prescription EVENT.

5) Indication EVENTs, such as "[headaches]" above, should always be labeled as OVERLAP, HYPOTHETICAL, and therefore shouldn't be linked to anything else in the document (more on this in Coreference links and modality). This is because we often don't know if the patient is actually experiencing the indication or not. We maintain this practice for all EVENTs that follow the Indication header, even if you can corroborate the experience of that elsewhere in the note.

6) Instructions headers most often introduce EVENTs that should be marked as OVERLAP, HYPOTHETICAL, since they are often copied-and-pasted instructions, and we typically don't know if the patient is actually following them or not. All the EVENTs here should be HYPOTHETICAL:

Take one [tablet] at [onset] of [headache].

Watch out, though! Instructions paragraphs sometimes include ACTUAL EVENTs, like [taken]$_{\text{NEG}}$ in the example above. These should be marked and linked as appropriate:

[Patient] has not [taken] in the last two years.

Here, [used]$_{\text{NEG}}$ would be in a CONTAINS-SUBEVENT link with the prescriptive [Advil] EVENT.)

For Instructions EVENTs, we operate under the general guideline that if, in the phrase in question, there is some kind of evidence that the events actually refer to the patient instead of being formulaic or copy-and-pasted instructions, then we can assign ACTUAL modality; otherwise stick with HYPOTHETICAL. Evidence for using ACTUAL modality may include, but isn't limited to: the presence of a specific date or time; an EVIDENTIAL; or a mention of an EVENT or entity that we can corroborate elsewhere in the note as actually happening or existing.

Evidence for assigning ACTUAL modality does not include use of the present tense, since formulaic instructions are often given in the present tense; in the following, [Take] would be HYPOTHETICAL:

Instructions: [Take] with meals

The challenge of medications (and allergies) sections is that often a template is followed, but the doctor can freely add information as well, making our job difficult when it comes to deciding what's actual and what's not.

7) [prescriber] shouldn't be linked to any other doctor in the note, unless it's explicitly stated who prescribed the medication. We don't have enough information otherwise to know which doctor prescribed which med. References to [the patient] and various body parts ([mouth], [tongue], etc.) may always be understood as referring to the actual patient and patient's anatomy, and linked accordingly with mentions in the rest of the note.

8) Dose increases and decreases also "count" as medication prescription SUBEVENTs; if you have "Ferrous [sulfate] was [decreased]" elsewhere in the note, [decreased] would be a SUBEVENT of the [sulfate] usage EVENT, rather than the [Sulfate] prescription EVENT. (This is because that ongoing event of the patient taking sulfate includes the dose decrease; it's not that the patient stopped and re-started the medication, they simply changed the dose.)

9) Sometimes there are two separate prescription references for the same medication, e.g.:

(242) Ferrous [Sulfate], 100-mg tablet 1 TABLET by mouth daily

Ferrous [Sulfate], 50-mg 1 TABLET by mouth daily

These should be linked as IDENTICAL. According to our consulting doctor, the medication is sometimes prescribed like this so that the patient can adjust the dose as needed. But the medication itself is the same, so we can understand the prescription events to be the same.

10) Finally, don't create any link between a collective medications reference and individual prescription events in the medications section:

(243) Patient continues **treatment** for diabetes

…

[section 20104]

Glucophage...

Novolin...

No link is possible here – even though [Glucophage] and [Novolin] are both used to treat diabetes – since [treatment] is a collective usage reference. This is because: a) we can't say, for example, [Glucophage] CON-SUB [treatment], because the treatment consists of Novolin as well, and Novolin isn't a subevent of Glucophage; and b) we can't say that [treatment] CON-SUB [Glucophage] because this contradicts the *prescription>usage* hierarchy that we've otherwise established. Under our current schema, it's best to simply leave these unlinked.

2) Allergies sections

Allergies sections look like this:

(244) Medication :

NO KNOWN MEDICATION [ALLERGIES]

Non-Medication / Food :

[Avocado] - [vomiting]

Cat [hair] - [hives]

Radiology :

NO KNOWN CONTRAST [MEDIA]

Allergies above current as of [Monday, August 2, 2012 at 10:49 AM]

This example demonstrates how an allergies section should look once it's been fully annotated for all markables. Annotation details are discussed below. EVENTs are shown in blue, SECTIONTIME in red. There are no MARKABLEs in the above example.

1) As with medications sections, you'll find several events marked as MARKABLEs instead of EVENTs. You should change them to EVENTs as appropriate. We understand all references to allergens – even very entity-like ones like [Avocado] – to be the implicit EVENTs of encountering those allergens.

As with the medications section, you'll also find other EVENTs and entities that aren't marked at all. We now want to fully annotate the allergies section for all EVENTs and entities.

2) Note that most EVENTs in these sections should be marked as OVERLAP, GENERIC, as we assume unless stated otherwise that the patient is not actually taking (or encountering) the allergen, nor are they actually experiencing a reaction:

- [Avocado]$_{\text{GENERIC}}$ - [vomiting]$_{\text{GENERIC}}$
- Cat [hair]$_{\text{GENERIC}}$ - [hives]$_{\text{GENERIC}}$

The only time we're linking a GENERIC markable to anything is when it has a SET-SUBSET relation with an ACTUAL markable (review SET-SUBSET if necessary). If elsewhere in the note you had something like "Patient ate [avocado] and [vomited]," you'd create S-SS links:

- [Avocado]$_{\text{GENERIC}}$ S-SS [avocado]$_{\text{ACTUAL}}$
- [vomiting]$_{\text{GENERIC}}$ S-SS [vomited]$_{\text{ACTUAL}}$

This will be rare, but does occur. You should delete any other links you find between GENERICs.

3) References to allergies themselves (as opposed to allergens), though, should be interpreted as such:

- Patient has many [allergies]$_{\text{ACTUAL, POS}}$
- NO KNOWN MEDICATION [ALLERGIES]$_{\text{HEDGED, NEG}}$
- NO KNOWN CONTRAST [MEDIA]$_{\text{HEDGED, NEG}}$

Note that we interpret [MEDIA] in this commonly-encountered phrase as referring to "media allergies." (Contrast media is often used in certain types of tests like MRIs.)

4) Finally, as stated in the introduction to this section, we don't mark headers or copied-and-pasted sentences like *Medication :*, *Non-Medication / Food :*, *Radiology :*, and *Allergies above current as of*...

Coreference links checks

This section lists several specific coreference issues that should be checked for and fixed if found. Remember you should also check each coreference link for accuracy.

1) Coreference links and modality

In this project, we are not doing coreference linking for HYPOTHETICAL or GENERIC markables. If you find a coreference link between HYPOTHETICAL or GENERIC markables, you should delete it. (Remember it's fine for ACTUAL and HEDGED markables to be in the same link with each other.) Of course, you should still check to make sure, say, a

HYPOTHETICAL markable isn't mixed up in the same chain as several ACTUAL markables; if it is, simply delete the HYPOTHETICAL markable from the chain.

The one exception to this: Recall that we permit linking GENERIC events to ACTUAL events with SET-SUBSET, when appropriate. For example:

> (245) Patients with [cancer]$_{\text{GENERIC}}$ are advised to avoid this medication. Our patient has kidney [cancer]$_{\text{ACTUAL}}$.

Here we'd still want to create a SET-SUBSET relation, where the GENERIC [cancer] is the SET (referring generally to all cancer events) and the kidney [cancer] is the SUBSET (referring to the patient's particular instance of cancer). Note that if there were multiple GENERIC [cancer] mentions, you'd have to link each one to the patient's ACTUAL cancer using S-SS (since there's no IDENT link between the GENERIC references, we can't infer the S-SS relation for each mention).

In other words, we still want to capture all possible links for ACTUAL and HEDGED EVENTs, but we do not need to link HYPOTHETICAL EVENTs to other HYPOTHETICAL EVENTs, nor do we need to link GENERIC EVENTs to other GENERIC EVENTs. Determining coreference for certain non-real EVENTs can be nearly impossible, and these relations are much less important to capture than the ones for actual (and hedged) EVENTs.

If, however, you happen to find a **TLINK** or **ALINK** that links HYPOTHETICALs or GENERICs, this is fine; leave it as is. This instruction is only for **coreference links** and for **the new TLINK type, CONTAINS-SUBEVENT.**

Finally, keep in mind that even though we don't have features for entities, we still have to determine their modality/polarity/etc. for the purpose of creating accurate coreference links:

> (246) I discussed with [the patient] the typical course of treatment for [patients with these comorbidities].
>
> a. [patients with these comorbidities] S-SS [the patient]

The doctor is speaking generally about patients with certain comorbidities, so this is a generic reference, even though that's not explicitly marked on the MARKABLE.

2) "Extra" PARTs and SUBSETs

Consider the following example from a preannotated note:

> (247) I discussed with the patient that the number of [growths] in his colon is concerning. The largest [mass] was removed during colonoscopy. Pathology showed this [mass] was benign.

a. [mass] IDENT [mass]

b. [growths]$_{S}$ - [mass]$_{SS}$, [mass]$_{SS}$

The key issue to observe here is that even though the two mentions of [mass] are appropriately linked as IDENT, both are included as SUBSETs in the S-SS link with [growths]. You will occasionally encounter cases like this, where multiple mentions that point to the same PARTs and SUBSETs are included in W-P and S-SS links.

While not inaccurate, this clutters the data and will make cross-document linking much more challenging. It's also unnecessary, since we can infer that if, say, two mentions of [mass] are IDENT, and one of them is a SUBSET of a SET, the other mention is therefore also a SUBSET.

Therefore, you should delete these "extra" mentions when you find them, leaving just **a single mention of the same referent for each relation**. Be sure, though, that the "extra" mentions you delete are in an IDENTICAL chain with the mention that remains! Also keep in mind that different terms can point to the same referent – a mass may elsewhere be referred to as a tumor or a carcinoma, but the EVENT is the same.

3) Equational clauses

Be especially on the lookout for EVENTs that serve as the predicates of equational clauses (that is, the *Y* EVENT in an *X is Y* construction). Because T1Coref did not mark these EVENTs, but T1Temp did, some type of coreference link will need to be added for nearly all of these EVENTs.

The type of link will depend on the semantics of each specific phrase. We will follow the RED (Richer Event Description) project's treatment:

- If *X is Y* is actually showing the equivalence of the identities of two mentions, then these are IDENT:

  (248) [One] of the growths was a [cyst] in the gallbladder.

- If *X is Y* asserts a SET-SUBSET relation, mark it as such. The "such-as" test is a good metric in this case:

  (249) The patient had a small [tumor] removed during endoscopy. Pathology showed [this] was a benign [mass].

Here, [tumor] is IDENT with [this]. The [mass] reference is a SET, and [this] is the SUBSET. You could re-state the relationship in the *X is Y* clause as "Benign masses such as this one…" (Therefore, [mass] should also have a modality of GENERIC, since it's not identifying a specific set of masses.) In other words, in our view, "benign [mass]" here is identifying a *type*.

The difference between this and the preceding example is that we can think of “benign masses” as a type, but we can’t do that with “a cyst in the [i.e., *this patient’s*] gallbladder.” We could substitute the definite article in “the gallbladder” with a possessive – “his gallbladder” or “Mr. Smith’s gallbladder.” We couldn’t accurately re-state the relationship as “Growths in his gallbladder such as this one” because there is only one growth in his gallbladder and it is the cyst.

In other words, if the *Y* clause includes EVENTs or entities that refer specifically to the subject (*X*) in question, the relation should be IDENTICAL.

- If X is Y is very predicative, then do not mark any relation at all.

(250) The [tumor] is very [large].

No relation.

Finally, because THYME did not mark entities, entities that function as the predicates of equational clauses have not been annotated at all. These should be marked as MARKABLEs and included in links, following the guidelines just described.

(251) [Ms. Brown] is a 47-year-[old] female with breast [cancer] and colon [cancer].

The phrase *a 47-year-old female with breast cancer and colon cancer* should be added as a MARKABLE (remember we capture the full phrasal span for entities), and in this case it would be in an IDENT relation with [Ms. Brown], because the ACTUAL EVENTs of [cancer] and [old] identify this as referring specifically to that patient.

4) Hard-to-spot SETs

Most SET-SUBSET relations are easily identifiable. Some SETs are grouped by what the members **are** (*patient has had three different cancers: colon cancer, kidney cancer, skin cancer*), while others are grouped according to some kind of shared function or attribute of the members (*the patient’s symptoms are pain, gas, bloating*). But both of these types are fairly intuitive.

The equational clause discussion above identifies one less intuitive SET-SUBSET relation, and there are two more you should watch out for:

a) “Chief complaint” should always be S-SS with whatever the chief complaint is

(252) Patient presents with severe abdominal [pain] for two weeks. We will run tests to help identify the source of his chief [complaint].

a. [complaint] S-SS [pain]

Doctors use this term somewhat imprecisely/variably, so we don’t want an IDENTICAL relation here as that could run us into trouble in the cross-document annotation stage.

b) Always use S-SS to link [positive] and [negative] test results to the things they're positive and negative for.

