SEWTA Scheme for Witness Testimony Annotation: Annotation Manual

Contents

1 Preliminary Note

This document sets out recommendations for using SEWTA (SchEma for Witness Testimony Annotation) to enrich any corpora of witness testimonies with manual annotations.

This document is designed for use by practitioners in legal psychology or criminology, as well as by other researchers engaged in corpus creation or the assessment of witness memory, whether in terms of credibility or accuracy.

Finally, the encoding method is based on XML markup. Here are some general basic principles of the XML formalism.

- An XML tag is a markup construct that begins with < and ends with >. A well-formed tag should be formatted as follows: an opening tag (<tag>) and a closing tag (</tag>).
- XML has a tree-like structure where each tag can contain nested tags, but opening and closing tags must follow a proper hierarchy, similar to the correct use of parentheses and square brackets. For example, the following structure is incorrect: <de><sd>We were leaving the theatre</de></sd> because the tag <de> is closed before its nested tag <sd>.
- Tags may include attributes that provide additional information, whose value is enclosed in quotation marks. For example: <df part="skin" who="victim">white</df>.

2 Overview of the SEWTA Schema

SEWTA, the Schema for Witness Testimony Annotation, is a comprehensive framework designed to annotate corpora of witness testimonies. Utilizing a modular and extensible XML-like markup, SEWTA captures information at various levels allowing for precise and flexible annotation. In fact, the schema covers both project metadata and the testimony content and style, enhancing the documentation, preservation, and re-usability of the corpora. This detailed annotation enables researchers to analyze testimonies at different levels, exploring narrative styles and their relation to truthfulness, as well as investigating witness credibility. As such, SEWTA is an invaluable tool for forensic psychology, legal studies, corpus creation, and related research fields, accommodating the diverse needs of various annotation projects and ensuring robust data documentation and analysis.

SEWTA's distinctive feature is its grounding in forensic psychology literature, particularly research on witness memory analysis. It incorporates criteria from the following models: Criteria-Based Content Analysis (CBCA) (**?**), Reality Monitoring (RM) (**?**) and Global Evaluation Schema (GES) (**?**).

This document describes SEWTA and a set of fundamental tags designed to be general enough for use across most annotation projects. However, the schema can be extended to include tags relevant to specific research objectives. This allows for an expanded set of available tags while ensuring that only the necessary ones are selected and that any potential tag incompatibilities within the study are carefully verified.

SEWTA is structured into two main sections within the XML file:

Metadata of the Project and Witness Demographic Data: This section, which we refer to as the *Global Information section*, includes information about the project and demographic data about the witnesses, ensuring comprehensive documentation and contextual understanding of the testimonies.

Transcriptions of the Testimonies: This section, also referred to as the *Corpus Data section*, contains the detailed transcriptions of the testimonies, annotated with content and stylistic information according to SEWTA's schema.

This structured approach ensures comprehensive and detailed annotation, enhancing the utility and integrity of the annotated corpus for various analyses and research endeavours.

```
<?xml version="1.0" encoding="UTF-8"?>
<corpus name="">
    <globalInformation>
        </project>
        </witnesses>
    </globalInformation>

    <corpusData>
        <testimonyTranscription id="" sbj="" lang=""
                collectionTime="" narrationDuration=""
                session="">
    </corpusData>
</corpus>
```

3 Global Information Section

This section of the corpus is internally organised into two sections.

Project Section: Captures information about the context in which the testimonies were collected.

Witnesses Section: Records demographic information about the witnesses.

3.1 Project Section

The project section is enclosed within the `<project>` tag. This section includes tags that provide context about the project, such as its name, duration, participants, and the nature of the testimonies collected.

