Annotation Guidelines

General Task Introduction

Why am I annotating?

This document details everything you need to know to start annotating. First off, it's important to know **why we ask you to label this data**. This data will be used for two Natural Language Processing (NLP)-tasks: 1. "Named Entity Recognition" (NER for short) and 2. "Sentiment Analysis" (SA for short). NLP is, as the name says it, a set of methods to digitally process natural text. For us humans, understanding language is a skill we're born with. We easily distinguish a river bank from a financial institute when we hear the word "bank" in context, we can distinguish Spanish from Dutch, and we can (mostly) distinguish when something is meant ironically or humorously. For computers however, language is one big enigma, and NLP engineers are trying to make natural language understandable to computers.

One of these NLP methods is **Named Entity Recognition** (NER): a set of automated methods to extract entities from a body of text. An entity can be anything that has a name. Entities which are often considered are people (e.g.: "Donald Trump") , organizations (e.g.: "Apple") and locations (e.g.: "Australia") – but it could also be more specific, such as body parts (e.g.: "nose") or even mythical animals (e.g.: "dragon").

Imagine you have a research project on witch trials. You have a large collection of 100 digitized court proceedings from the 16th Century, and you want to create an overview of all the people mentioned in these texts. Using traditional methods, this would take you hours of reading and logging all of the names in these documents. NER is a type of software which can speed up that process for you by automatically extracting all the names in these texts in a matter of minutes.

NER is currently quite popular in humanities settings. It can help libraries to automatically index their catalogue based on content, and it can also aid researchers with big digital text collections (also called "corpora") to perform their research faster and on a wider scale.

Another popular NLP-technique is **sentiment analysis** (SA). Sentiment analysis allows us to classify a text according to its expressed sentiment, namely "positive", "negative" or "neutral". This is different from **emotion analysis**, which allows for a more fine-grained classification (e.g.: "anger", "sadness", "jealousy", …).

We will be performing **aspect-based sentiment analysis**. This means that we will link the **entities we labeled before** to the **tokens/words which express an emotion about this entity**. Say we annotate all the **LOCATIONS** as entities. We'll also add the sentiment that is expressed regarding that location. In the example: "The Niagara Falls are extraordinarily beautiful.", "Niagara Falls" would be labeled as the entity LOCATION, "extraordinarily beautiful" would be annotated as "positive" and linked to the entity.

SA, and more specifically aspect-based sentiment analysis, is currently mainly used in the context of customer reviews. Websites such as Amazon analyze their customer reviews to see how people feel about certain aspects of the company, and where they can improve their services (CUSTOMER SERVICE, PRICE, DELIVERY, …). In the context of literature, it is currently very rarely applied! If you're interested in learning more on ABSA, certainly ~~check out the work by one of their colleagues~~ ~~on literary reviews in social media~~; and this interesting paper on the creation of an ABSA-annotated corpus for Swedish.
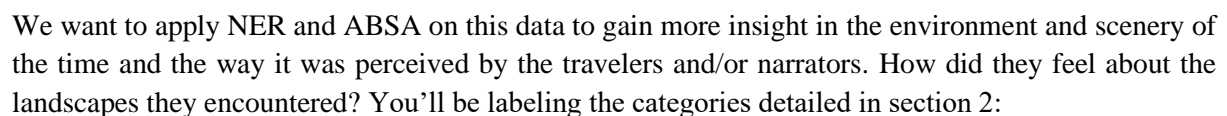
The downside? This software is of course not perfect. NER- and SA- software is currently mainly built to be applied to texts written in **modern languages**, and makes more mistakes when applying it to texts written in say, 18th century Spanish. The **digitization process** of these texts can also cause problems for NER. After scanning the texts with a scanner, Optical Character Recognition (OCR) is used to convert these texts to file type that can be read by the computer, and thus by NLP-tools (.txt). However, OCR is also not very good at reading historical texts. Think about the German Fraktur script as shown in the image below, and have a look at the transcriptions of "f" and "s". That's right, they're pretty much exactly the same.