(253) Margins are [negative]. ...Surgical margins are negative for [dysplasia]$_{NEG}$ and [tumor]$_{NEG}$.

a. [negative] S-SS [dysplasia]$_{NEG}$, [tumor]$_{NEG}$

This is also to facilitate smoother cross-document linking.

(Note that [positive] and [negative] are themselves only marked as EVENTs when that's all the sentence tells us about the test results. They aren't marked when the conditions that were tested for are explicit in the text – positive for metastatic [disease]. This is fine.)

Additionally, [positive] and [negative] should themselves never be SUBEVENTs of cancer.

5) WHOLE-PART links

LOCATIVE W-P

Recall that WHOLE-PART links capture when one entity is compositionally part of another entity. In clinical notes, these will likely most frequently occur between body parts – [eyes]$_W$ and [pupils]$_P$, for example. WHOLE-PART links are also referred to here as "compositional WHOLE-PART" or "normal WHOLE-PART" links.

You'll find the data also appear with preannotated SUBEVENT WHOLE-PART and LOCATIVE WHOLE-PART links. SUBEVENT W-P links occur between two EVENTs – e.g., [evaluation]$_W$ - [MRI]$_P$. SUBEVENT W-P links are different from the CONTAINS-SUBEVENT links we are adding in this pass. **They should be entirely ignored.** To that end, when you open a new note, go to the menu option in the program's left panel and click on the CorefChains links to view the menu of coreference links. From there, de-select the SubeventWP box

LOCATIVE links occur between an entity and an EVENT, typically a body part and an abnormality – e.g., [leg]$_W$ - [pain]$_P$ or [liver]$_W$ - [tumor]$_P$. Due to idiosyncrasies of the previous projects and the data merger, there are some compositional WHOLE-PART relations that have been inaccurately labeled as LOCATIVE W-P, and some locative relations that have been inaccurately labeled as compositional W-P. Importantly, **while the LOCATIVE links will not be part of this project's output, the compositional W-P links are, and therefore they should be extracted from the mixed-up links and preserved.**

First, note that locative WP and compositional WP are entirely different relations. For a simple example, consider the room you're sitting in right now. The ceiling of that room is part of the room. If you removed the ceiling, you've taken away part of the room. This is a compositional W-P relation. But now consider your computer's relation to the room. It's located in the room, but it's not part of the room. You could take it out of the room and nothing about the room would be altered. This is a locative relation.

For a clinical example, compare the semantic relation [the patient’s colon] has to [abnormal growths] with the relation it has to [the sigmoid colon]. The [growths] are abnormalities that are located in the colon – even attached to it – but are in no way part of the normal composition of the colon. The sigmoid colon, however, is an essential part of the structure of the colon. So [the patient’s colon] and [the sigmoid colon] are in a compositional whole-part relation, but [the patient’s colon] and [abnormal growths] are in a locative whole-part relation.

So, say you find the following:

(254) [Tumor]$_E$ in [the colon]$_M$ invades [the colonic wall]$_M$.

a. [the colon]$_W$ LOCATIVE W-P [Tumor]$_P$, [the colonic wall]$_P$

As you can see, the current annotation says [the colon] is in a LOCATIVE W-P relation with both [Tumor] and [the colonic wall]. While this is accurate for [Tumor] – an EVENT – the colonic wall is actually part of the colon, so this should be in a normal WHOLE-PART relation. The process here, then, is:

a) Create a new, compositional W-P relation between [the colon] and [the colonic wall].

b) Delete [the colonic wall] from the LOCATIVE W-P link.

As mentioned above, typically these different types of PARTs – locative PARTs and compositional PARTs – will be straightforwardly represented by the difference between EVENTs and entities. At the risk of oversimplifying: Compositional W-P links should most often occur between *[normal body part]*$_W$ - *[other normal body part]*$_P$. LOCATIVE W-P links should most often be between *[normal body part]*$_W$ - *abnormal [event/implicit event]*$_P$.

In the example above, [Tumor] is an EVENTs in the data, representing an abnormality that occurs somewhere along a patient’s clinical timeline; it is not compositionally part of the patient’s colon. [the colonic wall] is an entity (MARKABLE); this is in a compositional whole-part relation with the colon. But remember this won’t always be the case, since sometimes MARKABLEs conceptually refer to events (refer back to Converting eventive MARKABLEs to EVENTs)! In that case, you’ll have to convert the MARKABLE to an EVENT and then make the W-P extraction as appropriate.

In addition to checking LOCATIVE W-P links for compositional relations, you should also do the reverse – check normal W-P links for locative relations. These should be deleted from the normal W-P links if found. However, because locative relations aren’t part of this project’s output, **you do not need to spend time creating new LOCATIVE links**. If, say, you find [tumor] in a normal W-P relation with [the colon], simply delete [tumor] from the link. Don’t create a LOCATIVE W-P link. Similarly, you do not need to spend any time checking Locative W-P links for accuracy, apart from extracting compositional relations from them.

A final word on this topic: Making this choice for references to [specimens] and similar terms can be difficult since specimens can of course consist of both normal tissue (entities) and abnormal tissue (EVENTs). Because we've decided to always treat these terms as entities (Implicit EVENTs), you should make sure compositional WHOLE-PART links for specimens consist only of other entities. This means if, say, the specimen consists partially of adenocarcinoma tissue, the specimen should not be WHOLE-PART with the adenocarcinoma reference. (LOCATIVE WP is fine.)

Otherwise, assume that preexisting anatomical site WHOLE-PART links are accurate

Our goal is to separate locative relations from compositional relations. You do not have to also check to make sure the compositional relations for anatomical sites are accurate.

Combined with the previous step, this'll play out as follows:

You find a LOCATIVE W-P link between [the colon]$_W$ and [the cecum]$_P$. Since they are both entities, you recognize that this should be a compositional relationship. You can simply create a normal WHOLE-PART link between [the colon] and [the cecum]. You do not have to do a lot of searching online or review of high school anatomy textbooks to make sure that the rectum is in fact part of the colon!

(Of course, if you encounter a blatant error that you are 100% sure is wrong, you may fix it – if [eye] and [lungs] are in a WHOLE-PART relation, delete it.)

The one exception is that lymph nodes, which should always be marked and treated as entities, should never be in WHOLE-PART relations with anatomical sites. If you find [lymph nodes] in a compositional W-P link, you should delete it from the link. If you find [lymph nodes] as part of a LOCATIVE W-P link, you should leave it.

(255) [Multiple lymph nodes] are identified within [the mesenteric fat].

There should be no compositional W-P link between these two MARKABLEs. If the data appear with a LOCATIVE WP link for it, that's fine – follow the practice of **ignoring LOCATIVE links apart from checking for compositionality**, and move on.

(Note that LOCATIVE W-P links are the only ones where there'll be a mix of EVENTs and entities in the same link. This is fine. But remember that *EVENTs and entities should never be in the same link for all relations that will be part of this project's output – IDENTICAL, APPOSITIVE, WHOLE-PART, SET-SUBSET, and CONTAINS- SUBEVENT*.)

6) APPOSITIVEs

The following error commonly occurs in the data:

(256) [Alicia O'Connor, M.D.]

This MARKABLE should be divided into two MARKABLEs, [Alicia O'Connor] and [M.D.], and an APPOSITIVE relation created between the two.

7) Doctors

Any IDENTICAL links between the authoring doctor and the referring doctor should be deleted. The authoring doctor of course is all the first-person mentions in a note – the "I" doctor. Section 20101 is the Referral Source section. The doctor mentioned here is the doctor who referred the patient to the current place of care, and therefore is not the same physician as the one writing the note.

8) "X negative for X"

Consider the following:

(257) [Lymph nodes] negative for [tumor].

This may seem oddly specific, but this construction comes up fairly frequently and was interpreted differently by the two prior projects.

We will follow T1Temp's read of [tumor] here as a NEGATIVE, ACTUAL EVENT – i.e., there is no tumor in the lymph nodes. You should base your decisions about the accuracy of associated links on this interpretation. For example, because T1Coref understood this construction differently, you'll likely find [tumor] in IDENT chains with GENERIC, POSITIVE tumors. These relations should obviously be deleted based on our interpretation.

TIMEX3, DOCTIME, and SECTIONTIME checks

The first five "checks" listed below should be looked for in each document and changed if found. The remaining five can be treated more as reference – if you run into something that strikes you as being marked oddly, you can refer to these sections and change the annotation accordingly, if need be. But don't spend a lot of time agonizing over them.

1) TIMEX3s marked as MARKABLEs

You'll sometimes run into MARKABLEs that represent temporal expressions, rather than events or entities. These should simply be deleted, along with any associated links.

(258) Patient had surgery on [March 1, 2014]$_M$. His recovery since [March 1]$_M$ has

been smooth.

a. [March 1, 2014] IDENT [March 1]

Delete the link and both MARKABLEs here. You do not need to create anything instead. These will typically already also be tagged as TIMEX3s as well.

Similarly, any coreference chains made up of TIMEX3s should be deleted, and any TIMEX3s that appear as part of a coreference chain should be deleted from the coreference chain. (Note this does not include pre- and post- expressions that are tagged as EVENTs.)

2) PREPOSTEXPs

These have already been discussed in Most common eventive MARKABLEs.

3) DOCTIME

Every document has a line like this at the very top:

(259) meta rev_date="11/02/2013" start_date="**11/02/2013**" rev="0002"

DOCTIME should always be the date associated with *start_date*, and it should only be that date. If you find the meta rev_date (or any other date in the document) marked as DOCTIME, delete it and make sure the start_date is DOCTIME.

4) SECTIONTIME

Certain clinical sections in these notes have their own SECTIONTIME, which essentially functions as “DOCTIME” for that section. The DocTimeRel of the EVENTs within that section have been assigned relative to SECTIONTIME, not DOCTIME. Therefore, we don’t need TLINKs between the SECTIONTIME and the EVENTs within that section; their temporal relationship is already known from DocTimeRel. So, you may delete SECTIONTIME TLINKs if found (these will most likely show up in the Vitals Signs sections).

There’s one exception to this, however: In cases where a SECTIONTIME does not hold for a single, clinically-delineated section, but appears within another section, we need to preserve or create TLINKs between the SECTIONTIME and a related EVENT that precedes the SECTIONTIME itself.

(260) [start section id="20103"]

We are waiting final biopsy results.

Followup [February 21, 2012]:
Biopsies returned negative.

[end section id="20103"]

Usually each note or section of note corresponds to a single visit/interaction with the patient. As you can see, that's not always the case – here, the doctor has tacked the subsequent visit information onto the section from the previous visit. So [February 21, 2012] has appropriately been marked as a SECTIONTIME, marking the different time of the later visit.

All EVENTs following the SECTIONTIME should have DocTimeRel assignments relative to the SECTIONTIME rather than DOCTIME. All EVENTs preceding it should have DocTimeRel assignments relative to DOCTIME.

The issue is what to do with [Followup]. We can train a system to know that DocTimeRel assignments for EVENTs following a SECTIONTIME are made relative to SECTIONTIME rather than DOCTIME (up until another SECTIONTIME or until a clinically-delineated new section begins). But [Followup] precedes SECTIONTIME. We could mark [Followup] as AFTER (i.e., AFTER DOCTIME), but this is not very intuitive. Instead, we want to mark DocTimeRel of [Followup] as OVERLAP, but also create a CONTAINS link saying [February 21, 2012] CONTAINS [Followup].

Similarly, in situations where there's no date to mark as SECTIONTIME, but an addendum to a section reports EVENTs based on a later time than DOCTIME, we will mark the "Addendum" header itself as an (implicit) SECTIONTIME.

(261) Patient will consult the Cardiology Department later today to make sure he is ok anesthesia.

[Addendum]

Dr. Jones from Cardiology called after consultation to say patient is ok for anesthesia.

We again want to mark the EVENTs' DocTimeRels as intuitively as possible – [consult] should be AFTER in the first sentence and [consultation] BEFORE in the second (and these two EVENTs should be IDENTICAL). But, we also have to mark [Addendum] itself as a SECTIONTIME; otherwise, we'd end up with a nonsensical timeline that says the same consultation was both BEFORE and AFTER DOCTIME!

Marking the headers will allow us to know that the DocTimeRel assignment for the following mentions are relative to a different time, even though we don't know what that time is. Note that the header itself may be called a variety of terms such as [addendum], [addition], etc.; mark the header as appropriate for this context regardless of what term is used.

Only mark headers as SECTIONTIME when absolutely necessary to avoid "breaking" the timeline.

5) Non-markable expressions

The following expressions were sometimes mistaken for TIMEX3s. They are marked below to demonstrate how they may appear in the data, but they should be deleted if found.

(262) {Date}/{Time}=Jan 6, 2010:

Words like "Date" and "Time" cannot be linked to actual calendar dates and times.