Example

```
<project>
    <projectName>Trial Testimony Collection (TTC)</projectName>
    <dataCollectionPeriod>
        <start>2023-01-01</start>
        <end>2023-12-31</end>
    </dataCollectionPeriod>
    <curators>
        <curator>
            <name>Dr. Jane Smith</name>
            <contact type="email">jane.smith@example.com</contact>
            <role>Researcher involved in data collation</role>
            <contactPerson>Yes</contactPerson>
        </curator>
        <curator>
```

```
            <name>Dr. John Smith</name>
            <contact type="phone">+1 (312) 657-5093</contact>
            <role>Attorney during trials</role>
            <contactPerson>No</contactPerson>
        </curator>
    </curators>
    <funding>Center for Crime Studies.</funding>
    <testimonies>
        <crime>Robbery</crime>
        <collectionContext>Actual Trials</collectionContext>
        <sessionCount>5</sessionCount>
    </testimonies>
    <dataAvailability>
        <ethicalStatements>Approved by Example Board</ethicalStatements>
        <license>CC BY-NC-SA 4.0</license>
        <repository>http://TrialTestimonyCollection.edu</repository>
    </dataAvailability>
</project>
```

Tags and Descriptions

- `<projectName>`: The name of the project.
- `<dataCollectionPeriod>`: The period during which data was collected. Internally organised into the `<start>` and `<end>` tags. The date format required is YYYY-MM-DD.
- `<curators>`: Information about the individuals involved in the project. Each person's information is reported within the `<curator>` tag using the following tags:
  - `<name>`: The person's complete name.
  - `<contact>`: The contact information of the person. The 'type' attribute specifies which type of contact detail is provided.
  -
  - `<role>`: The role of the person within the project.
  - `<contactPerson>`: A boolean value indicating whether that is the person to contact to obtain further information about the project and the data.
- `<funding>`: Information in free text about project funding.
- `<testimonies>`: General information about the testimonies collected, reporting the following information:
  - `<crime>`: The crime(s) discussed in the testimonies.
  - `<collectionContext>`: The context in which the testimonies were collected (e.g., actual trials or scientific study).
  - `<sessionCount>`: The total number of sessions required to collect the testimonies.

- `<dataAvailability>`: The section reporting information about the ethical context, licensing terms, and access methods and includes the following sub-tags:
    - `<ethicalStatements>`: Describes the ethical approvals and consent procedures followed during data collection.
    - `<license>`: Specifies the licensing terms under which the dataset is available.
    - `<repository>`: Provides information about where the dataset is stored and how it can be accessed.

3.2 Witnesses Section

The witnesses section is enclosed within the `<witnesses>` tag. Each witness's data is captured within a `<witness>` tag, identified by a unique attribute.

Example

```
<witnesses>
    <witness id="W001">
        <gender>Female</gender>
        <yearOfBirth>1985</yearOfBirth>
        <ageAtTestimony>37</ageAtTestimony>
        <motherTongue>English</motherTongue>
        <educationLevel>Master’s Degree</educationLevel>
        <eyewitness>Yes</eyewitness>
    </witness>
    <witness id="W002">
        <gender>Male</gender>
        <yearOfBirth>1990</yearOfBirth>
        <ageAtTestimony>33</ageAtTestimony>
        <motherTongue>Spanish</motherTongue>
        <educationLevel>Bachelor’s Degree</educationLevel>
        <eyewitness>No</eyewitness>
    </witness>
</witnesses>
```

Tags and Descriptions

- `<witness>`: Contains the demographic information for an individual witness. Each 'witness' tag has the attribute `id`, which assumes as value the unique identifier for the witness.
- `<gender>`: The gender of the witness.
- `<yearOfBirth>`: The year of birth of the witness.

- `<ageAtTestimony>`: The age of the witness at the time of the testimony.
- `<motherTongue>`: The mother tongue of the witness.
- `<educationLevel>`: The highest level of education attained by the witness.
- `<eyewitness>`: Indicates whether the witness attended the crime.

Customizing the Schema

SEWTA is designed to be extendable to accommodate project-specific requirements. While the basic tags cover general information applicable to most scenarios, additional tags can be included as long as they adhere to the XML formalism.

