LLT Lab Annotation Protocol (Public): Sentence Type

The following guidelines and examples govern our annotation of spans of text using the type “Sentence”.

Part I: Guidelines[1]

We adopt a fairly standard definition of a “sentence” from the linguistic theory that deals with written linguistic structures. A sentence is defined as a textual segment or span consisting of one or more words that are grammatically linked, and which expresses (in principle, and at least implicitly) a complete thought (even though it may be necessary to take context and conventions into account to understand that thought). Sentences might express a declarative statement, a question, an exclamation, a request, a command, or a suggestion.

Normally, an explicitly expressive sentence contains a grammatical subject and a grammatical predicate. The grammatical subject is normally a noun phrase (a group of words that are in dependency relations with a single noun), and refers to the person, place or thing (including abstract things) that the sentence is about. The grammatical verb is normally a verb phrase (a group of words that are in dependency relations with a single verb, which in turn refers to an action, process or state). An example of a normal sentence structure is “The veteran filed a claim for disability benefits,” where “the veteran” is the grammatical subject and “filed a claim for disability benefits” is the grammatical predicate.

Not all sentences are explicit in what they express. An example of a one-word sentence with implicit meaning is “Yes.” when it is an answer to the interrogatory sentence “Did you seek medical attention for your condition?” In context, the sentence “Yes.” has the same meaning as the explicit sentence “I did seek medical attention for my condition.” Other examples of spans that we will annotate as sentences (see the examples below in List C) are:

- Data fields, such as “Decision Date: 03/28/17” - which we understand as having the same meaning as “This decision was issued on March 28, 2017.”

[1] Normally, a Lab protocol contains more explicit and precise procedures for identifying an annotation type and its boundaries, but this protocol builds on an English writer’s intuitions about sentence content, grammar and sentence boundaries.

- Headings, such as “FINDINGS OF FACT” - which we understand as having meaning similar to “The sentences in the following section state the findings of fact of the tribunal.”
- List numbers or letters, such as “1.” - which in context we understand as meaning “The following is the first item on the list.”

In selecting the span of characters that we annotate as type “Sentence”, be sure to highlight and annotate the entire sentence. There are normally several linguistic features to guide you. Any **white space** before the sentence is not part of the sentence span. Normally a sentence in English begins with a **capital letter** (i.e., the first character within the span selected as being a sentence is normally a capital letter). The **punctuation** at the end of the sentence is part of the sentence span. Hence, no annotation span should start or end with a white space.

Sentences in English normally end with one of three sentence-ending punctuation characters: a period (also called a “full stop”), a question mark, and an exclamation mark. We seldom see exclamation marks (“!”) in legal texts, but they do occur occasionally. In BVA decisions, you will sometimes see sentences ending in a question mark, especially in a section of the decision specifying issues to be addressed on remand. Normally, sentences end with a period (full stop) - although periods also cause extensive segmentation error, because many periods in legal texts do not signal the end of a sentence (e.g., periods within abbreviations or citations). You should assume that a sentence never ends with a comma, a semi-colon, or a colon (although in the case of a colon, we will see an exception in the examples below when the colon introduces a numbered list or a block quote, and should be regarded as ending a sentence).

Part II: Lists of Examples (Organized into 3 lists)

List A: Normal Form -- A span of characters is a "normal form" if we are highly confident that it constitutes an annotation of the type “Sentence”, and this confidence is based on some evidence or feature from the span text itself (the "linguistic cue"). Also, a sentence in "normal form" has a certain fixed format or pattern, which we find recurring numerous times. These should be the easiest types of sentences for computer software to identify.

In the case of annotations for the type “Sentence”, the normal form is a grammatical subject consisting of a noun phrase followed immediately by a grammatical predicate consisting of a verb phrase. The noun phrase and verb phrase can contain subordinate clauses, provided the sentence as a whole is relatively easy to parse by parts of speech.

EXAMPLES:

The Veteran's chronic adjustment disorder with depressed and anxious features is related to service.

[NOTE: the noun phrase is “the Veteran’s chronic adjustment disorder with depressed and anxious features”; the verb phrase is “is related to service”.]

The Veteran does meet the criteria for a diagnosis of posttraumatic stress disorder (PTSD).