A text like: *ich liebe fie* is thus easily misread by the OCR as "ich liebe **fie**". And that's only one minor example. Old texts are often stained or torn, which further complicates the digitization process. This erroneous digitization also gravely impacts the quality of the NLP-tools such as NER. So as you can see, this task still needs some attention!

The software behind the NLP-tools often thrives on **Machine learning** (or **deep learning**, which is a more advanced version of ML). This means that we feed the software some examples of data annotated by humans, and the software automatically learns the patterns to extract these entities automatically. Of historical texts, this manually labeled data (or "training data") is sparse, which is also one of the main reasons why the available tools don't work so well in historical contexts. Training data is also important to evaluate these tools. Manually annotating these examples takes a lot of time and effort, but it's one of the most important steps to make the tools better for historical texts – which may uncover a wealth of untapped historical information.

So please know: the quality of your work is very important to us!

What am I annotating?

██████████████████ ██████████████████ is conducting historical research on works of **travel literature**. These are written accounts of travelers ranging from the 15th – 20th centuries in multiple languages (Normally the corpus includes English, Dutch, German and French, but you may also encounter other languages such as Latin, Spanish or Portuguese).

We want to apply NER and ABSA on this data to gain more insight in the environment and scenery of the time and the way it was perceived by the travelers and/or narrators. How did they feel about the landscapes they encountered? You'll be labeling the categories detailed in section 2:

1. **Life**: flora, mythical flora
2. **Life**: fauna, mythical fauna
3. **Environment**: weather phenomena, natural phenomena, land cover, human landforms, natural landforms, biomes
4. **People**
5. **Locations**
6. **Organisations**

Be aware that the **OCR-quality of the texts** may vary. You'll most likely encounter weird symbols and words which don't make sense. If one of the entities you come across **contains errors which are not due to its historical nature but due to the OCR-process, you can correct it.** However, when unsure, leave the entity as is.

Named Entity Recognition (NER) on travel literature

This section details **what you should (not) be annotating** and includes some real annotated examples from the corpus.

PEOPLE (PER)

Category	Example sentences	Label
Proper names ! Include titles such as Mr., Mrs, …	It is a dangerous business, but one which Mr. Roosevelt was fully nerved for by previous experiences in his extensive hunting trips […]	**PER**
Fictional person names	[…] suddenly sprang forth in full perfection, like Athene from the brain of Zeus, without our knowing of any forerunner […]	**MYTH**

Special cases to label

Category	Example sentence(s)	Label
Ethnic groups **e.g.:** tribes, denominations for ethnic groups, …	On the part of the Inuit, the Esquimaux, they were after this fashion :— " We promise that we will not steal. We promise we will bring you fresh meat. We promise we will sell or lend you dogs. We will keep you company whenever you want us, and show you where to find the game." On the part of the Kablwnah, the white men, the stipulation was of this ample equivalent:— "We promise that we will not visit you with death or sorcery, nor do you any hurt or mischief whatsoever.." "Remember that righteousness and our real ultimate self-interest demand that the blacks be treated justly." "The Masai are now great cultivators."	PER
The people of…	The American people likes to reward its heroes. […] and here we find the ancient people of Africa adopting the same methods.	PER
Half-animal, half-person **! In case of creatures which are half-animal, half-person (e.g.: centaur, faun, …), we use the MYTH label**	We saw what looked like a faun in the woods, stretching its legs.	MYTH

LOCATIONS (LOC)

Category	Example sentences	Label
Proper names Think of any location bearing a proper name: **countries, cities, states, villages, forests, bodies of water, mountain ranges, buildings, roads, bridges, …**	When you arrive before the Church of St. Mark's you realise that at last, after all your travels throughout the length and breadth of the globe, you have before you a building in which colour and design unite in forming perfection. The vessels were for the most part built upon the Merrimac river. During the whole period of our stay in the river Kawa Kawa, our crews were abundantly fed with excellent fish [...] There is no city more written about, more painted, and more misrepresented, than Venice.	**LOC**
Fictional location names	There is much in precious stones, music, sculptured figures, in pictures of heaven and hell, that appeals to these people.	**MYTH**