Example of Adding a Custom Tag

```
<witness id="W003">
    <gender>Non-binary</gender>
    <yearOfBirth>1993</yearOfBirth>
    <ageAtTestimony>30</ageAtTestimony>
    <motherTongue>French</motherTongue>
    <educationLevel>Doctorate</educationLevel>
    <eyewitness>Yes</eyewitness>
    <occupation>Research Scientist</occupation> <!-- Custom tag -->
    <nationality>Canadian</nationality> <!-- Custom tag -->
    <diagnosis>Depression</diagnosis> <!-- Custom tag -->
</witness>
```

Custom Tags and Descriptions

- `<occupation>`: A custom tag to include the witness's occupation.
- `<nationality>`: Records the nationality of the witness, which can be relevant for understanding cultural or legal contexts.
- `<diagnosis>`: Includes any relevant medical or psychological diagnosis that might impact the witness's testimony or its interpretation.

Similarly, custom tags can be added within the `<project>` tag to provide further details specific to the project's needs. Here are some examples.

- `<trial>`: Provides details about the trial context if the project involves actual court testimonies. This tag might be internally articulated into the following tags.
  - `<court>`: The court where the trial was held.
  - `<judges>`: The names of the judges presiding over the trial.
  - `<juryCount>`: The number of jurors.
  - `<trialDate>`: The date of the trial.

- <taskDescription>: Describes the task presented to participants in a scientific study. A possible value for this tag: ‘Participants were asked to recall and describe a robbery scenario based on a picture given as prompt’.
- <otherMedia>: Describes other media accompanying the testimonies, e.g. audio or video recordings of the testimonies or evidence associated with the crime, such as photographs or documents.

4 Corpus Data Section

The <corpusData> tag serves as a container for organizing the transcriptions of individual testimonies within the corpus. This tag is essential for structuring and managing the dataset, ensuring that each testimony is accompanied by detailed metadata for clarity and context.

The core of this structure is the <testimonyTranscription> tag, which is repeated for each testimony in the corpus. Each <testimonyTranscription> tag is enriched with several attributes that provide essential metadata, facilitating the organization and retrieval of information.

Example

```
<corpusData>
    <testimonyTranscription id="T001" sbj="W001" lang="English"
    narrationDuration="15m32s" session="1">
        <!-- Here a single testimony transcription. -->
    </testimonyTranscription>
    <testimonyTranscription id="T002" sbj="W002" lang="Spanish"
    narrationDuration="20m45s" session="1">
        <!-- Here a single testimony transcription. -->
    </testimonyTranscription>
</corpusData>
```

Attributes and Their Descriptions

- id: This attribute assigns a unique identifier to each testimony transcription. By ensuring that each transcription can be distinctly referenced, this identifier is crucial for maintaining the integrity of the dataset and avoiding any mix-ups or duplication of data.
- sbj: Standing for ‘subject’, this attribute links the transcription to the specific witness who provided the testimony. By using the witness’s unique identifier, this attribute ensures that each testimony can be traced back to its source, providing a clear connection between the witness’s demographic information and their testimony.
- lang: The language attribute indicates the language in which the testimony was given. This is vital for understanding and analyzing the content

accurately, especially in multilingual datasets where testimonies might be provided in different languages.

- `narrationDuration`: Specifying the length of the testimony, this attribute indicates how long the witness spoke. This information can be important for various analyses, such as examining the depth of detail provided or the witness's level of engagement.
- `session`: In cases where testimonies are collected over multiple sessions, this attribute notes the session number. It is useful for tracking the progression of testimonies and understanding how information might evolve over time.

5 Annotating Testimonies

The transcription of each testimony can be enriched with a set of tags designed to capture both the content and stylistic aspects of the narratives. Annotations may apply to a single word, a sequence of words, a sentence, or even span multiple sentences. Additionally, a phenomenon can be enclosed within two nested tags.

