ANNOTATION GUIDELINES FOR THE FLORIDA ANNOTATED CORPUS FOR TRANSLATIONAL SCIENCE (FACTS)

Amanda Hicks, Selja Seppälä, Nathan Boire, Chloe Herring

Version: March 2, 2017

1 INTRODUCTION

These annotation guidelines are created for the development of the Florida Annotated Corpus for Translational Sciences (FACTS). This document specifies how to annotate patient level statements in case reports with ontology classes.

2 DEFINITIONS AND NOTATION USED IN THIS DOCUMENT

individual – A particular thing in reality that instantiates (i.e., is an instance of) some universal. Individuals are non-repeatable entities in reality. Examples include, Theodore Roosevelt (an instance of the universal HOMO SAPIENS), the Eifel Tower (an instance of a BUILDING or PIECE OF ARCHITECTURE), and Roosevelt's polio (an instance of POLIO). Individuals are also members of classes.

universal – A repeatable feature of reality that exists only as instantiated in a respective individual.[1] For example, the universal POLIO is instantiated in Roosevelt's polio. Universals are often denoted by general nouns.

denotes – A relationship between a symbol (such as a string and a unique entity in reality that it represents. The name 'Theodore Roosevelt' denotes the man who went by that name. Denotation picks out a single, unique entity.

word – A string coupled with the meaning of that string. The words *quick* and *fast* have the same meaning, but are different words since they consist of different strings.

token – A single occurrence of a word, string, or symbol in a text. The sentence "The quick brown fox jumped over the lazy dog." contains two tokens of the word *the*.

class – The collection of all things to which a particular word can be said to correctly apply. That is, the extension of a word. The class of all human beings is the collection of things to which *human being* can be said to apply. The class of cases of polio is the collection of each disease that is polio in some organism.

delimiter – A character that serves as a boundary.[2] Spans of text are delimited by boundaries or delimiters, thereby indicating which spans of text can be annotated and which spans cannot.

The following can serve as delimiters:

- the beginnings and ends of documents,
- punctuation and special characters (-,!.;?:#@%^&*()_+=[]\|<>),
- white spaces such as spaces, carriage returns and tabs,

Letters and numbers are *not* delimiters.

single quotes

Single quotes (‘yyy’) will be used to refer to strings in text.

small caps

Small caps (YYY) will be used to refer to universals.

italics

Italics (*yyy*) will be used to refer to words.

angle brackets

Angle brackets (<yyy>) will be used to refer to annotation tags.

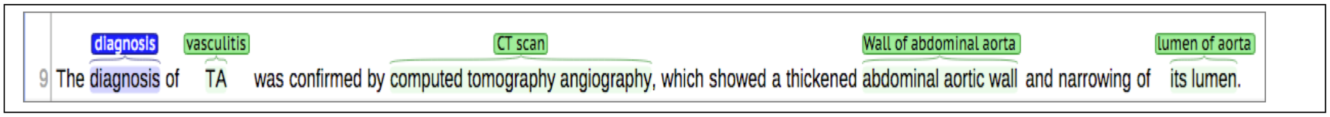
backslashes

Backslashes (\yyy\) will be used to refer to representational elements in an ontology.

Example: When a token ‘A’ is annotated with the tag <B> corresponding to the ontology class \B\, it indicates that ‘A’ denotes an individual that is an instance of universal B (or that ‘A’ is a member of the defined class B).

3 WHAT THE ANNOTATIONS MEAN

When a token ‘A’ is annotated with class \B\, it indicates that ‘A’ denotes an individual that is an instance of universal B (or that ‘A’ is a member of the defined class B).



9 The diagnosis of TA was confirmed by computed tomography angiography, which showed a thickened abdominal aortic wall and narrowing of its lumen.