(263) At {the time of surgery} patient was not experiencing [nausea].

This phrase indicates a temporal relationship between two EVENTs – the surgery and the non-nausea – rather than identifying a location on a timeline, so {the time of surgery} should be deleted.

Compare with the following, though, where we mark {that point} as it's an anaphoric reference to a time:

(264) At {that point}, patient requested a second opinion.

6) Prepositions marked as TIMEX3s

This section and the four following may be used as a reference; you don't have to explicitly go looking for these scenarios.

(265) Patient will return {following} wisdom teeth surgery.

Prepositions like *following*, *after*, *during*, *while*, *before*, etc., should not be marked at all. They indicate temporal relationships between two markables, not points on a timeline. These and related links should be deleted. These errors should be rare.

Relatedly, prepositions should never be the headword of a TIMEX3:

(266) Patient will return {in four weeks} for chemotherapy.

{four weeks} is the span here, not {in four weeks}.

Finally, pay attention to whether a preposition separates two different TIMEX3s or is part of an expression designating a single TIMEX3:

(267) During {the month of July}, she will come visit.

(268) From {May 1st} to {the 3rd}, she will refrain from eating solid food.

In (267), {the month of July} points to a single location on the timeline – it's simply another way of saying "July." In (268), there are two different timepoints – May 1st and May 3rd. Rather than annotating this as {May 1st to the 3rd} and labelling it a DURATION, we mark the dates separately, as shown, and both are DATEs. (Note we can still capture the durative nature of the EVENT by saying that [eating]$_{NEG}$ BEGINS-ON {May 1st} and ENDS-ON {the 3rd}.)

7) TIME versus DATE

Any temporal expressions that point to explicit times of the day should be marked as TIME rather than DATE, even if a calendar date mention is part of the span. (Remember that two adjacent temporal expressions that together identify a single time are annotated as a single TIMEX3.)

(269) {June 11, 2014 10:30}

(270) Cardiology will see her {this afternoon}.

The preceding TIMEX3s are both TIMEs.

8) QUANTIFIER spans

If a QUANTIFIER appears in its own phrase, mark the entire phrase, just as we do for all TIMEX3s:

(271) The patient vomited {twice} before the surgery.

(272) We have seen Mr. Lastname {three times} for his ulcerative colitis.

(273) On {two to three incidents} she has had blood in the stools.

When a QUANTIFIER appears in the same phrase as the EVENT it numbers (typically as a premodifier), only include the number in the span. The quantifier isn't the headword of these phrases.

(274) The patient has had {two} [seizures].

(275) Mr. Lastname has made {three} [trips] to the ER for his ulcerative colitis.

Finally, a word about how treatment cycles should be annotated, since these terms come up frequently in the data:

(276) Patient has had {four cycles} of [FOLFOX].

Note that {four cycles} is the span of the QUANTIFIER ("cycles" essentially just means "times something happens") and [FOLFOX] is the EVENT. As with all QUANTIFIERs and the EVENTs they quantify, the two should not be related to each other by any temporal or coreference link.

"Cycles" should also be marked as an EVENT. See Chemotherapy and radiation.

9) SETs

SETs represent frequencies – the number of times an event occurs for a given period of time. They appear in a variety of ways.

SETs that include the full frequency information – that is, both the quantifier (the number of times an event occurs), and the timespan in which that event occurs that number of times – are the easiest to spot and mark:

(277) Will administer Lariam {twice daily}.

(278) Patient has checked into the ER {roughly three times a month}.

(279) Simvastatin ZOCOR 20-mg tablet 2 tablets by mouth {one-time daily}.

Sometimes the frequency information is discontinuous:

(280) At that point, he was experiencing {8 to 9} [stools] {per day} according to patient report.

a. {per day} = SET

b. {8 to 9} = QUANTIFIER

Temporal expressions like this one should be checked, as they may sometimes appear as DATEs or DURATIONs in the data. When these phrases – {a day}, {per day}, {daily}, {each day} – refer to how often an EVENT occurred, they should always be annotated as SETs. (Compare with "Patient will return in {a week} for MRI," where {a week} is a DURATION.) Note that the quantifying information is also separately marked as a QUANTIFIER.

Relatedly, remember that QUANTIFIER only applies for number of occurrences of an event, not for quantifiers like "She has two eyes." Sometimes it's ambiguous as to whether a quantifier is quantifying objects or quantifying events. Follow the rule of thumb that if something we consider to be an EVENT in our schema is being quantified, you should treat the quantifier itself as a TIMEX3 QUANTIFIER. If an entity is being quantified, do not mark the quantifier as a TIMEX3 QUANTIFIER. Because we interpret [stools] as implicitly pointing to the passing event in the example above, {8 to 9} should be marked as a QUANTIFIER as shown.

Finally, we still mark SETs for which the timespan information has been elided and we only have the quantifying information:

(281) Heart rate is {112}.

Compare with *Heart rate is {112 /min}$_{SET}$*; {112} in the preceding example represents the same thing, but the "per minute" is implicit.

10) DATE versus DURATION

DATEs refer to a point on a timeline (albeit sometimes a general one), whereas DURATIONs refer to a span of time. This difference is usually fairly clear and intuitive, as in the following:

(282) The patient continuously experienced nausea for {nearly two weeks}.

a. {nearly two weeks} = DURATION

(283) Since {August}, she has not had any episodes.

a. {August} = DATE

There is one specific situation, however, where the temporal expression may be interpreted as either a DATE or DURATION. Consider the following non-clinical example:

(284) Rachel Nelson was spotted in Manila Wednesday, {weeks} after her disappearance.

Here {weeks} could be viewed as either a DATE identifying the calendar date when Rachel Nelson was spotted (i.e., the same calendar date as Wednesday), or it could be interpreted as the period of time between when the disappearing event and the spotting event. For consistency's sake, we've determined that any temporal expression that occurs in this specific context should be treated as a DURATION:

{quantity of time} – preposition – noun phrase

weeks after her disappearance

In other words, when a temporal expression is followed by a prepositional phrase that tells the relation between the temporal expression and another time or event, the temporal expression is a DURATION.

You should change the class of TIMEX3 if you find these incorrectly marked in the data. This may also involve changing associated TLINKs. The following should be annotated as shown:

(285) She will return for [followup] {the day} before her [vacation].
a. {the day} = DURATION

b. {the day} BEGINS-ON [followup]

c. {the day} ENDS-ON [vacation]

Pathology note checks

While clinical notes discuss a patient's current state, medical history, plan of care, and so forth, pathology notes present the findings from studying tissue that was removed during some kind of procedure.

The following checks occur most frequently in pathology notes, but because parts of pathology notes may be paraphrased or directly quoted in clinical notes, you should also bear these in mind as you annotate clinical notes.

Note that there are no TLINKs in the pathology notes. This was intentional as there is little temporal information in these documents. You should not create new TLINKs for pathology notes, except for CONTAINS-SUBEVENT as appropriate.

That said, pathology notes will have very few SUBEVENT links. Of the four categories for which we annotate subevent relations, medications sections aren't present in pathology notes, and chronic diseases other than the cancer usually aren't discussed. For the remaining two categories – cancer events and cancer treatment events – pathology notes often talk about the subevents (resections, tumors, etc.), but rarely mention the "containing" EVENTs (surgery, cancer, etc.). We can't annotate what's not there, so there will be very few CONTAINS-SUBEVENT links for most path notes. (We'll be able to make these relations in cross-document annotation later on.)

The most frequent CONTAINS-SUBEVENT link you will find in these notes will likely be between an [Intraop] reference and subprocedures of the operation:

(286) Frozen Sect-[Intraop]

A. Soft tissue, [excision]:

B. Colon, [resection]:

Recall that “Intraop," short for "intraoperative," is considered a PREPOSTEXP; and recall that because we understand PREPOSTEXP EVENTs as referring to the root event within the term (i.e., the operation itself), [Intraop] counts as a general procedural term (see Surgical procedures and Most common eventive MARKABLES). So for this example we'd have:

a. [Intraop] CONTAINS-SUBEVENT [excision]

b. [Intraop] CONTAIN-SUBEVENT [resection]

Together these links say: “This resection and this excision were subprocedures of this operation.”

Following are the other specific issues you should watch for, particularly for pathology notes.

1) Delete relative positions

These are terms like: [2.7 cm from the distal surgical resection margin] – i.e., MARKABLEs that identify a location relative to an entity. These count as egregious errors and should be deleted (but first delete them from any links they're part of).

2) Lymph nodes and SET-SUBSET relations

Consider the following example:

(287) [Multiple ([5 of [31]$_{M}$]$_{M}$) regional lymph nodes]$_{M}$ show adenocarcinoma… [Five regional lymph nodes]$_{M}$ demonstrate adenocarcinoma.

a. [31] S-SS [5 of 31]

b. [5 of 31] IDENT [Five regional lymph nodes]

This example demonstrates the way the preannotated data will consistently appear for situations like this, where there is parenthetical information that tells us the number of a group of lymph nodes and a subset of that number.

Note that two of the MARKABLEs in the first sentence refer to the same entity – [5 of 31] and [Multiple (5 of 31) regional lymph nodes] both refer to the same five lymph nodes. Therefore, we want to delete the [5 of 31] MARKABLE and keep the full-span MARKABLE, [Multiple (5 of 31) regional lymph nodes].

However, the [5 of 31] MARKABLE participates in an accurate relation with [31], where [31] is the SET and [5 of 31] is the SUBSET. We want to preserve this relation, so before deleting [5 of 31], you should first change the S-SS relation such that [31] is the SET and [Multiple (5 of 31) regional lymph nodes] is the SUBSET.

3) Specimen, section, margin

Because these terms are prevalent in pathology notes, it bears repeating from Implicit EVENTs that the terms [margin], [section], and [specimen] are entities and should be converted from EVENTs to MARKABLEs where appropriate.

(288) [The surgical resection margins]$_{M}$ are negative.

(289) This final pathology report is based on the gross/macroscopic examination and the frozen section histologic evaluation of [the specimen(s)]$_{M}$.

(290) [All sections]$_{M}$ are submitted.

4) Delete formula paragraph links

Many of the pathology notes contain this copied-and-pasted paragraph:

(291) This final pathology report is based on the gross/macroscopic examination and the frozen section histologic evaluation of the specimen(s). Hematoxylin and Eosin

(H&E) permanent sections are reviewed to confirm these findings. Any substantive changes identified on permanent section review will be reflected in a revised report.

Because this paragraph doesn't contain information unique to the patient, we don't need to spend time annotating it. If found, you should:

- Delete any markables in this paragraph from any links they may be part of (and delete the entire link if it involves only markables from the paragraph). This includes both coreference links and TLINKs.
- Leave the markables themselves.

5) Tumors in different locations are not IDENTICAL

This is usually fairly intuitive – a [mass] in the colon isn't the same as a [mass] in the liver. But it sometimes gets harder to sort through in pathology notes:

(292) A. Rectum #1, biopsy: Grade 3 (of 4) [adenocarcinoma] consistent with colon [primary].

B. Rectum #2, biopsy: Grade 3 (of 4) [adenocarcinoma] consistent with colon [primary].

C. Rectum #3: Grade 3 (of 4) [adenocarcinoma] consistent with colon [primary].

a. [primary] IDENT [primary], [primary]

As shown, all the [primary] references here should be IDENT, since each one is talking about the primary colon tumor.

However, according to our medical consultants, the adenocarcinoma mentions refer to "pieces" of the tumor that all originally belonged to the same mass (the primary) but are now both separate from it and distinct from each other. Therefore, we treat these as different implicit EVENTs: the presence of adenocarcinoma in the part of the rectum biopsied in A, the presence of adenocarcinoma in the part of the rectum biopsied in B, etc.

(While we treat masses as implicit EVENTs so we can put them on a timeline, they also of course refer to physical objects. It may be easier to conceptualize this from the physical-object standpoint; to put it crudely, a lump of tissue in one place isn't the same as a lump of tissue in another place.)

Since none of the three mentions of [adenocarcinoma] should be IDENTICAL to each other, you should delete an IDENT chain if found. Rather, each mention of [adenocarcinoma], as well as a mention from the [primary] IDENT chain, should go in CONTAINS-SUBEVENT links with a mention of the overall cancer, if present.

We follow the same practice for negated tumor EVENTs:

(293) Liver, segment VI, biopsy: Negative for [tumor]$_{NEG}$.

Liver, segment II, biopsy: Negative for [tumor]$_{NEG}$.

a. [tumor]$_{NEG}$ IDENT [tumor]$_{NEG}$

Recall that a general rule of thumb we follow is to understand negated things in the context in which they're presented. Because the author is noting these two tumor references as not existing relative to two different locations, we don't want to link them as IDENT.