It is recommended to perform a fine-grained annotation. For example, it is preferable to use:

```
The thief had <df part="hair" who="thief">long hair</df> and
<df part="hair" who="thief">blonde hair</df>
```

rather than:

```
The thief had <df part="hair" who="thief">long hair and blonde
hair</df>
```

This approach enhances data analysis by distinguishing between witnesses who provide more detailed descriptions of a specific characteristic. Moreover, it facilitates the extraction of precise features relevant to the study.

5.1 Content Annotation

The SEWTA tagset is based on forensic psychology literature, enabling the modelling of multiple aspects of witness testimony that are crucial for assessing accuracy and credibility in legal proceedings.

The schema is structured into coherent groups of tags, described below. For a detailed discussion on how each tag contributes to accuracy and credibility evaluation, as well as an overview of the relevant literature, please refer to the SEWTA paper [**paper under evaluation**].

5.1.1 Episodic Autobiographical Memory Tags

Spatio-Temporal Information

These tags provide details about the location (`<spat>`) and timing (`<temp>`) of events, situating the testimony within a spatial and temporal framework. Both tags require the use of the `type` attribute (possible values: *orig* or *alt*, depending on whether the information is specific or generic, respectively).

Example

```
- I went <spat type="orig">to Rome</spat> <temp type="orig">
on Sunday afternoon</temp>.
- <temp type="alt">Later</temp> I decided to go <spat type="alt">
back</spat>.
```

Interactions and Dialogues

The <inter> and <convers> tags describe actions and dialogues involving multiple people, capturing the dynamics of interactions and conversations within the testimony. The <inter> tag is used for detailed descriptions of non-verbal actions involving at least two people. The text must describe two interconnected actions, highlighting the cause-effect relationship between them and specifying the individual responsible for each action. The <convers> tag captures direct or indirect speech reported in the testimony and is particularly useful for annotating questioning between the witness and an attorney. In such cases, an attribute might be added to specify the speaker.

Example

```
- <inter>Jack spilled his juice on me, so I yelled at him</inter>.
- <convers>The thief told me to give him the wallet</convers>.
```

Sensory Impressions and Mental States

The <sens> and <ms> tags document the internal experiences of the witness, capturing their sensory perceptions (e.g., what they heard, smelled, viewed, touched, or tasted) and psychological states (their feelings, thoughts, and overall psychological disposition). The <ms> can reflect also the witness's impression on the psychological state of other people involved in the story.

Example

```
- I <sens>heard noise</sens> from the other room.
- The child <ms who="victim">was frightned</ms>.
```

Physical Description

Physical descriptions of individuals are annotated using the <df> tag. These might include details about the individuals' appearance and characteristics. The attributes defined for this tag include who (the referent) and part (the body part or characteristic described).

Example

```
- The thief had <df who="thief" part="hair">red hair</df>.
```

Everyday Life Routines

The <life> tag captures contextual integration or how the speaker connects the event to their daily routine, providing context about their habits, circumstances, and social relationships.

Example

```
- <life>I always go to work at 9</life> and that day, on the way, I witnessed the car crash.
```

5.1.2 Script-Deviant Details Tags

Superfluous, Unusual and Unclear Details

The `<sd>` tag identifies peripheral details, not essential to the main narrative or accusation. Unusual details, annotated using `<ud>`, enable the identification of unexpected or surprising elements within the described context. Details that the witness declares to not comprehend are annotated using `<adnc>`. They identify information perceived as confusing or unclear.

Example

```
- A <sd>dog</sd> was <sd>barking</sd> <sd>to a man</sd>.
- <ud>The doctor took out a toy from the first aid kit</ud>.
- <adnc>I don’t know why he offered me tea before it all happened</adnc>.
```

External Associations and Unexpected Complications

The `<assoc>` tag is used to document connections made by the witness between the main event and other similar actions that occurred at different times. These associations can involve the same people or different individuals and serve to contextualize the witness’s experiences. When the witness reports situations that arise unexpectedly during the crime, disrupting the sequence of events, the `<compl>` tag can be used.