[NOTE: the noun phrase is “the Veteran”; the verb phrase is “does meet the criteria for a diagnosis of posttraumatic stress disorder (PTSD)”.]

A disability which is aggravated by a service-connected disability may be service-connected.

[NOTE: the dependent clause “which is aggravated by a service-connected disability” modifies “disability” and is part of the noun phrase that is the grammatical subject of the sentence. The verb phrase is “may be service-connected”.]

List B: Linguistic Transforms of Normal Forms -- A span of text that is a "linguistic transform" of a normal form is one for which we are also highly confident that it constitutes an annotation of the type “Sentence”, a confidence based on some evidence or feature from the span text itself (the "linguistic cue"). However, while a sentence in normal form has a straightforward format or pattern of <grammatical subject noun phrase><grammatical predicate verb phrase>, a "linguistic transform" has a more linguistic pattern that is in principle transformable into sentences that do have normal forms. There might be some linguistic rules that would make it easier for computer software to identify such forms. So here we are looking for merely grammatical transformations of the normal form for sentences.

EXAMPLES:

Consequently, as outlined in a February 2013 Formal Finding, the RO requested information from the Joint Services Records Research Center (JSRRC) and the US Army Crime Records Center; however, those sources provided negative responses for the requested date range.

[NOTE: This sentence has a complex grammatical structure that departs from the simple normal form. It can be parsed into two independent clauses, the second of which has a sentence normal form. The first clause, however, contains the grammatical subject in the middle of the clause.]

Establishment of service connection for PTSD in particular requires: (1) medical evidence diagnosing PTSD; (2) credible supporting evidence that the claimed in-service stressor actually occurred; and (3) medical evidence of a link between current symptomatology and the claimed in-service stressor.

[NOTE: This is a single sentence that contains a numbered list. This is to be distinguished from the numbered lists discussed in List C.]

See, e.g., Young v. McDonald, 766 F.3d 1348, 1353 (Fed. Cir. 2014) ("PTSD is not the type of medical condition that lay evidence . . . is competent and sufficient to identify.").

[NOTE: This example illustrates that citation sentences are indeed sentences, even though they do not have the normal form of a sentence. But they do express complete thoughts in highly encoded form. This citation sentence, for example, means the following: “The decision by the U.S. Court of Appeals for the Federal Circuit in 2014 named ‘Young v. McDonald’ contains, at page 1353 of volume 766 of the third series of the Federal Reporter, the sentence ‘PTSD is not the type of medical condition that lay evidence . . . is competent and sufficient to identify.’.” The fact that the Bluebook of formats for legal citations is so thorough leads us to put citation

sentences in List B (linguistic transforms of normal forms) instead of in List C (aberrant forms of normal forms).]

List C: Aberrant Forms of Normal Forms -- A span of text can have a very unusual linguistic structure, but we are still confident (but maybe not "highly confident") that it constitutes an annotation of the type "Sentence". This confidence might be based more on the context than on linguistic cues within the span itself (e.g., co-references with words or phrases in other sentences). These we will have to study to determine what reasons we can give for annotating them as sentences.

There are certain classes of examples on which we can generalize.

Headings or Organizational Sentences

Some spans of text that we will annotate as "sentences" organize the text (chunk it into meaningful segments). We will annotate them as "sentences" (see the guidelines at the beginning of this protocol).

EXAMPLES:

ISSUES (or THE ISSUE)
INTRODUCTION
FINDINGS OF FACT
CONCLUSIONS OF LAW
REASONS AND BASES FOR FINDINGS AND CONCLUSIONS
ORDER
REMAND

Data-Field Sentences

Some spans of text that we will annotate as "sentences" provide a data field and a value (see the guidelines at the beginning of this protocol). They implicitly assert the value of the data field.

EXAMPLES:

Citation Nr: 1710389
Decision Date: 03/28/17
Archive Date: 04/11/17
DOCKET NO. 12-12 279
Veteran represented by: Veterans of Foreign Wars of the United States

RYAN T. KESSEL
Veterans Law Judge, Board of Veterans' Appeals Department of Veterans Affairs

[NOTE: As a matter of pure convention, for BVA decisions we will annotate the signature block of the decision as two sentences, the first stating the judge's name, and the second stating the judge's title (even if "Department of Veterans Affairs" is on a separate line in the decision).]