A couple other points on this topic:

- Any growth that's identified as being "metastatic" is never IDENT to the primary (original) tumor. "Metastatic" inherently means secondary growths that develop somewhere different from the original mass.
- Any growth that's discussed as being present in the lymph nodes is never IDENT to the primary tumor:

(294) Colon, sigmoid, resection: Grade 3 (of 4) [adenocarcinoma]...Multiple (13 of 27) lymph nodes positive for [adenocarcinoma].

a. ~~[adenocarcinoma] IDENT [adenocarcinoma]~~

- We also differentiate between recurrent tumors and primary tumors – a distinction of time rather than location. A recurrent tumor should not be linked as IDENT to the original tumor, even though, again, histologically they're the same.

You should look for and adjust links that do not align with these guidelines.

6) Postmodifiers versus separate noun phrases

Let's look at a similar example, this time paying attention to the MARKABLEs. You'll likely find them marked as follows:

(295) A. [Colon, transverse], resection: Adenocarcinoma...

B. [Abdomen, right lateral sidewall], biopsy: Benign tissue...

Sometimes for this type of phrase in pathology notes, the term after the comma is a postmodifier – meaning there's a single entity being identified – but other times it actually does point to a separate entity. We want to make sure the entities are marked appropriately.

In *Colon, transverse*, "transverse" is a postmodifier identifying the part of the colon being discussed, so: [Colon, transverse]. In *Abdomen, right lateral sidewall*, "right lateral sidewall" isn't modifying “Abdomen” (it's not a "right lateral sidewall abdomen"); it's referring to a specific part of the abdomen. So here we’d want: [Abdomen], [right lateral sidewall].

While there are bound to be edge cases, generally speaking, if the word after the comma is an adjective, you can treat it as a postmodifier:

- [Colon, additional ascending]
- [Colon, sigmoid]
- [Perihepatic region, right]

If it's a noun, treat it as a separate entity:

- [Colon], [cecum]
- [Vagina], [left apex]
- [Sigmoid colon], [rectum]

For our original example, this means we want to divide the long single MARKABLE in (B) into two individual ones, with the idea being that they refer to two different parts of the body:

- [Abdomen], [right lateral sidewall]

7) Organs: Parts versus wholes

Consider a similar example:

(296) A. [Mesentery]$_{M}$, [small bowel], biopsy: Benign tissue…

B. [Mesentery]$_{M}$, [sigmoid colon], biopsy: Benign tissue…

In this particular context, we’re being told what the source is for the tissue removed by each of these individual biopsies. So the question is whether to read a term like [mesentery] here as referring to the entire organ (the same connected tissue), or specifically as the part of the organ from which that biopsy was taken.

We interpret it the first way, that is, we understand unmodified references like the mesentery mentions here as referring to the whole organ, and therefore they should be linked as IDENTICAL.

This is the only situation where we’re not assuming accuracy of the preexisting WHOLE-PART anatomical site links (Assume existing anatomical site W-P links are accurate). You may need to add or change associated IDENTICAL and WHOLE-PART links for this context. You’ll also likely not run into this in the clinical notes.

8) Path note phenomena that are not errors

“Received” phrases

The Gross Description section of pathology notes includes phrases like the following:

(297) Received fresh labeled colon is a 26.0 cm portion of colon.

You'll typically find that everything to the left of the copula is unmarked (except [Received], appropriately referring to the event of receiving the specimen). This is not an error. This part of the phrase ("fresh labeled colon") is understood to be a label and doesn't point to an actual entity. Therefore, these should remain unmarked.

AJCC staging codes

AJCC staging codes are codes that provide important information about the cancer's severity and location in the body. They are correctly marked as EVENTs and look like the following:

(298) Multiple (7 of 21) regional lymph nodes involved by adenocarcinoma. AJCC [pT3N1MX].

(299) With available surgical materials, AJCC [pT3N0].

2 The second annotation stage: Cross-document coreference

In this annotation stage, we'll make cross-document coreference links using the annotated data from the previous within-document passes. (The within-document stage is referred to below as both "within-doc" and "single-file.") As stated in the introduction, cross-document linking will take place for each individual patient's set of three notes, two clinical and one pathology.

Introduction and overview

The three notes will be concatenated into a single document, with divisions showing where one note ends and the next begins. In the following discussion, “Note A” refers to the topmost note, which is the chronologically first clinical note. “Note B” refers to the middle note, which is the second clinical note. “Note C” refers to the final note, which is the pathology note. Typically, though not always, the pathology note (Note C) is written before the second clinical note (Note B). Notes B and C are not in chronological order in the cross-doc file because Note B tends to have the most entities and EVENTs in common with both notes A and C. Both clinical notes tend to discuss the patient’s medical history, and the course, evaluation, planning, and treatment of the current disease; and Note B often gives a summary of the pathology report. Putting Note B in the middle is therefore most efficient for linking.

The EVENTs and entities (MARKABLEs) under discussion are shown in bold or colored print in the examples. Markables that are irrelevant (apart from providing context) aren’t marked.

You’ll start by linking Note A to both notes B and C, using a single link for all three notes as appropriate for a given relation. For example:

(300) **Note A**: CT-scan showed a **recurrence** in the ascending colon.

Note B: Biopsy confirmed recurrent **adenocarcinoma** on the right side.
Note C: Ascending colon, partial resection: Recurrent moderately-differentiated **adenocarcinoma**.

All three bolded EVENTs are referring to the same abnormal mass. These should be put into a single IDENTICAL chain, with the mention from Note A as the First Instance. (Of course, if the adenocarcinoma was only mentioned in any two of the notes, you’d still create an IDENT chain between the two.)

After you check Note A for coreference links with both notes B and C, you’ll then compare Note B to Note C. Of course, many of the links for notes B and C will have already been created – using the above example, because [adenocarcinoma] in both notes B and C have already been put in an IDENTICAL chain with each other and with the mention in Note A, it would be redundant to create yet another IDENT chain for just the mentions in B and C.

However, it’ll sometimes be the case that notes B and C both refer to an entity or EVENT that wasn’t present in Note A, and we need to capture these relations as well. For example:

(301) **Note B:** Surgery included **excision** of 18 perirectal lymph nodes.

Note C: Perirectal lymph nodes, **excision**: All benign.

There is no mention of the lymph node excision in Note A. Therefore, an IDENT relation should be created here, with Note B's mention as the First Instance.

Cross-document links are identified with a "C" in our online annotation tool to distinguish them from within-document links. That is, when you click on, say, the [excision] EVENT from Note B above, and the pop-up menu of links associated with that [excision] appears, the cross-document IDENTICAL chain will be identified with a "C" so that you can easily distinguish it from the within-document IDENTICAL chain for the same excision EVENT.

We will be using the same coreference relations we used for within-document linking: IDENTICAL, SET-SUBSET, and WHOLE-PART. (APPOSITIVE by definition can't be created cross-document.) We'll also be creating one type of TLINK, CONTAINS-SUBEVENT, because it conveys structural information as well as temporal. No other type of TLINK will be annotated in this pass. For ease of discussion, all four of these links may sometimes be referred to in these guidelines simply as "**coreference links**." SET-SUBSET, WHOLE-PART, and CONTAINS-SUBEVENT links are often collectively referred to as "**structural links**," as they all indicate some type of hierarchical relationship, unlike IDENT.

As with within-document annotation, we will not be creating any links for anatomical sites, body systems or properties (*musculoskeletal*, *height*, *temperature*, etc.), or normally-occurring body tissue, including terms that refer to tissue removed from the body, such as specimen, section, and margins. (By "normally-occurring tissue," we mean simply any tissue that's compositionally part of the body, versus abnormal tissue like tumor tissue. Abnormal tissue mentions should all be marked as EVENTs, however, making them easy to distinguish.) Medically-trained annotators will create these links in a later pass. Since body part and tissue references make up the majority of entities in these documents, and since WHOLE-PART links are used only for entities, this means that cross-document **WHOLE-PART links will be very rare**. IDENT and SET-SUBSET, by contrast, are used for EVENTs as well, and CONTAINS-SUBEVENT is of course used only for EVENTs.

You should still check for other compositional relations that fit the definition for WHOLE-PART, but these will be quite uncommon. Examples may include a department that's part of a hospital, or a doctor that's part of a department (remember people may be considered to be in a WHOLE-PART relation with organizations they work for).

Finally, the only WHOLE-PART relation we'll be using in this pass is compositional WHOLE-PART (i.e., "normal" WHOLE-PART). You may notice certain files also contain LOCATIVE and SUBEVENT WHOLE-PART links, but these are irrelevant to the current project and should be entirely ignored. To that end, when you open a new note, go to the menu option in the program's left panel and click on the CorefChains links to view the menu of coreference links. From there, de-select the LocativeWP and SubeventWP boxes.

Process

The following rules should guide the technical aspect of your cross-document linking decisions:

a) For all cross-document links, link the topmost mention in the first note to the topmost mention in the following note(s).

In other words, only the first mention of the patient in Note A needs to be linked to the first mention of the patient in Note B. The fact that both mentions also participate in their own IDENT chains within the individual notes means that we'll be able to infer the cross-document identical relationships of all of the other mentions simply by linking the top two together.

Using the topmost mention is for efficiency's sake – if you need to double-check a cross-document link you've made, you'll know which mention to select, rather than having to click on all the identical mentions in a within-document chain in order to find which one you've linked cross-document.

b) If there's a within-doc structural link between two markables, you do not need to create that same link cross-doc for the same two markables.

Again, if the mentions are linked via an IDENTICAL chain, the other relations they have can be inferred.

For example, say in Note A there's a mention of all the cancer screening procedures the patient's undergone, but no individual mentions of those procedures. In Note B, there's a mention of the same group of screening procedures, and also a specific mention of one of those tests, a colonoscopy:

(302) **Note A:** ...screening tests...

Note B: ...screening tests_{S}...$\text{colonoscopy}_{\text{SS}}$

Because in Note B the screening tests and the colonoscopy are linked via SET-SUBSET, and because the test mentions between Note A and Note B are linked via a cross-document IDENT chain (shown here by the same color), you do not need to create another link saying that the screening tests in Note A is in a SET-SUBSET relation with the colonoscopy in Note B.

The same is likewise true if the PART, SUBSET, or SUBEVENT has an IDENTICAL cross-document relation, i.e., if there's a mention of the colonoscopy in both notes but a mention of the group of screening tests in only one note. Provided there's a within-document SET-SUBSET link between the screening procedures and the colonoscopy in that note, the same relation can be inferred.

Put differently: Do not link markable A in one note to markable B in another note with a structural relation if markable B is already in a within-doc structural relation of the same type with a cross-document IDENT mention of markable A.

c) But, do link the SUBSET, PART, or SUBEVENT mentions to each other using IDENT.

To continue with the same example, say in Note A there's a mention of all the cancer screening procedures a patient has undergone, and also a mention of one of those procedures, a colonoscopy. Now say you also have both mentions in Note B. You should link the mentions of the group of procedures to each other via IDENT, and you should also link the mentions of the specific colonoscopy to each other via IDENT.

(303) **Note A:** ...screening tests$_{S}$...colonoscopy$_{SS}$ on June 14th, 2010

Note B: ...screening tests$_{S}$...her June 14th colonoscopy$_{SS}$

The reason for this is because the patient could have had multiple colonoscopies. So even if the above two mentions are both members of the same group of tests, an IDENTICAL link is necessary to show they're referring to the same colonoscopy.

In other words, ***always create IDENTICAL links whenever appropriate***, regardless of the other links in which the mentions may also participate.

d) Create cross-document structural links when both components of the relation do not have a cross-document IDENTICAL link.

Based on the above, most of the cross-document links you make will be IDENTICAL links. However, you should also be on the lookout for cross-document WHOLE-PART, SET-SUBSET, or CONTAINS-SUBEVENT relations, which will be necessary if neither part of the relation participates in a cross-document IDENTICAL chain.

In other words, if Note A talks about the group of screening procedures but not the colonoscopy from June 14th, 2010, and Note B talks about that colonoscopy but not the group of screening procedures, you should create a cross-document SET-SUBSET link between the two, since that relation has not yet been created at all:

(304) **Note A:** ...screening tests$_{S}$...

Note B: ...colonoscopy$_{SS}$...

e) Link mentions of the same event to each other even if the DocTimeRel is different.

This might be obvious, but [proctocolectomy] in Note A with a DocTimeRel of AFTER, ACTUAL should be linked via IDENT to [proctocolectomy] in Note B with a DocTimeRel of BEFORE, ACTUAL, if they in fact refer to the same procedure.

f) Keep cross-document links separate from within-document links.

In other words, create new links for cross-document relations even if you could add a mention to a within-document link.

For example, say in Note A you have a mention of the group of screening procedures and it's in a SET-SUBSET relation with a colonoscopy, while in Note B you have a mention of a different screening test the patient had, a mammogram:

(305) **Note A:** ...screening tests$_{s,s}$...colonoscopy$_{ss}$

Note B: ...mammogram$_{ss}$...

As the example shows, you should create a new, different SET-SUBSET relation between the screening tests and the mammogram instead of adding the mammogram to the within-document SET-SUBSET link already created for Note A.

g) Don't change any within-document links.