Example

```
- <assoc>I already knew what would happen next, after what happened
 last time</assoc>.
- <compl>She tried to unlock the door with a key, but when it didn’t
fit, she used a hairpin instead</compl>.
```

5.1.3 Memory-Related Shortcomings Tags

Memory Failings

The `<fm>` tag marks cases where the witness acknowledges a lapse in memory, highlighting uncertainty and potential gaps in the testimony. The `<rem>` tag captures the witness’s attempts to recall an event or detail, whether successful or not. Note that these efforts are often accompanied by speech pauses or sounds that signal cognitive processing, such as thinking or hesitating (see Sec. 5.2).

Example

```
- It was in Rome, but <fm>I don’t remember the name of the street</fm>.
- <rem>Wait a minute... let me think...</rem>.
```

Self-Questioning and Guilt

The annotation of instances of self-questioning may involve the use of two distinct tags, depending on the context. The <rd> tag is used when a witness acknowledges potential doubts about their testimony but ultimately affirms its truthfulness. This tag has two attributes: type="test" is used when the witness recognizes that their account may seem implausible but insists it is true; type="pers" covers cases where the witness asserts the reliability of their testimony despite having a questionable past or previous behaviours that might cast doubt on their credibility.

If the witness sincerely questions their own credibility or the reality of the events as they recall them, use the <rc> tag instead.

Use `<sdep>` when the witness expresses internal conflict, self-criticism, guilt, self-blame, or takes partial responsibility for the events.

Example

```
- <rd type="test">I know it's unlikely that someone could disappear so
quickly, but that's exactly what happened</rd>.
- <rd type="pers">I made mistakes in the past, but I swear this time I'm
telling the truth</rd>.
- <rc>I remember hearing a loud noise, but now I'm not even sure if it
actually happened or if I imagined it</rc>.
- <sdep>If only I had looked at his face more carefully</sdep>.
```

Perpetrator Exoneration

The `<pp>` tag captures the witness's attempts to exonerate or mitigate the actions of the perpetrator by mentioning positive behaviours, apologising on behalf of the perpetrator, or downplaying the consequences of their action.

Example

```
- <pp>Everyone would have done the same in that situation</pp>.
```

Spontaneous Corrections

The <err> tag is used to mark spontaneous corrections made by the witness without external prompting. The `corresp` attribute is used to indicate the corrected version of the statement. For example, if the witness initially states that a person's hair is long but then immediately corrects themselves to indicate the hair is short, the `corresp` attribute would capture this correction as in the example below.

Example

```
- He has got <err corresp="short">long</err> hair.
```

5.1.4 Cognitive Operation and Factual Error Tags

Cognitive Operations

The `<co>` tag is used to annotate various cognitive operations and mental processes involved in recalling and narrating events. When the witness engages in reflective thought or memory recall, the attribute `type="ins"` is used. If the witness explains reasons or justifies actions, the annotation should include `type="caus"` to capture causation-related expressions. When expressing confidence in their testimony, the `type="cert"` attribute is applied, whereas expressions of uncertainty, hesitation, or doubt are marked with `type="insecur"`.

Example

```
- <co type="ins">I remember seeing a white car</co>.
- The crash occurred <co type="caus">due to the driver’s speeding</co>.
- <co type="cert">No doubt</co>, it was him.
- <co type="insecur">Maybe</co> it happened earlier.
```

Commission Errors and Distortions

The `<ce>` and `<de>` tags are used to identify and annotate inaccuracies in witness testimonies. `<ce>` tag marks instances where false information is introduced, while the `<de>` tag is applied to altered or distorted details. Bboth tags require knowledge of the actual events.

Example

```
- The woman <ce>was driving a red bike</ce>.
- This happened <de>in Main Street</de>.
```

5.2 Stylistic-Level Annotation

5.2.1 Elements of Discourse

False Start

The `<fc>` tag is used when the speaker begins a sentence but then abruptly starts a new sentence instead. In the example below, the witness started to describe their activity but then provided a more specific account of who was involved.