Numbered or Lettered Lists

Numbered or lettered lists require special treatment, whether the list items are themselves sentences or not. By convention, we will annotate the numbers or letters themselves as separate sentences (see the guidelines at the beginning of this protocol).

EXAMPLES:

FINDINGS OF FACT
1. The Veteran does meet the criteria for a diagnosis of posttraumatic stress disorder (PTSD).
2. The Veteran's chronic adjustment disorder with depressed and anxious features is related to service.

[NOTE: This passage contains 5 sentences. The heading "FINDINGS OF FACT" is a sentence (see the guidelines and examples above). The two numbers introducing list items ("1." and "2.") are each sentences (see the guidelines and examples above). Then there are two list items, which are each sentences in normal form (see List A above). Notice the trouble we would create for machine learning or for our own understanding if we included the list-item number ("1.") as part of the sentence that is the list item. The "1." expresses a thought separate from the sentence it introduces. The numbering of the list could change, but the role and meaning of each sentence on the list would remain the same. We don't want an ML program treating "1." as part of the sentence it introduces, or trying to POS parse the sentence including the "1."]

I. Laws and Regulations

[NOTE: A numbered or lettered heading should be two sentences. In the example above, "I." should be one sentence (making a statement about where the heading is in a sequence of headings). The heading itself should be one sentence (its meaning normally unaffected by its order in the sequence of headings). See the reasoning for the previous example.]

NOTE: the following sentence is not a numbered list of sentences, but a single sentence that contains three clauses, which happen to contain numbers. The following example is a single sentence, and is included as an example of a linguistic transform in List B above:

Establishment of service connection for PTSD in particular requires: (1) medical evidence diagnosing PTSD; (2) credible supporting evidence that the claimed in-service stressor actually occurred; and (3) medical evidence of a link between current symptomatology and the claimed in-service stressor.

Colons Ending Sentences

There are at least two situations in which we will regard the colon as sentence-ending punctuation, whereas presumptively it is not (see, e.g, the previous example, which contains a included colon). The first is when the author introduces a block quotation with a sentence ending in a colon. For example:

> For example, in a June 1977 service personnel record a counselor opined that:
>
> I have personally interviewed this SM and found him to have a good attitude towards the Army. However, he has a serious academic problem.

This example contains 3 sentences. The first sentence ends with a colon: “For example, in a June 1977 service personnel record a counselor opined that:”. There are several reasons for deciding this. First, the sentence could have been written ending in a period with no loss of meaning, as in “... opined as follows.” As a matter of style, however, authors often use a colon to connect this sentence with what comes after. But more importantly, the assertion that the counselor opined what follows is a different thought than what the counselor opined. Better to keep the two thoughts in distinct sentences. Third (and most importantly), if we do not end the first sentence after the colon, where would we end it? To end it after the first quoted sentence is misleading, because there is a further sentence that the colon also includes within its scope. But including all three spans in a single sentence seems totally unacceptable. Finally, the quoted sentences should be annotated as sentences in their own right, parsed separately, etc., independent of the sentence that introduces the block quote. There seems to be little choice except to end the first sentence with a colon.

The second situation is when the span of text ending in a colon introduces a numbered list. The reasoning similar to that for block quotes. For example:

> Accordingly, the case is REMANDED for the following action:
>
> 1. When disability ratings and effective dates have been determined for all service-connected disabilities, to include those granted herein, and all development that the RO deems necessary is undertaken, the Veteran's request for a TDIU should be readjudicated.

This example contains 3 sentences: “Accordingly, the case is REMANDED for the following action:”; “1.”; and the sentence beginning with “When disability …”. Any alternative boundaries have strong reasons against them.

Grammatical or Typographical Errors

It occasionally happens that the human author of the document omits punctuation. If we can still determine from context and content that the span of text is a sentence (often because of what comes after it, together with its content), we should still annotate it as a sentence. For example, the following should be annotated as a sentence, despite the missing period at the end, because this span was followed by a normal sentence.

> 38 U.S.C.A. § 1111 (West 2014)