Only cross-document links may be made in this pass. Within-document links may not be deleted, changed, or added.

However, due to idiosyncrasies of cross-document annotation (discussed more below), you may sometimes run into situations where you feel you cannot create a reasonable cross-doc link because of how it would conflict conceptually with a given within-doc link. These cases should be rare, but if you encounter one, you should communicate it to the adjudicator. (Making within-document changes during annotation will cause major technical problems with the online tool later on in the pipeline, but these changes can be made without issue during adjudication, so we permit them during adjudication only, and only when absolutely necessary.)

Cross-document specific phenomena

In cross-document annotation, the definition of the relations themselves are the same as they are for within-document annotation. Everything you know about what an IDENTICAL link means or represents within-document is the same cross-document, and the same is true for the other relations.

However, several other peculiarities of cross-document annotation make it a somewhat different beast from within-doc annotation. For one thing, the notes for a given set are typically written weeks or months apart, often by different authors (because of the de-identification process, we don't know when the authors are the same or different). This means annotation choices necessarily rely less on certain linguistic cues – such as narrative discourse – and more on modifying details we can corroborate in the other note(s). Things like dates and locations become very important. For example, we can know whether two cross-doc biopsy mentions are referring to the same biopsy event in part if the dates are the same for both; and whether one tumor is the same as another if the described location in the body is the same. Of course, these are all clues we look for in single-file annotation as well, but the necessity of this kind of "sleuthing" is accentuated in cross-doc work since the number of other clues we have available are reduced.

Furthermore, it's naturally the case that when you have three documents that all discuss roughly the same set of events, you've got more information about those events than when you only had one document. The information we have about a given event's structure suddenly becomes much more granular and nuanced. While not surprising, this phenomenon does have interesting consequences for annotation – within-document links that were entirely reasonable based on the information present in a single file may not make quite as much sense when viewed in light of the other two files.

We've already incorporated certain tools and guidelines in our within-document passes to help accommodate this – the CONTAIN-SUBEVENT relation was added for just this reason, as well as several specific annotation rules. However, it's simply not possible to anticipate all the scenarios where cross-doc information renders a within-doc linking choice questionable or inaccurate. This means a couple things for cross-document annotation:

- **Paying attention to within-document linking choices is crucial.**

For example:

(306) **Note A**: For neoadjuvant **therapy**, Mr. Smith will receive **Avastin**.

a. [therapy] IDENT [Avastin]

This is a reasonable within-doc link – the event of administering the neoadjuvant therapy is the event of administering the Avastin.

However, say we learn this in the later clinical note, written months later:

(307) **Note B**: Patient underwent surgery following neoadjuvant **therapy** in the form of **Avastin** and concurrent **radiation**.

a. [therapy] CON-SUB [Avastin], [radiation]

Apparently since making the initial plan in Note A, the patient's care team decided that radiation was needed as well, so both the chemotherapy (Avastin) and the radiation are subevents of the total adjuvant therapy the patient received.

What does this mean for cross-document linking? Importantly:

- ~~neoadjuvant [therapy] IDENT neoadjuvant [therapy]~~

The two mentions of neoadjuvant [therapy] should not be IDENT. Because in Note A [therapy] is IDENT to [Avastin], and in Note B [therapy] is CON-SUB with [radiation], a cross-doc IDENT link here would entail that the radiation is a subevent of Avastin (and that Avastin is a subevent of itself)!

Instead, the only cross-doc link to be made here is:

- [Avastin] IDENT [Avastin]

Taken with the within-doc relations, this link entails that neoadjuvant therapy from Note A is a subevent of neoadjuvant therapy in Note B. This is fine. Terms like *therapy* and *treatment* are used variably all the time in these notes, sometimes to refer to an overall treatment course and sometimes to refer to specific types of treatment, and therefore it's not problematic to nest them with each other. The thing to pay attention to is how they've been linked within-document to specific types of treatment, so that you can ensure your cross-document links don't wind up telling lies about the relation of the specific treatment types.

The main takeaway here is that within-doc linking choices should guide your cross-doc analyses. That said, if you run into a situation where you're sure you can't create an accurate cross-doc link without it being logically at odds with the within-document links, tell the project's adjudicator. We are allowing the possibility of changing some within-doc links in adjudication. This should be quite rare, however.

- **Paying attention to within-document context is crucial**

This might sound obvious, but a general guiding principle we have in both single-file and cross-document annotation is to "trust" the author, that is, to make annotation decisions based on the way the author presents information in the text. Always apply all specific instructions from the guidelines; then, take into account within-doc linking choices; finally, if those two guideposts don't apply to the specific situation at hand, you should create links that make sense to you as a thoughtful reader, taking into account all available clues in the text.

Here's an example of how this might play out:

(308) **Note A** (DOCTIME May 2, 2012): Patient presents with a variety of **symptoms**.

Note B (DOCTIME July 19, 2012): Ms. Dubois sought medical attention in May for rectal **bleeding**, **nausea**, and **bloating**. She has since developed severe RLQ **pain**.

You notice that [bleeding], [nausea], [bloating], and [pain] are all medical problems. However, you also notice that [pain] can't be among the group of symptoms referred to in Note A, because the author explicitly says it developed later, during or after the colonoscopy that itself occurred after Note A was written. Therefore, the appropriate cross-doc link is:

a. [symptoms] S-SS [bleeding], [nausea], [bloating]

In other words, while we've lost some linguistic cues, we're not annotating in a vacuum; we still want to interpret EVENTs in the context in which they're discussed within-doc, not in an abstract, generalized way.

Of course, this will take ongoing fine-tuning, and there are going to be many borderline calls. When you've taken into account all available information, and you're still on the fence about a potential link, default to linking rather than not linking – it's possible to go back and get rid of erroneous relations, but we can't recover information that hasn't been linked at all.

That said, don't resort to guessing! Even borderline cases must have some kind of support from the text, as in (308).

Finally, use the Needs_Medical_Opinion feature as necessary to flag relations you're not sure of due to lack of clinical knowledge.

In addition to these general principles, following are several specific guidelines to keep in mind as you work on cross-document annotation.

Modality

We maintain the same modality-linking rules in the cross-document stage as the within-document stage, summarized again here:

- For all TLINKs except CON-SUB:
    - GENERICs may be linked to GENERICs and HYPOTHETICALs to HYPOTHETICALs. Neither may be linked to ACTUAL or HEDGED EVENTs.
    - ACTUAL and HEDGED EVENTs may be linked to each other when appropriate.
- For all coreference links, including CON-SUB:
    - Don't link HYPOTHETICAL or GENERIC markables at all, except for GENERIC-ACTUAL SET-SUBSET links.

What we know about events – whether they will or did or might happen – changes over time, and we are now dealing with documents written at different times. So it may be desirable down the road to link HYPOTHETICAL to ACTUAL EVENTs, but this is beyond the scope of the current project.

For SET-SUBSET links between GENERICs and ACTUALs, note that cross-doc GENERIC linking will be pretty rare, as it'll only occur when the SUBSET doesn't have an IDENTICAL link. However, if there are multiple GENERIC mentions that require a cross-doc relation, *each* of them needs to be linked since we don't link GENERICs to each other. For example:

(309) **Note A**: **Patients**$_1$ with **cancer**$_1$ are advised to avoid this medication…

Patients$_2$ with **cancer**$_2$ are at higher risk for severe side effects.

Note B: **Mr. Cruse** has kidney **cancer**.

Assuming neither Mr. Cruse nor his kidney cancer are mentioned in Note A (and therefore there's no feasible IDENT link), the cross-doc relations should be as follows:

- [Patients]$_1$ S-SS [Mr. Cruse]
- [Patients]$_2$ S-SS [Mr. Cruse]
- [cancer]$_1$ S-SS kidney [cancer]
- [cancer]$_2$ S-SS kidney [cancer]

Finally, recall that entities don't have features. When in doubt about whether an entity should be understood as actual, hypothetical or generic, you should first refer to its within-document links to see how it was interpreted during single-file annotation. If that doesn't shed any light, make the call based on what you know about each modality and the way the entity is discussed in the text. (All concepts in this discussion pertain to entities as well as EVENTs.)

Polarity

Because of this cross-document phenomenon that different notes often present information differently, once in a great while EVENTs that point to the same real-life event may have different polarities:

(310) **Note A**: April 16, 2014: We attempted **resection**$_{NEG}$, but patient's blood pressure spiked and procedure was halted.

Note B: Prior to the April 16, 2014 **resection**$_{POS}$, patient had exhibited no symptoms.

Here, even though the resections have different polarities, other contextual clues (such as the date) tell us that the event under discussion is the same, so these may be linked as IDENT. **This should be extremely rare.** Don't go looking for POS-NEG links.

EVIDENTIALs

We will not be linking EVIDENTIAL EVENTs to each other cross-document (nor within-document).

(311) **Note A:** CT-scan on June 4, 2015 **demonstrated** large rectal mass.

Note B: June 4, 2015: CT **showed** large mass in the rectum as well as an abdominal hernia and indeterminate hepatic lesions.

No relation should be created between [demonstrated] from Note A and [showed] from Note B. (Of course, the two CT scan mentions here should be linked via IDENT since they refer to the same test, as well as any of the findings that point to the same EVENT, such as rectal [mass].)

This is because it's very difficult to get agreement on when two showing EVENTs are the same, and because the more important piece is that all the tests and findings be linked to each other.

Pathology notes

Recall that while clinical notes discuss a patient's current state, medical history, plan of care, and so forth, pathology notes present the findings from studying tissue that was removed during some kind of procedure.

Pathology notes frequently mention the subprocedures, but not the overall procedure. Similarly, they tend to mention physical manifestations of a cancer (adenocarcinomas, masses, etc.), but not the cancer itself. In other words, for both cancer EVENTs and cancer treatment EVENTs, they tend to discuss the SUBEVENTs, but not the larger EVENT (the narrative container). Refer to CONTAINS-SUBEVENT in the within-document annotation guidelines.

This means you should especially be on the look-out for cross-doc CONTAINS-SUBEVENT links between either one of the clinical notes and the path note. However, note that this cross-doc link still won't be frequent because many of the SUBEVENTs will have cross-doc IDENT links (see under Process above: *If there's a within-doc structural link between two markables, you do not need to create that same link cross- doc for the same two markables*).

Compare the following two examples:

(312) **Note A**: PLAN: Low-anterior **resection** and gallbladder **removal**. Patient is scheduled for **surgery** on October 15th, 2015.

- Note A within-document links:

a. [surgery] CONTAINS-SUBEVENT [resection]
b. [surgery] CONTAINS-SUBEVENT [removal]

Note C: Gallbladder, **excision**: Multiple cysts identified.

- Cross-document links:
  a. [removal] IDENT [excision]

Because [excision] is in a cross-doc IDENT link with [removal], which itself is in a within-doc CONTAINS-SUBEVENT link with [surgery], we do not need to create a cross-doc CON-SUB link between [surgery] and [excision].

(313) **Note A**: Patient presents with colon **cancer**.

Note C: Omentum, biopsy: Negative for **tumor**$_{\text{NEG}}$.

- Cross-document links:

  a. [cancer] CONTAINS-SUBEVENT [tumor]$_{\text{NEG}}$

In this example, there's no cross-doc IDENT link for the negated tumor. Therefore, the hierarchical relation between the cancer and the negated tumor has not yet been captured, and you should create a cross-doc CONTAINS-SUBEVENT link.

Medications and allergies sections

Our task for these sections are the same as for other sections, which is to link coreferential mentions cross-document. However, medications and allergies sections have some peculiarities that warrant brief discussion here.

Allergies sections

Recall that most EVENTs in allergies sections are GENERIC, as they're not really referential to any particular events of the patient actually encountering the allergen or experiencing the reaction – they represent more of a *when X, then Y* type of statement. Because we don't link GENERICs to each other, cross-doc links to anything in the allergies section will likely be very rare.

Per our guidelines above, the main cross-doc link you'll create for allergies sections is a S-SS relation between the GENERIC reference to an allergen or reaction and an ACTUAL reference to the patient actually encountering the allergen or experiencing the reaction. While unusual, this does occur in these notes. So if Note A refers to the patient reacting to taking plavix and Note B lists [Plavix] in the allergies section, there would be a S-SS relation between the two medication references.

Sometimes a patient has no allergies:

(314) Medication :

NO KNOWN MEDICATION **ALLERGIES**
Radiology :
NO KNOWN CONTRAST **MEDIA**
Allergies above current as of Friday, September 30, 2015

Here the events are HEDGED, NEG, as the patient apparently doesn't have any medication or contrast allergies (contrast is a chemical substance used in certain scans). These negated allergies would only be in cross-document links if the other note lists the same time for the observation (September 30, 2015, in this case). If the time is different, no link should be made. This is discussed more under Potentially continuous EVENTs below.