ORGANISATIONS (ORG)

Category	Example sentences	Label
proper names	Even when in 1888 I made my first passing acquaintance with Mombasa (before the days of the Imperial British East Africa Company), as well as other African ports [...]	**ORG**

LIFE: fauna, mythical fauna

Category	Example sentences	Label
Scientific names	I may add now, that in the course of my investigations I discovered that at least once it is found in its boring state by other authors, viz. by Ridlev & Dendv and described by them (1887 p. 224 — 229) under the name <mark>Cliona dissimilis</mark>.	**FAUNA**
Common names	Het haar geleek veel op dat van de sik van een <mark>geit</mark>. When we first rowed into this bay it was in quiet possession of herds of <mark>walruses.</mark>	**FAUNA**
Fantastical animals	<mark>Djanggi</mark> schijnt een fabelachtig dier te zijn , dat niemand juist weet te beschrijven.	**MYTH**

Special cases to label

Category	Example sentence(s)
Hypernyms (overarching animal denominations) **e.g.:** "birds" is the hypernym of "pigeon", "raven", "chicken",…	Op het dak ligt een platte mand, waarop <mark>visch</mark> te drogen wordt gelegd. As it was, this bed of snow was so steep that, had we missed our footing, we must have rolled down, and been precipitated into the sea, as invariably happened with the <mark>birds</mark> we shot, […].
Food	When the flesh of <mark>animals</mark> that have died of this disease is eaten, it causes a malignant carbuncle.

LIFE: flora, mythical flora

We label **plants** (trees, herbs, …) and **parts of plants** (vegetables, flowers, …) using the **FLORA** label. Fungi are by definition not plants, since they have different characteristics regarding their nutrient uptake and molecular composition. But since this distinction is not important in our case, we also label fungi as flora. We will include **everything that has to do with plants**.

Included: **trees, flowers, fruit, nuts, vegetables, herbs, cacti, seeds, grains, legumes, grasses**, …

Category	Example sentences	Label
Scientific names	[…] oranges, peaches, and other fruit trees, ferns, especially Gleichenia linearis, weeds of cultivation, miscellaneous shrubs and trees, including Pterocarya stenopter […] […] the Fungus Melitenjis, which is defcribed [sic] in Linneus's Amanit […]	**FLORA**
Common names	The seed of the Maidenhair tree, " Peh-k'o " [Ginkgo hiloha) , after being roasted is esteemed as a dessert nut. Rubroboletus satanas, commonly known as Satan's bolete or the Devil's bolete, is a basidiomycete fungus […]	**FLORA**
Fantastical plants	He plucked the spaghetti tree.	**MYTH**
Fruits, nuts & vegetables	Oranges thrive remarkably well in this red sandstone; and the extensive orchards are a wonderful sight during the month of December.	**FLORA**

Special cases to label

Category	Example sentence(s)
Hypernyms	[…]we found it as easy to ascend to the tops of these ==trees== as to climb the steps of a ladder […] "The ==soil== is so rich that no manure is required."
Species names (sing.: taxon, pl.: taxa)	[…] but we cannot explain why no one of the family of ==melastomas== vegetates north of the parallel of thirty degrees, or why no rose-tree belongs to the southern hemisphere.
Food	The ==mushrooms== were cooked in a broth.

! Unsure about an entity?

If you're unsure whether an entity is a plant, a fungus, an animal or neither, a good reference for disambiguation is https://www.catalogueoflife.org/ .

ENVIRONMENT: weather, natural phenomena, landforms, biome, land cover

We also want to label elements of the environment.