FIGURE 1 EXAMPLE OF ANNOTATIONS THAT INDICATE THAT THE TEXT DENOTES AN INDVIDUAL THAT IS AN INSTANCE OF A UNIVERSAL

The annotation <diagnosis> on the string ‘diagnosis’ indicates that ‘diagnosis’ denotes an individual that is an instance of DIAGNOSIS. That is, in this context the token ‘diagnosis’ directly refers to a specific diagnosis. Likewise, the annotation <vasculitis> on the string ‘TA’ indicates that ‘TA’ denotes an individual disease that is an instance of the universal VASCULITIS. That is, in this context, ‘TA’ denotes a specific disease, the patient’s vasculitis. The annotation <CT scan> on the string ‘computed tomography angiography’ indicates that ‘computed tomography angiography’ denotes an individual that is an instance of CT SCAN. That is, ‘computed tomography angiography’, in this context, denotes the specific procedure that was performed on the patient, and this procedure is an instance of CT SCAN, and likewise for the remaining annotations.

4 WHAT TEXT TO SELECT FOR ANNOTATION

4.1 DO NOT ANNOTATE TEXT IN THE TITLE.

4.2 ONLY WORDS IN SENTENCES THAT DENOTE INDIVIDUALS SHOULD BE ANNOTATED.

For example,

1. 'the patient complained of nausea'

is about an individual patient and her symptoms. Some of these will be annotated provided that there are appropriate classes in the ontology.

2. 'Takayasu arteritis (TA) is an inflammatory process frequently associated with stenosis and obliteration of the aorta and its primary branches.'
3. 'The patient had a suspected cortical adenoma'

Example 2 describes Takayasu arteritis generally, not TA associated with a particular person or patient. Since it does not contain any individual level statements about a particular individual, it will not be annotated.

Example 3 describes the epistemological status of the physician's diagnosis rather than an adenoma. Where there is epistemological uncertainty – that is, where it is not known or asserted to actually be the case that something exists – there should be no annotation. Likewise, when something is asserted to *not* be the case, there should be no annotation.

4.3 SPANS OF TEXT MUST BE DELIMITED ON BOTH SIDES.

4. ' blood pressure '

Example 4 contains three delimiters: the white space before 'blood', the white space between 'blood' and 'pressure', and the white space after 'pressure'. This means that there are three spans of text that can be selected for annotation, each of which is bounded on both sides by a delimiter: 'blood', 'pressure', and 'blood pressure'. Note that 'bl' and 'd pr' are not candidates for annotation. This is because they do not have delimiters on both sides.

5. ' 120/80 mmHg '

Example 5 contains four delimiters: the white space before '120', '/', the white space after '80' and the white space after 'mmHg'. This means that there are six spans of text that can be selected for annotation: '120', '80', 'mmHg', '120/80', '80 mmHg', and '120/80 mmHg'. Notice, not all of these spans would be annotated since they do not all correspond to something meaningful (e.g., '80 mmHg' does not mean anything). Nevertheless, this illustrates how the delimiters provide boundaries around what can potentially be annotated. Finally, note that 'mm' and 'Hg' are not candidates for annotation. This is because they do not have delimiters on both sides.

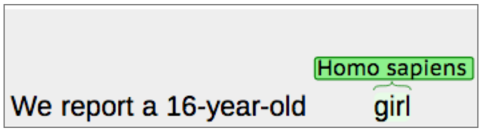
4.4 ANNOTATE THE SHORTEST SPAN OF TEXT THAT MOST CLOSELY FITS THE MEANING OF THE CHOSEN CLASS.

For example, in

6. 'We report a 16 year old girl'

‘girl’ is the shortest span to be annotated with <Homo sapiens>. Although the entire phrase ‘a 16 year old girl’ above refers to an instance of HOMO SAPIENS, rather than annotate the entire phrase, we will annotate only ‘girl’ since it is the shortest span that still refers to the same individual (cf. Figure 2).

Measurement values are often reported with units of measure. The units are an important part of the measurement datum. A length of 5 is not meaningful to somebody who does not know if this is five centimeters, five feet, or five miles. When measurement values are reported, include the units of measure in the span if they appear in the text. Figure 3 shows an annotation on a blood pressure measurement datum that includes the text denoting the unit of measure mmHg.



We report a 16-year-old girl

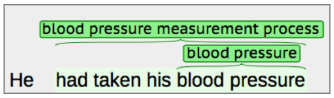
FIGURE 2 'GIRL' IS ANNOTATED WITH \HOMO SAPIENS\ RATHER THAN \FEMALE

4.5 INCLUDE AUXILIARY VERBS

For example, the string ‘had taken his blood pressure’ in

7. ‘He had taken his blood pressure’

should be annotated with the class \blood pressure measurement process\ rather than the shorter string ‘taken his blood pressure’ (cf. Figure 3).