Apart from unusual cases where doctors veer from the typical template-filling format for this section, these are the only two cross-doc links you'll need to make for allergies sections.

Medications sections

Consider the following example:

(315) **Note A**: **Plavix** 20-mg capsule 1 **capsule** by mouth every 3 days.

Note B: **Plavix** 20-mg capsule 1 **capsule** by mouth every 3 days.

Recall that [Plavix] is considered to be the event of the patient having a prescription for that *type* of medication. [capsule] is considered to be the event of the patient having a prescription for that particular *dose* of medication. They are in a SET-SUBSET relation, where [Plavix] is the SET and [capsule] is the SUBSET.

Most cross-doc medication section relations will be straightforward, following our cross-doc linking practices described above. In this case:

a. [Plavix] IDENT [Plavix]

b. [capsule] IDENT [capsule]

No need for a cross-doc S-SS link because these are already present within-doc.

Sometimes they continue the medication but change the dose:

(316) **Note A**: **Plavix** 20-mg capsule 1 **capsule** as **directed** by **prescriber**.

Note B: **Plavix** 20-mg capsule 2 **capsules** as **directed** by **prescriber**.

Cross-doc links:

a. [Plavix] IDENT [Plavix]
b. ~~[capsule] IDENT [capsules]~~
c. ~~[directed] IDENT [directed]~~

d. ~~[prescriber] IDENT [prescriber]~~

Note that the prescription type EVENTs (Plavix) are IDENT because the patient is still on that type of medication regardless of the dosing change. But the prescription dose EVENTs here shouldn't be linked, since they're different. Similarly, [directed] and [prescriber] references may be linked as IDENT cross-doc if the prescription stays the same, but shouldn't be linked if the prescription changes. This is because the change in prescription may have been done by a different physician; we don't know.

Sometimes the patient is not on any medications:

(317) January 23, 2012: Patient stated he is on no current **medications**.

As with allergies, if the observation of no medications is made again in a later note at a different time, the negated medications should be left unlinked. If the observation of no medications is made again in a later note, but the time listed is the same (January 23, 2012, in this case), the negated medications may be linked as IDENT.

Potentially continuous EVENTs

It's usually relatively easy to make linking choices about EVENTs that are more or less punctual. We know that a colonoscopy that a patient had back in the 1980s isn't the same as the colonoscopy she had last month. EVENTs that may or may not be durative are much trickier, however.

Consider the following example:

(318) **Note A** (March 25, 2012): Heart rate was **regular**.

Note B (April 1, 2012): Heart: **Regular** rate.

Here we have two instances of heart rate regularity that are observed at different times. The question is, are they identical? Or, more generally: When the same type of potentially continuous EVENT is observed at two different timepoints, should those mentions be considered to be the same, ongoing EVENT, or two different EVENTs that happen to be of the same type? If the former is true, we would want to create an IDENTICAL link between the two; if the latter is true, there should be no link at all.

Generally speaking, our approach is that when stative or attributive EVENTs – things that vary in value – are measured or identified at two different times, they should not be linked, unless there's explicit linguistic evidence they're the same event. This means that for the example above, we wouldn't link [regular] to [Regular]. It's possible the heart rate was at some point irregular between Note A and Note B, so we can't say for sure that this is the same, ongoing event of regularity.

A few other examples:

(319) **Note A**: November 11 CT-scan shows the liver mass is **small**.

Note B: Hepatic mass is **small** on March 4 scan.
~~a. [small] IDENT [small]~~

[small] and [small] are not IDENT. Again, it's possible that between these two observations the nodules grew and then shrunk again.

(320) **Note A**: Abdomen: **Obese**.

Note B: Abdomen: **Obese**.

a. ~~[Obese] IDENT [Obese]~~

(321) **Note A**: Mr. Long is a **pleasant** 58-year-**old** man with multiple cancers.

Note B: Patient is a **pleasant** 58-year-**old** male here for treatment recommendations.

a. ~~[pleasant] IDENT [pleasant]~~

b. [old] IDENT [old] (*do link [old] if the numerical age is the same*)

We include **symptoms** under this umbrella as well. If we substitute both mentions of [regular] with [irregular] in the first example, we would still not link them, for the same reason – we can't know that the events of being irregular are the same. Similarly, if the patient presents with [nausea] in Note A and is also [nauseous] in Note B, these should also not be IDENT – it's quite possible the patient stopped being nauseous in between the two visits, and therefore these would be two different nausea events.

The same policy applies to negated EVENTs:

(322) **Note A**: Scan shows no **metastases**$_{\text{NEG}}$.

Note B: No **metastases**$_{\text{NEG}}$ observed on testing.
a. ~~[metastases]$_{\text{NEG}}$ IDENT [metastases]$_{\text{NEG}}$~~

Not having metastases is a state. Because these EVENTs are being observed at two different timepoints, they should not be linked.

(Note that this is why observations of the patient not having allergies or not being on medications shouldn't be linked, unless the time of observation is the same in both notes. See Medications and allergies sections above.)

However, we should link stative/attributive/symptomatic EVENTs as IDENT if the author provides explicit linguistic evidence that they in fact refer to the same event:

(323) **Note A**: Patient is **nauseous**.

Note B: Patient's **nausea** has continued since his previous visit.

a. [nauseous] IDENT [nausea]

The same is true for negated EVENTs – saying something doesn't exist at Time A and it doesn't exist at Time B (which is what we have in example 322 above) is different from saying the same thing has never existed:

(324) **Note A:** Ms. Day has never had a **mammogram**$_{NEG}$. We have ordered one for next week, February 11.

Note B (DOCTIME 2/11/14): Patient had had no **mammogram**$_{NEG}$ before today.
a. [mammogram]$_{NEG}$ IDENT [mammogram]$_{NEG}$

Both [mammogram] EVENTs highlighted here refer to the mammogram event that never happened prior to February 11. These are IDENTICAL. As you can see from both this example and the [nausea]$_{POS}$ example (example 323), the "explicit linguistic evidence" that identifies two potentially continuous EVENTs as the same will frequently be the use of the present perfect tense ("nausea has continued," "has never had a mammogram"). Often at least one of the EVENTs will have a DocTimeRel of BEFORE/OVERLAP, though this won't always be the case. You may and should consider other contextual clues as well (like the date in the mammogram example) when making your decision.

By contrast, EVENTs that are characterized by or cause these variable states, attributes and symptoms should get linked as IDENT. This includes diseases/diagnoses and abnormal implicit EVENTs. (Recall that by "abnormal implicit events" we mean terms like [mass], [tumor], [adenocarcinoma], etc. – abnormalities that are entity-like in that they have physical substance, but that we interpret as the event of the patient's having those abnormalities.) We certainly want a mention of the patient's colon cancer or arthritis or heart disease linked to another mention of the same disease in a different note, and mentions of the same abnormal mass linked to each other. So:

(325) **Note A**: **Diabetes**, **stable**.

Note B: **Diabetes**, **stable**.

a. [Diabetes] IDENT [Diabetes]
b. ~~[stable] IDENT [stable]~~

The disease mentions here should be linked, but the stability of the disease – i.e., its state – should not.

Finally, on a related note, just as EVENTs observed at different timepoints are not IDENTICAL (unless stated otherwise), neither are EVENTs observed relative to different locations. This is obvious for positive EVENTs – having a [mass] in the colon is not the same as having a [mass] in the liver – but a brief (within-doc) example is warranted here for negated EVENTs:

(326) A. Liver, segment II, biopsy: Negative for **tumor**$_{NEG}$.

B. Liver, segment III, biopsy: Negative for **tumor**$_{NEG}$.
a. ~~[tumor]$_{NEG}$ IDENT [tumor]$_{NEG}$~~

These two negated tumor EVENTs are not IDENT, as they are noted as not existing relative to two different locations, one in segment II of the liver, the other in segment III.

Authoring doctors

It's tempting to say the first-person references to the authoring doctor are IDENT cross-document:

(327) **Note A: I** discussed with the patient that adjuvant chemotherapy is a viable option in his case.

Note B: I saw patient in followup prior to his visit with Cardiology.

But because patients often see different doctors, we can't make this assumption and therefore **authoring doctors should never be in a cross-document IDENTICAL link**:

a. ~~[I] IDENT [I]~~

Also be careful about linking first-person plural entities (we, us, our). You should only link these if it's clear based on within-document links and context that these groups consist of the exact same members. (Watch out! In many cases not all the SUBSETs of a SET are explicit in the text, so just because all the SUBSETs listed are the same doesn't mean two SETs are the same. You might have a [we] mention with only one SUBSET, [I]. This doesn't mean the entire SET consists of one person; it means the other members aren't explicitly stated in the text.)

IV Appendices

Appendix A: Clinical Note Section Labels

- 20100 - Revision History

- 20101 - Referral Source
- 20102 - Chief Complaint/Reason for Visit
- 20103 - History of Present Illness
- 20104 - Current Medications
- 20105 - Allergies
- 20106 - System Review
- 20107 - Past Medical/Surgical History
- 20108 - Social History
- 20109 - Family History
- 20110 - Vital Signs
- 20111 - Physical Examination
- 20112 - Impression and Report and Plan
- 20113 - Diagnosis
- 20114 - Administrative
- 20115 - Special Instructions
- 20116 - Advance Directives
- 20117 - Service Actors
- 20118 - Immunizations
- 20119 - Admission Findings and Test Results
- 20120 - Problem Oriented Hosp. Course
- 20121 - Final Physical Examination
- 20122 - Adverse Reactions
- 20123 - Diet / Nutrition
- 20124 - Discharge Condition
- 20125 - Condition at Discharge
- 20126 - Ongoing Care Orders
- 20127 - Admission Physical Exam
- 20128 - Ongoing Care
- 20129 - Follow Up Agreements
- 20130 - PHF and CVI Dates

- 20133 - Admission Medications
- 20135 - Anticipated Problems and Interventions
- 20136 - Post-op Services
- 20137 - EMTALA Statement
- 20138 - Patient Education
- 20147 - Dismissal Medications
- 20148 - Patient Consent

Appendix B: First pass checklist

Following is a quick-reference list of the major annotation rules and the specific issues you should check for if you are doing first-pass (data synching stage) annotations. This should not be used as a substitute for the complete guidelines, but as a supplemental checklist once you are thoroughly acquainted with the issues as fully discussed in the guidelines.

Major annotation rules

- EVENTs and entities (MARKABLEs) should never be in the same link (either coreference links or TLINKs).
    - Exception: LOCATIVE W-P
- Ignore SUBEVENT W-P links.
- Temporal links may only be used for EVENTs and TIMEX3s (not entities).
- Coreference links are used between EVENTs and other EVENTs, or between entities and other entities. However, WHOLE-PART is only for entities.
- CONTAINS-SUBEVENT links may only be used for the four categories of events listed below. We are not annotating all subevent relations.
- HYPOTHETICALs and GENERICs should not be in coreference links or CONTAINS-SUBEVENT links.
    - Exception: GENERIC-ACTUAL SET-SUBSET links
- Don't double-tag (i.e., create two different labels for the same term), except in the following specific cases:
    - PREPOSTEXPs (double-tagged as EVENTs and TIMEX3s)
    - Physical properties in the Vital Signs section, section ID 20110 (double-tagged as EVENTs and entities)
    - Physical properties in other sections when they are the only way we can capture a testing event (double-tagged as EVENTs and entities)

Checklist of tasks

1) Mark singleton entities

2) Add CONTAINS-SUBEVENT links for:

2a) Cancer events

- **Cancer terms** ("containing" events; not comprehensive):
    - *cancer*
    - *disease*

- **Physical manifestation terms** (subevents; not comprehensive):
    - *carcinoma*
    - *adenocarcinoma*
    - *metastasis*
    - *recurrence*
    - *tumor*
    - *lesion*
    - *nodule*

        - *mass*

Don't nest cancer subevent links, except for between tumors and parts of tumors: *[part] of [tumor]*: [tumor] CON-SUB [part].

2b) Cancer treatment events

- **Surgical procedures**
    - **Things that count as "general" procedural terms** (not comprehensive):
        - *surgery*
        - surgical *intervention*
        - surgical *management*
        - *operation*
        - surgical *procedure*
        - *preoperative*
        - *postoperative*
        - *intraoperative*
    - **Things that count as "specific" procedural terms** (not comprehensive):
        - *resection*, *resected*
        - *hemicolectomy*
        - *colectomy*
        - *ileostomy*
        - *colostomy*
        - *excision*
        - *removal*
        - *APR* (= abdominoperineal *resection*)
- **Chemotherapy and radiation**
    - Always use CON-SUB (not S-SS) to link "larger" chemo/rad events to "smaller" ones.