Category	Example sentences	Label
Weather *Descriptions of weather phenomena.* **(e.g.: sun, rain, snow, cold, chilly, hot, …)**	"When I came on deck the following morning the ==rain== fell in heavy showers." "A ==cloud== appeared to open directly over our heads, and let down the water almost in one body, but at 7.15, as the violence of the rainfall had somewhat abated, we departed from Syracuse […]" It was a ==bright== and ==sunny== day; the atmosphere being purified by a strong but refreshing ==breeze==. […] when suddenly the "heaven grew black again with the storm-cloud's frown," and a flash of ==lightning== illuminated the sky with crimson radiance.	**WEATHER**
Natural phenomena *Descriptions of dynamic natural phenomena.* **(e.g.: wave, earthquake, rainbow …)**	At our feet, full sixty feet below, lying between two walls of rock, which looked as though an ==earthquake== had rent it apart to leave space for the sparkling water […]	**NAT_PHENOMENON**
Landforms *Naturally occurring landforms.* **(e.g.: hill, mountain, river, ice, …)**	[…] covering an elliptic space whose surface had a coat of ==ice== nearly an inch thick. The principal business thoroughfare, Main Street, is in the heart of the city, and crosses the ==river== over a handsome iron bridge.	**NAT_LANDFORM**
Biomes **(e.g.: tundra, desert, tropical forest, …)**	However, there is deep, ==wet moss== or ==tundra== everywhere, as one soon discovers to his sorrow if he attempts to leave the plank walks.	**BIOME**
Human-made landforms (e.g.: building, street, …)	Genesee Street is the principal thoroughfare lined with large blocks of ==commercial houses==. […] conducted us all over the spacious ==building==, and at last, struck by a bright thought, threw	**HUM_LANDFORM**

	open the door of an ==apartment== which he said would be free in a few hours, as the gentleman occupying it was packing up his belongings preparatory to his departure […] The principal business thoroughfare, Main Street, is in the heart of the city, and crosses the river over a handsome ==iron bridge==.	
Elements of the environment (land cover?)	The fine deposits of ==sand== and ==clay== extending at different places along the shore to a distance of twenty miles inland, by their contrasts added to the scenic beauty, exhibiting a variety of magnificent views.	**LAND_COVER**

Aspect-based Sentiment Analysis on travel literature

Aspect-based sentiment analysis labels

We label the sentiment based on a scale from 1 (very negative) to 5 (very positive).

1	2	3	4	5
Very negative	**Negative**	**Neutral**	**Positive**	**Very positive**

The difference between these categories can be very vague, and prone to subjective decision-making by the annotator. Don't worry about it, that's completely normal. You can trust your intuition. A couple of ground rules:

- There is an explicit or implicit clue in the text suggesting that the speaker is positive (i.e., happy, admiring, relaxed, forgiving, etc.) or negative (i.e., sad, angry, anxious, violent, etc.) regarding the entity.
- Watch for textual clues and intensifiers e.g.: "very", "extremely", "exceptionally", "unusually", …

Example sentences per sentiment

Entities are marked in yellow. Sentiment markers regarding the entity are underlined and in bold.

1	"Externally, the hyenas have somewhat the appearance of **extremely ugly** and **unattractive**-looking dogs"	**FAUNA**
2	"I led them over his spoor, again and again, but to no purpose; the dogs seemed **quite** **stupid**, and yet they were Wolf and Boxer, my two best."	**FAUNA**
	"The rhinoceros is **bad-tempered** and **resentful** of interference."	**FAUNA**
3	"Under the microscope the shell presents a **surprising** appearance."	**FAUNA**
	"Therefore the snakes are **far less an annoyance or an impediment** to the exploration of these forests"	**FAUNA**
4	"This is the administrative center of the British Protectorate of Uganda, an **interesting** country with which we must deal in a chapter by itself."	**LOCATION**
	"[...] you see Lake Helen, a **beautiful** sheet of water, stretching away to the north for eight miles [...]"	**LOCATION**
	"[...] the **gorgeous** red-purple Dissotis flowers, a yellow composite like a malformed daisy [...]"	