He had taken his blood pressure

FIGURE 3 ANNOTATIONS SHOULD INCLUDE AUXILIARY VERBS

5 HOW TO SELECT CLASSES FOR ANNOTATION

5.1 ANNOTATE FROM THE CORRECT ONTOLOGICAL CLASS.

This will often require looking at superclasses in the hierarchy. For example,

8. ‘girl’

should be annotated with HOMO SAPIENS rather than FEMALE since the person denoted by ‘girl’ is a Homo sapiens and organism but is not a biological quality (cf. Figure 2 and Figure 4). Similarly in the sentence

9. ‘The patient presented with hypertension.’

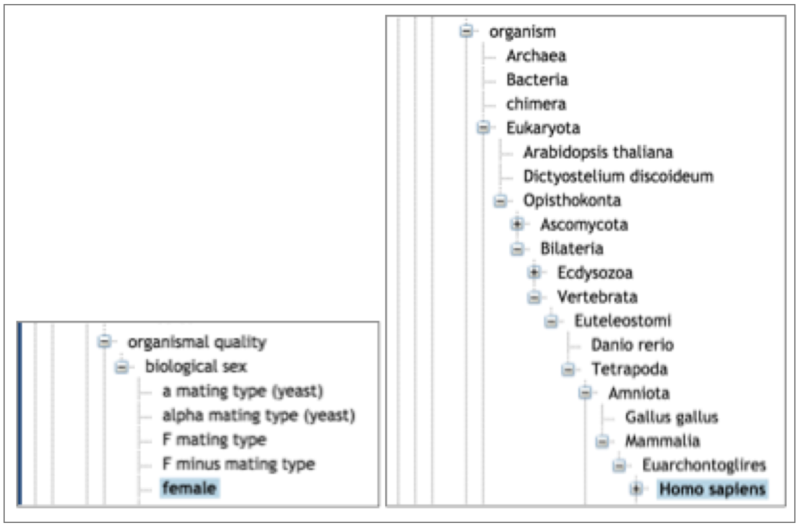


FIGURE 4 COMPARISON OF THE HIERARCHIES FOR \FEMALE\ AND \HOMO SAPIENS\ IN THE ONTOLOGY FOR BIOMEDICAL INVESTIGATIONS

the word ‘hypertension’ should be annotated with a class that describes the disease or disorder hypertension. It should not be annotated with clinical indicators of hypertension such as \blood pressure measurement datum\ or with classes that are conceptually related but indicate a distinct type of entity such as \blood pressure\.

Note that collections of instances of some universal A are usually not instances of A. For example,

10. ‘24 hour monitoring of blood pressure’

denotes multiple instances of measuring blood pressure and so denotes a set of blood pressure measurements. A collection of blood pressure measurement processes is not an instance of the universal BLOOD PRESSURE MEASUREMENT PROCESS, but instead is an instance of COLLECTION OF BLOOD PRESSURE MEASUREMENT PROCESSES. Therefore, it should not be annotated with the VSO class \blood pressure measurement process\.

Caution should be exercised when annotating measurement value ranges such as

11. ‘60-70 bpm’.

The text in Example 11 refers to a range of beats per minute, not a single measurement datum and so should not be annotated with a Vital Sign Ontology[3] (VSO) measurement datum class. In VSO, a measurement datum is always the output of some measurement process and always refers to a single entity. A measurement datum may have a value that falls within the range, but is distinct from the range of values itself.

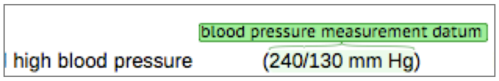


FIGURE 5 EXAMPLE OF AN ANNOTATION OF A MEASUREMENT DATUM. THE ANNOTATED SPAN INCLUDES THE UNITS OF MEASURE AND EXCLUDES PUNCTUATION MARKS.