2c) Chronic disease events (episodes)

2d) Medications events

3) Check for the following specific issues:

- Convert eventive MARKABLEs to EVENTs (and vice versa)
- Double-tagged events
- Properties and their states
- Vital Signs sections
- Delete markables and links from non-annotated sections:
    - 20116 Advance Directives

  - 20138 Patient Education
  - 20148 Patient Consent
  - 20123 Patient Diet and Nutrition
- Events of location
- Conjoined NP and list NP MARKABLEs
- Test results
- Spans
- Annotate medications sections
- Annotate allergies sections
- Delete HYPOTHETICAL and GENERIC coreference links
- Delete “extra” PARTs and SUBSETs
- Annotate equational clauses
- Annotate hard-to-spot SETs (*positive/negative*; *chief complaint*)
- Extract compositional WHOLE-PART links from locative ones and vice versa
- Check APPOSITIVEs for accuracy
- Make sure authoring doctor isn’t linked to referring doctor
- “X negative for X”
- Delete temporal expression MARKABLEs and delete TIMEX3s from coreference links
- Make sure PREPOSTEXPs are double-tagged as EVENTs and linked appropriately
- Check DOCTIME for accuracy
- Check for unusual SECTIONTIMEs
- Delete non-markable “temporal expressions”
- Delete relative positions
- Check for accurate S-SS relations for lymph nodes
- Make sure all tissue references (like *specimen*, *section*, *margin*) are entities
- Delete formula paragraph links
- Make sure tumors in different locations are not IDENT
- Make sure body parts are marked accurately based on our treatment of organ terms and postmodifiers versus separate noun phrase

Appendix C: The rest of the pipeline: Coref-from-scratch, the anatomy passes, and post-processing

The preceding guidelines were written based on how roughly two-thirds of the files for this project should be annotated. This appendix discusses the annotation for the remaining one-third, which were never annotated by T1Coref. For these files, we have the original T1Temp annotations (with EVENTs, TIMEX3s, SECTIONTIMEs, DOCTIMEs, and TLINKs), but no coreference links and no entities. Therefore, they are being sent through a somewhat different pipeline, described below. This is referred to as the "Coref-from-scratch" pipeline, and the pipeline/guidelines for the other two-thirds – i.e., everything in the main body of this document – is referred to as "Merged," reflecting the fact that those notes consist of annotations from T1Temp and T1Coref that have been merged.

Secondly, there are two separate anatomy passes not discussed in the main body of the guidelines. One is a single-file anatomy pass that is the final stage of the Coref-from-scratch within-document pipeline. The other is a cross-document anatomy pass. Both are also described below.

I The Coref-from-scratch pipeline

1 The first pass: T1Temp corrections

This is a single-annotated, corrections-style pass with the purpose of bringing all **temporal annotation** into alignment with the guidelines above. For example, because terms like specimen, section, and margin were inconsistently annotated in T1Temp (sometimes as EVENTs, sometimes left unmarked), this pass allows us to make these terms consistent with the way we are treating them in the other two-thirds of the notes (i.e., as entities; therefore they're deleted here if found as EVENTs). Because T1Temp didn't annotate medications or allergies sections, annotators are also adding EVENTs for these sections in this pass. No entities or coreference links are created in this pass, and it is not adjudicated.

Specifically, annotators in this pass:

- Check every EVENT and TIMEX3 briefly for consistency with all the EVENT and TIMEX3 guidelines discussed above. They may fix any kind of error (features like modality, polarity; EVENT or TIMEX3 existence; etc.), but in all cases the error must be egregious to warrant changing it.
- Do not check every TLINK, but may fix egregious errors if found. CONTAINS-SUBEVENT TLINKs are added in the next pass.

- Annotate medications and allergies sections for EVENTs only, based on the guidelines for these sections in Reconciling the merged data above. Entities and coreference links for these sections are added in the next pass.

2 The second pass: Adding entities, coreference links, and CONTAINS-SUBEVENT links

This pass is double-annotated and adjudicated. Annotators add all entities (i.e. MARKABLEs), coreference links (IDENTICAL, APPOSITIVE, SET-SUBSET, WHOLE-PART), and CONTAINS-SUBEVENT TLINKs in accordance with the guidelines above.

In this pass annotators are not permitted to:

- Delete or change any markables or links that are already present.
- Add anything other than the categories listed above (i.e., they may not add EVENTs or TIMEX3s, etc.)
- Create any links involving anatomical sites. A medically-trained annotator is doing this post-adjudication. (They are still creating the entities for these terms, however.)
- They can completely ignore the following sections from the guidelines:
    - "Extra" PARTs and SUBSETs
    - WHOLE-PART links
    - TIMEX3, DOCTIME, and SECTIONTIME checks
- For all other sections of the guidelines, language that discusses changing EVENTs, TIMEX3s, SECTIONTIME, DOCTIME, or TLINKs should be ignored. Again, annotators are only adding things in this pass, as listed above, not deleting or changing anything.

3 Adjudication

Adjudication takes place for the annotations created in the second pass. EVENT and TIMEX3 corrections are permitted in adjudication as well, even though the first pass on temporal annotation isn't technically being adjudicated.

4 Anatomy pass

In this pass, a medically-trained annotator is adding all coreference links for body part and tissue entities, as this is much more efficient and accurate than three non-medically trained people (two annotators and one adjudicator) attempting to research and create these links, which of course rely heavily on knowledge of anatomy.

The anatomy pass guidelines are linked under Appendix E.

II The cross-document anatomy pass

This pass takes place at the very end of the pipeline, after cross-document adjudication occurs. It is single-annotated by a medically-trained annotator.

In this pass, the annotator will create all cross-document links involving anatomical sites and tissue references. The annotator will also be permitted to change within-document links as necessary. The reason for this is that the anatomy links in the Merged notes aren't consistent with the anatomy links in the Coref-from-scratch notes, not because either one is necessarily wrong, but because there's variability and disagreement in the medical community as to how the terms relate to each other. Permitting within-document changes is necessary in this stage to make these links consistent with each other. This pass will not be adjudicated, but will be immediately set to gold.

The cross-doc anatomy pass guidelines are linked under Appendix E.

III Post-processing

We implemented the following quality-control steps as the final annotation step in producing cross-document gold:

Automatic post-processing tasks

(1) Script identifies when two EVENTs are related by both a TLINK and IDENT link. If found, the TLINK is deleted.

- Only a within-doc issue.

(2) Script identifies when two EVENTs or TIMEX3s have been linked by more than one TLINK type.

- The script looks at IDENTICAL mentions of the two events (via both within-doc and cross-doc IDENTICAL chains).
- If the script finds two events linked as both CONTAINS and CONTAINS-SUBEVENT, it automatically deletes the CONTAINS link.
- If the script finds any other combination of TLINK types linking the same two events, it identifies the relations and entities involved so that an annotator can then compare and resolve the conflict.

(3) Script identifies empty/negative-span markables as well as empty and half-empty links. Deletes empty markables and links. Identifies half-empty links for an annotator to fix.

Manual post-processing tasks

After the automatic post-processing tasks, a single annotator:

(1) Resolves conflicting links (see #2 under Automatic post-processing tasks above).

(2) Converts CONTAINS to NOTED-ON for *test-result* TLINKs.

> *Note: In the following examples, EVENTs are not marked exhaustively; only the ones being discussed are marked.*

The T1Temp project specified that test EVENTs should be linked to their associated result EVENTs using CONTAINS:

(a) [Biopsy] demonstrated [cancer].

[Biopsy] CONTAINS [cancer]

In this post-processing step, we converted these links to a new TLINK type, NOTED-ON, which is a subtype of OVERLAP:

[cancer] NOTED-ON [Biopsy]

This change was motivated in part by the addition of coreference relations, which exposed the temporally-conflicting information produced by the test-result CONTAINS links. A *cancer* reference like the one above is frequently IDENTICAL to other mentions throughout a given note, which themselves often have a DocTimeRel or other attached temporal information that shows the cancer exists outside the temporal bounds of the test. When combined with the original *test-CONTAINS-results* relations, this meant our annotations said: *X-result is temporally contained by Y-test* and *X-result temporally exists before/after Y-test*, a logical impossibility.

NOTED-ON, being a subtype of OVERLAP, eliminates this conflict and additionally represents an attribution/reporting relation.[7] Informally, this TLINK allows us to say "*X* temporally overlaps *Y*, and we saw event *X* on/by event *Y*." Specific guidelines our annotators used for this post-processing conversion follow:

- While the vast majority of T1 test-result relations were CONTAINS, per the T1Temp guideline, in practice, OVERLAP was sometimes used. These should also be changed to NOTED-ON:

  (b) Patient here for further discussion of the [mass] seen on CT-[scan].

  ☐ Delete: [mass] OVERLAP [scan]
  ☐ Create: [mass] NOTED-ON [scan]

- Continue to maintain the standard of not linking HYPOTHETICAL or GENERIC EVENTs to ACTUAL or HEDGED ones:

  (c) [MRI]$_{\text{ACT}}$ demonstrated possible [adenocarcinoma]$_{\text{HYPO}}$ but further testing needed.

  ☐ No link in T1 or T2.

[7] It's therefore similar to CONTAINS-SUBEVENT in that both links represent temporal information and additional semantic information.

- For the purposes of this conversion, links involving EVIDENTIAL EVENTs (*shows*, *demonstrates*, etc.) should be ignored. In T1, EVIDENTIALs were usually linked to tests via CONTAINS, and this is fine from a temporal standpoint. Because we do not annotate coreference links for EVIDENTIALs (see EVIDENTIALS), no conflicts with temporal information will be generated.
- Because of annotation rules that changed during or since T1, certain *test-CONTAINS-results* may be "missing"; NOTED-ON links may be added in these cases when obvious. For example, T1Temp guidelines originally specified that modifiers shouldn't be marked as EVENTs (see pp. 9-10 and p. 14 of the T1Temp guidelines). In practice, again, this rule changed over the course of the T1 project in an effort to capture clinically-significant information. As a result, some premodifying EVENTs that did not exist in T1 were added in T2 in the synchronization pass in order to make the annotated markables as consistent as possible. These EVENTs, therefore, might not be in test-CONTAINS-results links:

  (d) [Pathology] noted 2 of 25 [metastatic] lymph nodes.

  - ☐ No existing link between [Pathology] and [metastatic], likely because [metastatic] wasn't an EVENT in T1.
  - ☐ *Create*: [metastatic] NOTED-ON [Pathology]
- Closely tied to the preceding guideline: When deciding whether a new (versus converted) NOTED-ON link is warranted, try to reflect the language in the same way that THYME 1 did. Informally, T1 applied the *test-CONTAINS-results* rule for cases where the author's language indicates "these are the things we saw/learned from this test" versus "this is here but we already knew it." For example:

  (e) Today's [scan] shows the [mass] has [grown].

  - ☐ *Delete*: [scan] CONTAINS [grown]
  - ☐ *Create*: [grown] NOTED-ON [scan]
  - ☐ No NOTED-ON relation between [mass] and [scan]
- If there are multiple *test-CONTAINS-results* links for IDENTICAL mentions of an EVENT, change only the first (local) one to NOTED-ON. For example:

  (f) [Pathology] showed [adenocarcinoma]. [Tumor] is non-invasive.

  - ☐ [adenocarcinoma] IDENT [Tumor]
  - ☐ *Delete*: [Pathology] CONTAINS [adenocarcinoma]
  - ☐ *Delete*: [Pathology] CONTAINS [Tumor]
  - ☐ *Create*: [adenocarcinoma] NOTED-ON [Pathology]
- Because different reviews may note different (and sometimes conflicting) findings, when a test has multiple reviews, link the results to the review EVENT instead of the test EVENT:

(g) Outside [review]$_1$ of Mr. Jones' CT-scan found [thickening] in the sigmoid colon but this was questionable on local [review]$_2$.

□ [thickening] NOTED-ON [review]$_1$

- Sometimes treatment procedures like surgeries were treated as "tests" in T1 in terms of the *test-CONTAINS-results* rule. Relations associated with these EVENTs should therefore also be checked and converted as appropriate:

(h) [Resection] showed fluid [collection] in the abdomen.

□ *Delete*: [Resection] CONTAINS [collection]

□ *Create*: [collection] NOTED-ON [Resection]

- General (non-diagnostic) test results (*negative*, *positive*, *normal*, *unremarkable*, etc.) can also be linked by NOTED-ON to the tests.
- Everything that we consider to be tests may be linked via NOTED-ON to their test results, including Vital Signs "tests" (see Vital Signs) and physical examinations:

(i) Pulse [Rhythm]=[regular]

□ [regular] NOTED-ON [Rhythm]

(j) Spine: [Examined] and [normal].

□ [normal] NOTED-ON [Examined]

- Terms like "auscultation," "palpation," etc., don't count as tests. We consider these to be more manner-like terms than tests. Delete CONTAINS links if present and don't add NOTED-ON links.

(k) Lungs [clear] to [auscultation].

□ No NOTED-ON link.