Example

```
- <fc>I was walking</fc> My friend Maria and I were walking.
```

Repetition

The `<rep>` tag marks instances where a word or phrase is repeated consecutively by the witness, indicating a moment of stress or emphasis. The attribute `n` indicates the number of times the idea is repeated. In the example, the word ‘her’ is repeated twice.

Example

```
- He stole <rep n="2">her</rep> purse.
```

Hesitation

The `<vac>` tag captures hesitations in speech, often marked by elongated words or filler sounds. he elongated word is annotated in its correct form using the `in_text` attribute.

Example

```
- The woman was at <vac in_text="an">aaan</vac> ice cream stand.
```

Grammatical Correction

The `<err>` tag, specifically with the attribute `corrgr`, is used when the speaker self-corrects a grammatical error in their speech.

Example

```
- We <err corrgr="were">was</err> walking through the city
centre.
```

6 Complete Schema of SEWTA

```
<?xml version="1.0" encoding="UTF-8"?>
<corpus name="">
    <globalInformation>
        <project>
            <projectName/>
            <dataCollectionPeriod>
                <start/>
                <end/>
            </dataCollectionPeriod>
            <curators>
                <curator>
                    <name/>
                    <contact type=""></contact>
                    <role/>
                    <contactPerson/>
                </curator>
            </curators>
            <funding/>
            <testimonies>
                <crime/>
                <collectionContext/>
                <sessionCount/>
                </testimonies>
                <dataAvailability>
```

```
                    <ethicalStatements/>
                    <license/>
                    <repository/>
                </dataAvailability>
        </project>

        <witnesses>
            <witness id="">
                <gender/>
                <yearOfBirth/>
                <ageAtTestimony/>
                <motherTongue/>
                <educationLevel/>
                <eyewitness/>
            </witness>
        </witnesses>
    </globalInformation>

    <corpusData>
        <testimonyTranscription id="" sbj="" lang=""
        narrationDuration="" session="">
            A transcription of a testimony annotated using SEWTA.
        </testimonyTranscription>
    </corpusData>

</corpus>
```

7 Summary of Testimony Annotation Tags

Macro-class	Class	Name	Tag	Attributes	Attribute values
Content Annotation	Episodic Autobiographical Memory	- Everyday Life Routines	<life>	none	none
		- Spatial Information	<spat>	type	"orig” "alt”
		- Temporal Information	<temp>	type	"orig” "alt”
		- Iteractions	<inter>	none	none
		- Dialogues	<convers>	none	none
		- Sensory Impressions	<sens>	none	none
		- Mental States	<ms>	who	person
		- Physical Descriptions	<df>	part, who	part of body, person
	Memory-related shortcomings	- Spontaneous Corrections	<err>	corresp	right word
		- Memory Lapses	<fm>	none	none
		- Efforts to Remember	<rem>	none	none
		- Reality Controls	<rc>	none	none
		- Raising doubts	<rd>	type	"test" "pers"
		- Self-Criticism	<sdep>	none	none
		- Perpetrator Exoneration	<pp>	none	none
	Script-deviant details	- Unexpected Complications	<compl>	none	none
		- Superfluous Details	<sd>	none	none
		- Unusual details	<ud>	none	none
		- External Associations	<assoc>	none	none
		- Unclear Details	<adnc>	none	none
	Cognitive Operations	- Insight - Causation - Certainty - Insecurity	<co>	type	"ins" "caus" "cert" "insecur"
	Memory errors	- Commission	<ce>	none	none
		- Distortion	<de>	none	none
Stylistic-Level Annotation	Elements of Discourse	- False start	<fc>	none	none
		- Repetition	<rep>	n	number of repetitions (integer)
		- Hesitation	<vac>	in_text	right word
		- Grammatical Corrections	<err>	corrgr	right word