5.1.1 TEXT THAT REFERS TO AN ABSENCE OF SOME ENTITY

It is often the case that clinicians report the absence of some entity, for example, the absence of a bilateral radial pulse as in the example below.

12. 'Bilateral radial pulses were absent'

Although 'bilateral radial pulses' appears in a statement about a patient, this text should not be annotated with the corresponding class. Such an (erroneous) annotation would entail that the patient *does* have bilateral radial pulses since all annotations in this project mean "X is an instance of the universal Y"; an absent pulse is not a pulse. If the ontology contains a class \absent bilateral radial pulse\, then the entire text 'bilateral radial pulses were absent' should be annotated with that class.

13. 'without any evidence of hypertensive retinopathy'
14. 'this would usually be associated with features of hypertensive retinopathy, which were absent in our patient'

Examples 13 and 14 state that the patient does not have hypertensive retinopathy, therefore the string 'hypertensive retinopathy' occurring in both examples will not be annotated a class \hypertensive retinopathy\.

5.1.2 TEXT SPAN THAT REFERS TO AN IMPORTANT ENTITY BUT DOES NOT HAVE A CORRESPONDING CLASS IN THE ONTOLOGY

There may be clinically relevant entities mentioned in the text that do not have a corresponding class in the ontology. For example,

15. 'The pulses in bilateral dorsalis pedis and posterior tibial arteries were weak.'

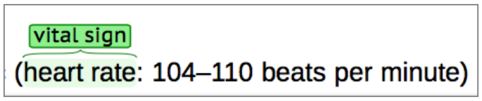
Indicates the strength of the patient's pulse. The VSO has a class \dorsalis pedis pulse rate\ but it does not have a corresponding class for the strength of the dorsalis pedis pulse. While this class is similar to strength of the bilateral dorsalis pedis pulse, strictly speaking the strength of a pulse is not the same as a rate of the pulse, so this text should not be annotated with the class \dorsalis pedis pulse rate\.

There may be a more general class that can be used to annotate the text. For example, the VSO has the class \quality\ from the Basic Formal Ontology. Just as the pulse rate is a quality of the artery, the pulse strength is a quality of the artery, so the span 'weak' can be annotated with the class \quality\.

Similarly, in Example 16

16. 'Initial vital signs revealed a tachycardic (heart rate: 104–110 beats per minute) and hypertensive (blood pressure: 179/121 mmHg) patient.'

\pulse rate\ is a class in VSO, but there is no corresponding class for heart rates. While 'heart rate' is sometimes synonymous with 'pulse rate' in clinical text, this string should only be annotated with the class \pulse rate\ if there is compelling reason to conclude that the pulse rather than the heart beat was measured. Otherwise, 'heart rate' is annotated with the more general class \vital sign\ (cf. Figure 6).



(heart rate: 104–110 beats per minute)

FIGURE 6 'HEART RATE' IS ANNOTATED WITH \VITAL SIGN\ IN THE ABSENCE OF EVIDENCE THAT IT WAS ACTUALLY THE PULSE RATE THAT WAS MEASURED

This raises the question of how far up the hierarchy one should go to find a proper class for annotation.

If there is a reasonable sibling class in the ontology (e.g., the class \pulse rate\is a reasonable sibling class of heart rate), then annotate the text with the immediate superclass. If there is no reasonable sibling class in the ontology, then do not annotate the text.

5.1.3 TEXT SPAN THAT IS A ONLY A PART OF A COMPLEX TERM

Language use can be quite flexible allowing us to take a part of a complex term to express the meaning of the complex term.

17. 'noninvasive blood pressure was 164/106 mm Hg'

In Example 17, the text does not mean that the blood pressure was noninvasive, rather that the blood pressure was measured using some noninvasive technique. In this case, the span 'noninvasive' will be annotated with the class \noninvasive blood pressure measurement process\.

5.2 Annotate with the most specific class such that *A is an instance of B* is true.

In Example 6, it is true that a girl is an instance of the universal MAMMALIA, but it is better to annotate 'girl' with \Homo sapiens\ since that is a more specific class.