- We do not need to check pathology notes for *test-CONTAINS-results* links nor add NOTED-ON links, since the only TLINKs present in pathology notes are CONTAINS-SUBEVENT relations. See Pathology Note Checks.
- To aid in finding tests-CONTAINS-results links for this task, here is a **sample** list of terms that may have been considered "tests" in THYME 1:

Broad	Surgery	Pathology	Labs	Imaging	Physical examination	Other
-evaluation	-surgery -surgical exploration -proctectomy -resection -hemicolectomy *etc.*	-pathology -path	-CBC -hemoglobin *etc.*	-x-ray -CT scan -MRI -ultrasound *etc.*	-exam -examined	-EGD -colonoscopy -barium enema study -fecal occult study -hemoccult stool testing

						etc.

(3) Adds TLINKs where needed for *procedure* semantic scripts.

- While we can't infer temporal relations for many clinical events unless explicitly stated in the text (such as the temporal relation between, say, a CT-scan and bloodwork during a given patient visit), others are inherently ordered. A piece of tissue must be [removed] prior to performing [pathology] on that tissue, for example, so we can say *[pathology] BEFORE [removed]*, even if that temporal relation is not expressed in the language of the text.
- T1Temp avoided this type of linking because it puts "a heavy burden on the annotation process" (Styler et al., 2014). However, in practice, as annotators learned these consistent temporal relations between certain medical events over time, many of these sorts of links were in fact already present in T1Temp gold data.
- Additionally, primarily through the addition of the subevent relation, the current project was already doing this to some degree – that is, we infer relations such as *[surgery] CONTAINS-SUBEVENT lymph node [removal]*, even if the text doesn't explicitly indicate temporal containment between two such events.
- For these two reasons and because it provides a more complete and consistent timeline of patient information, we intentionally added TLINKs for inherently ordered elements of all procedural events where they weren't already present. These included:
  - surgical procedure/diagnostic test CONTAINS biopsies/removal
  - procedure/test BEFORE pathology
  - pathology BEFORE report
  - pathology BEFORE staging
  - pathology BEFORE results/findings
- A note on the term *biopsy*: "Biopsy" can mean the act of removing tissue for study, the examination of the tissue, or the tissue itself. In an attempt to be as consistent with T1Temp as possible, we interpret all references to "biopsy" as being to the event of removing the tissue; so, *[colonoscopy] CONTAINS [biopsy]* rather than *[colonoscopy] BEFORE [biopsy]*. (We also always treat "biopsy" as an EVENT, not as an entity referring to the tissue specimen itself.)
- We did not expand this type of TLINKing to other types of narratives in the text, although this could certainly be an area of future work if desired.

Appendix D: Schemas at a glance

Tables 1-3 below, with following comments, summarize the T1Temp, T1Coref, and THYME 2 annotation schemas, along with the major modifications that occurred between THYME 1 and THYME 2. Figure 1 visually elaborates on the CONTAINS and OVERLAP TLINK changes between THYME 1 and THYME 2.

Table 1: **THYME 1 temporal relations schema (T1Temp)**

T1Temp markables	**T1Temp relations**	
• EVENTs (*any conceptual event, regardless of POS*)[1] • TIMEX3s • DOCTIMEs • SECTIONTIMEs	*TLINKs*	*ALINKs*
	• BEFORE • OVERLAP • CONTAINS • BEGINS-ON • ENDS-ON	• CONTINUES • INITIATES • REINITIATES • TERMINATES

Table 2: **THYME 1 coreference relations schema (T1Coref)**

T1Coref markables	**T1Coref relations**
• MARKABLEs (*any non-singleton noun, noun modifier, pronoun, or nominalized verb; may be an event, entity, or temporal expression*)[1]	• IDENTITY • APPOSITIVE • SET-SUBSET • WHOLE-PART

Table 3: **THYME II temporal and coreference relation annotation**

Markables		
• EVENTs • Entities (*including singleton entities*[1, 2]) • TIMEX3s • SECTIONTIMEs • DOCTIMEs		
Single-file relations		
TLINKs	*ALINKs*	*Coreference links*
• BEFORE • OVERLAP • NOTED-ON[2, 3] • CONTAINS • CONTAINS-SUBEVENT[2, 4] • BEGINS-ON • ENDS-ON	• CONTINUES • INITIATES • REINITIATES • TERMINATES	• IDENTICAL • APPOSITIVE • SET-SUBSET • WHOLE-PART[5] • CONTAINS-SUBEVENT[2, 4]
Cross-doc relations		
• IDENTICAL • SET-SUBSET • CONTAINS-SUBEVENT • WHOLE-PART[5]		

Comments on tables 1-3:

[1] Note that T1Coref's markables were determined by POS and whether there was a coreferential within-document mention; T1Temp's markables were determined by eventiveness, regardless of POS or coreference. Singleton entities were therefore unmarked

by either original project and were added new in THYME 2. (All entities are called MARKABLEs in THYME 2's output due to T1Coref's labeling practices. See Terms: entities, markables, and MARKABLEs.)

[2] These terms indicate **new annotation categories** added in THYME 2.

[3] NOTED-ON is a subtype of OVERLAP. See discussion of NOTED-ON under Manual post-processing tasks in Appendix C.

[4] CONTAINS-SUBEVENT is a subtype of CONTAINS. It appears under both TLINKs and Coreference links because it represents both temporal information (temporal containment) and partial-identity information (event-subevent). While other TLINKs have distance restrictions, CON-SUB does not. It was restricted to four semantic categories: cancer treatment events; cancer events; medication usage events; and chronic disease events. See Adding CONTAINS-SUBEVENT links.

[5] WHOLE-PART application dramatically changed between THYME 1 and THYME 2. T1Coref usage subsumed compositional, locative, and subevent relations; THYME 2 usage restricts to compositional whole-part relations between entities only. THYME 2 cross-document application restricts this even further; it was used only for non-anatomy WHOLE-PART relations. See THYME 2 Anatomy Linking guidelines, Cross-Document Anatomy Pass, for discussion.

Figure 1: THYME 1 to THYME 2 TLINK modifications

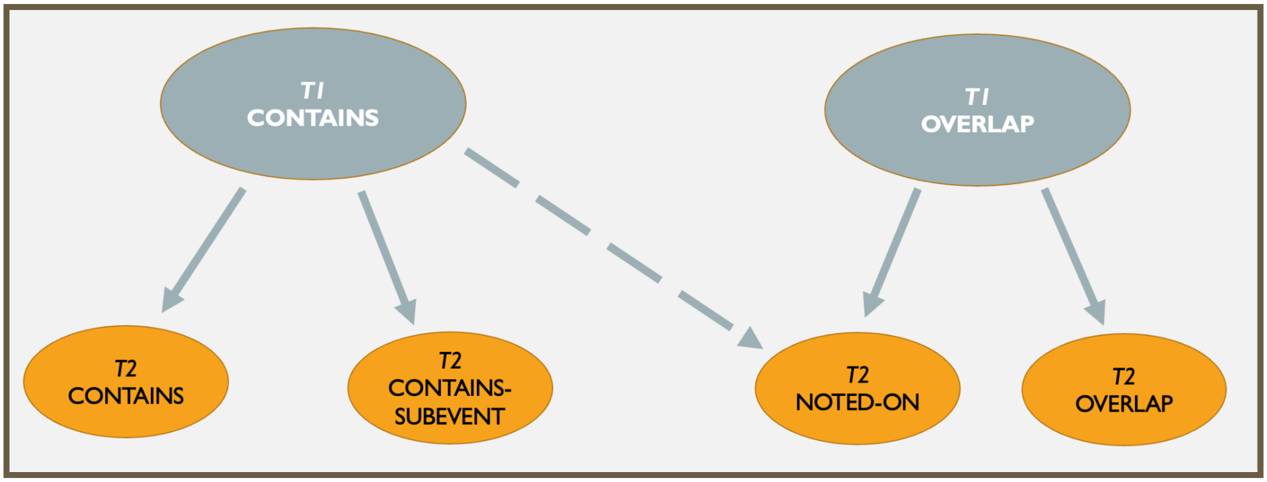

Figure 1 comments: **THYME 1** had a single CONTAINS relation and a single OVERLAP relation. **THYME 2** divided T1 CONTAINS links into two subtypes: CONTAINS, which represents temporal containment only; and CONTAINS-SUBEVENT, which represents temporal containment plus partial identity. THYME 2 also re-categorized a specific subset of T1 CONTAINS links as a subtype of OVERLAP and called this subtype NOTED-ON. This subset consists of temporal relations between tests and their results. THYME 1 specified these TLINKs should always be CONTAINS, such that the test CONTAINS its results (*Biopsy demonstrated cancer* > biopsy CONTAINS cancer); THYME 2 converted these to NOTED-ON. This is discussed in detail in Appendix C. The dotted-line arrow above therefore indicates a previous (T1) relationship; the solid-line arrows indicate current (T2) relationships.

Appendix E: Anatomy Linking guidelines

- THYME 2 Anatomy Linking Guidelines: Within-doc and cross-doc linking guidelines for the anatomy passes for this project.

Appendix F: References

Massimo Poesio, *The MATE/GNOME Proposals for Anaphoric Annotation, Revisited*. Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL, 2004. Association for Computational Linguistics.

Massimo Poesio and David Traum, *Conversational Actions and Discourse Situations*, Computational Intelligence 13(3), 1997.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary, *Iso-timeml: An international standard for semantic annotation*, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10) (Valletta, Malta.), European Language Resources Association (ELRA), 2010.

James Pustejovsky and Amber Stubbs, *Increasing Informativeness in Temporal Annotation*, Linguistic Annotation Workshop, 2011, pp. 152–160.

Guergana Savova, Wendy Chapman, Jiaping Zheng, and Rebecca Crowley, *Anaphoric relations in the clinical narrative: corpus creation*. J Am Med Inform Assoc. 2011;18(4):459. doi: 10.1136/amiajnl-2011-000108.

William F. Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova and James Pustejovsky, *Temporal Annotation in the Clinical Domain*. Transactions of the Association for Computational Linguistics, 2, pages 143–154. 2014. Association for Computational Linguistics.

Appendix G: Document Revision History

11/3/16 by Kristin Wright-Bettner:

- Compiled preannotated data info (gathered from T1Temp, RED, and T1Coref guidelines), first pass, second pass, and third pass into single document. Made the following major changes:
    - Took out SingletonEVENTs section since we converted all SingletonEVENTs back to normal EVENTs in Anafora. (SingletonEVENTs were EVENTs that did not participate in a coreference relation and had been visualized differently for the first part of first-pass production.)
    - Updated entire document to account for addition of subevents/M-conversion annotation pass
    - Updated section on Conjoined Noun Phrases to reflect that we are now preserving these and adding WHOLE-PART relations for them.
    - Added that Section 20148 shouldn't be annotated (apart from checking links for MARKABLEs already created by Coref)
    - Added discussion of metachronous and synchronous cancers
    - Changed handling of IDENT links for surgical procedures, treatments
    - Clarified that all disease and cancer mentions should be IDENT, if they refer to the same disease
    - Updated Properties and their states section to note that sounds are EVENTs and that sometimes properties should be double-tagged as EVENTs
    - Added Quantifiers section to Doubly-marked events
    - Moved medications/allergies sections coreference checks from first pass to second pass
    - Added Delete relative positions section
    - Added negated SUBEVENTs section to second pass guidelines
    - Added appendices B, C, and D

6/6/17 by Kristin Wright-Bettner:

- Merged the first and second passes of the Merged notes pipeline (i.e., combined the data synching pass and the subevents pass into a single pass)
- Added Appendix D, The rest of the pipeline (*later changed to Appendix C*).
- Made the following significant changes as well as smaller ones:
    - Added III 1.3.1.6 Events of location
    - Clarified that nonsense strings shouldn't be marked
    - Updated SUBEVENT sections to reflect that we're also now annotating medications and chronic disease subevents
    - Clarified annotating "cycles of chemotherapy"
    - Added that we are no longer doing coreference or subevent linking for HYPOTHETICAL or GENERIC modalities (this is a change)
    - Clarified handling of general/specific cancer and cancer treatment terms (also a change)

- Added how to fully annotate medications and allergies sections (prior to this we were only going to add what had a cross-document coreference link)
- Added that we are not annotating the formula paragraph in pathology notes
- Added how to handle tumors in different locations

December 2018-January 2019 by Kristin Wright-Bettner:

- Added Appendix E and post-processing information under Appendix D (*later changed to Appendix C*).
- Updated modality linking guidelines in cross-document guidelines.
- Clarified and updated medications section guidelines.
- Clarified how to mark normally-occurring bodily entities such as blood and stool.
- Corrected marking body systems instructions.
- Added condensed categories of consistent hierarchical links with examples.
- Corrected *Organs: Parts vs. Wholes* guidelines.

December 2021 by Kristin Wright-Bettner:

- Expanded Appendix D, Schemas at a glance.
- Added NOTED-ON guidelines to Appendix C.
- Edited the Introduction for clarity.
- Changed project labels throughout, in main guidelines and anatomy-linking guidelines: Converted “THYME” to “T1Temp” and “Coref” to “T1Coref.”