5	"The <mark>roar of the lion</mark> is **extremely grand** and **striking**"	**FAUNA**

NULL-aspects

You might run into an expression of a sentiment where the entity is not explicitly mentioned, but it is clear that the sentence/group of words is **about** an entity class.

In that case, you can label the entity as "NULL_ENTITY_SENTIMENT".

Examples:

"These <mark>savages</mark> roamed the mountains."	**NULL_PER_2**
"These <mark>animals</mark> would not let us in their camps."	

In the example above, the writer refers to the native people as "savages" or "animals". The word is clearly negative, and the reader understands that the text refers to people (PER).

However, the word in itself is not an entity – and there are no other words in the sentence which explain to the reader that this word carries a negative meaning. Note that we **do not see these type of denominations as synonyms**.

"I was <mark>shaking</mark> when stepping outside"	**NULL_WEATHER_2**

"I was shaking", in its context, refers to the weather being cold. The reader gets that this is about weather through the context, but weather phenomena (snow, breeze) aren't explicitly mentioned in the text.

Irony

While rare in this type of literature, expressions may be ironic. If you come across an ironic sentence, that means the sentiment is flipped upside down. You can then **label it without considering the irony**, but then add an "irony"-label **on the sentence level** to indicate that the sentiment is actually its inverse.

Anaphoric references and synonyms

You will find that sometimes entities are referenced and referred to across sentences.

Anaphoric references

Example with reference word "it":
In cases like this, we add a "refers_to"-label to connect the **entity** and its **reference word.**
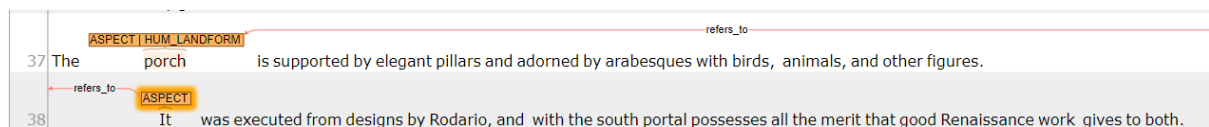
Synonyms

Sometimes, there will be several denominations referring to the same concept.

Example of synonymy:
In the example below, we see two denominations ("Chatham Island" and "Wairi Kaori") which refer to the same concept.

If you come across this, you can label both instances ("Chatham Island" and "Wairi Kaori") as entities. Use the "refers to" label to link the synonyms.

If there is a sentiment linked to the concept, you can link it using "sentiment_of" to the most used denomination throughout the text, if there is one.
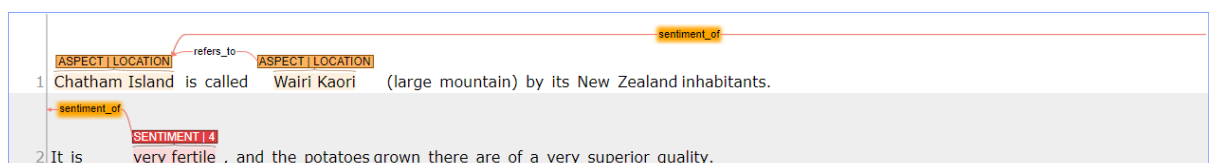
Working with InCeption

1. Logging in
2. Selecting a text
3. Annotation of entities and sentiment
4. Annotation of coreferences
5. Annotation of irony

Logging in

We are annotating in the software platform InCeption on the server ~~of LTS (the Language and~~ ~~Translation Technology Team of UGent~~).

The link to the server InCeption instance: ~~https://inception.lt-ugent.be/login.html~~ . You will see this screen:

INCEpTION

Welcome!