This principle sometimes involves some degree of interpretation. For example,

18. 'Secondary causes for hypertension other than renal disease were ruled out'

states that the patient has some renal disease. However, we know from the title and the description of the case that the patient has more specifically End-Stage Renal Disease (ESRD). In this case, 'renal disease' is annotated with the corresponding more specific \XXX\ class.

5.3 Multiword expression with compositional meaning

When a text span corresponds to a multiword expression (MWE) for which there is no corresponding class in the ontology, but a part of the MWE does have a corresponding class in the ontology, annotate that subpart with the corresponding ontology class whenever the meaning of the MWE is compositional. For example, if the class \treatment-resistant hypertension\ is not in the ontology, but \hypertension\ is, and the meaning of 'treatment-resistant hypertension' is the sum of the meaning of its parts, then annotate 'hypertension'.

6 Special cases of annotations

6.1 Multiple Annotations on a Single Span

It is sufficient to have only one annotatation a span that indicates an individual. Some exceptions may be made for spans that have multiple annotations where the text denotes more than one individual and more than one universal is instantiated.

19. 'both legs'

Example 19 can be annotated with \left leg\ and \right leg\ (Figure 7). Where both classes are in the ontology, this is preferable to annotating only the token 'legs' with the class \leg\. Multiple annotations on a single span are permissible as long as one class used in the annotation is not a subclass of any of the others.

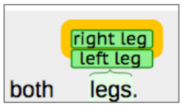


FIGURE 7 AN EXAMPLE OF MULTIPLE ANNOTATIONS ON A SINGLE SPAN

6.2 Nested Annotations

Nested annotations are permissible as long as one class used in the annotation is not a subclass of any of the others.

20. 'achieving normal BP values'

Example 20 is an example of a piece of text that can have nested annotations (Figure 8). The string 'BP' can be annotated with the class \blood pressure\, and the string 'achieving normal BP values' can be annotated with the class \regulation of blood pressure\.

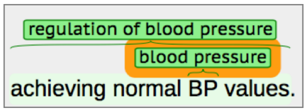


FIGURE 8 AN EXAMPLE OF NESTED ANNOTATIONS

6.3 Discontinuous Annotations

Discontinuous annotations are permissible when the relevant pieces of text are separated by text that denotes a different individual or is irrelevant and does not denote text at all.

21. 'blood pressure was monitored and controlled'

In Example 21, the string 'blood pressure was controlled' denotes an instance of REGULATION OF BLOOD PRESSURE. However, this span is discontinuous and interrupted by ' monitored and '. Since the string 'monitored' denotes an instance of a different universal (BLOOD PRESSURE MEASUREMENT PROCESS), it should not be included in the annotation. In this case the continuous string 'blood pressure was monitored' should be annotated with \blood pressure measurement process\ (Figure 9). The discontinuous string 'blood pressure was ... controlled' should be annotated with \regulation of blood pressure\ (Figure 9).

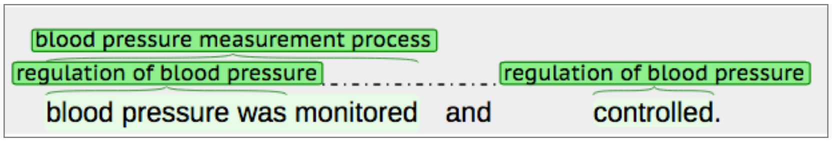

blood pressure measurement process
regulation of blood pressure
regulation of blood pressure
blood pressure was monitored and controlled.

FIGURE 9 AN EXAMPLE OF A DISCONTINUOUS ANNOTATION

7 Works Cited

1. Arp R, Smith B, Spear AD. Building ontologies with Basic Formal Ontology: MIT Press; 2015.

2. Bada M, Hunter LE, Eckert M, Palmer M. An overview of the craft concept annotation guidelines. Proceedings of the Fourth Linguistic Annotation Workshop; Uppsala, Sweden. Stroudsburg, PA: Association for Computational Linguistics; 2010. p. 207-11.

3. Goldfain A, Smith B, Arabandi S, Brochhausen M, Hogan WR. Vital sign ontology. Bio-Ontologies 2011 [Internet]. 2011 2011 July 15-16:[71-4 pp.]. Available from: http://bio-ontologies.knowledgeblog.org/155.