User ID

Password

SIGN IN

Log in with the **login credentials** you received from ~~------~~

Selecting a text

Next, this screen will pop up. The project is called **"Travel_literature"**. Click on "**Annotation**" to go to your annotation interface.
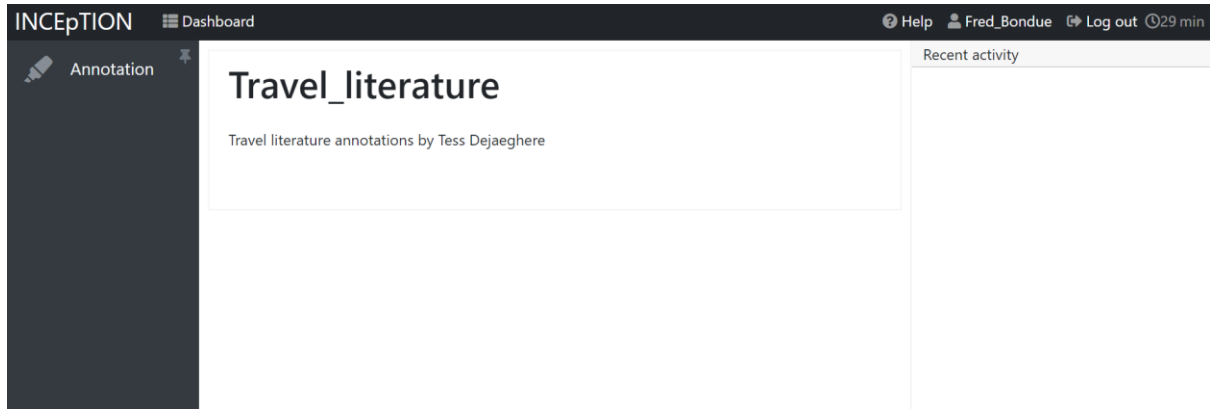
Choose a text to annotate and click "**Open**".
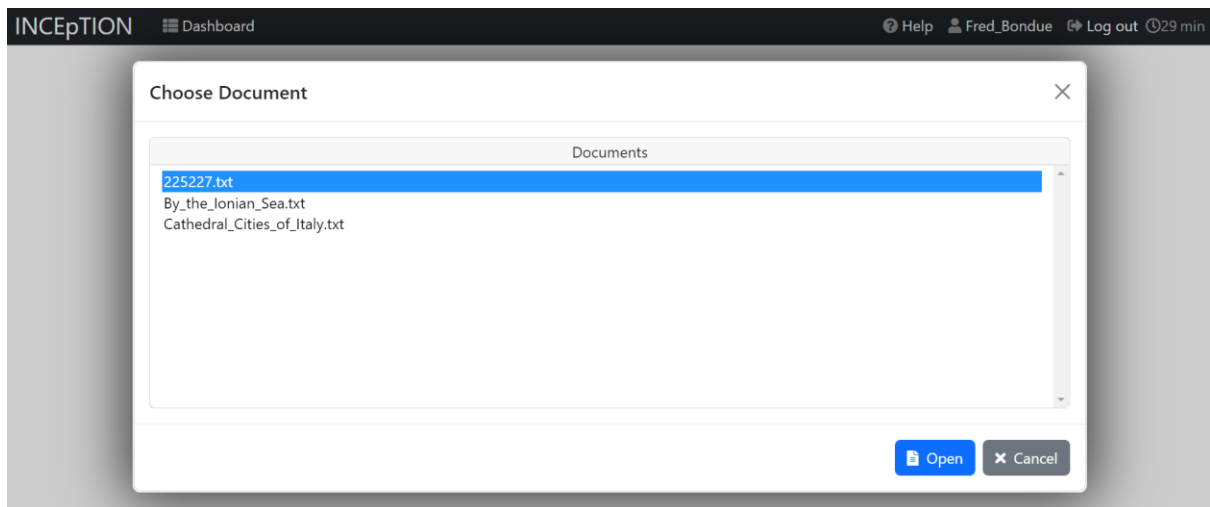
Annotation of entities and sentiment

The annotation interface shows you the text divided in several blocks of sentences. Make sure you have **Emotions_Entities** selected as a layer.
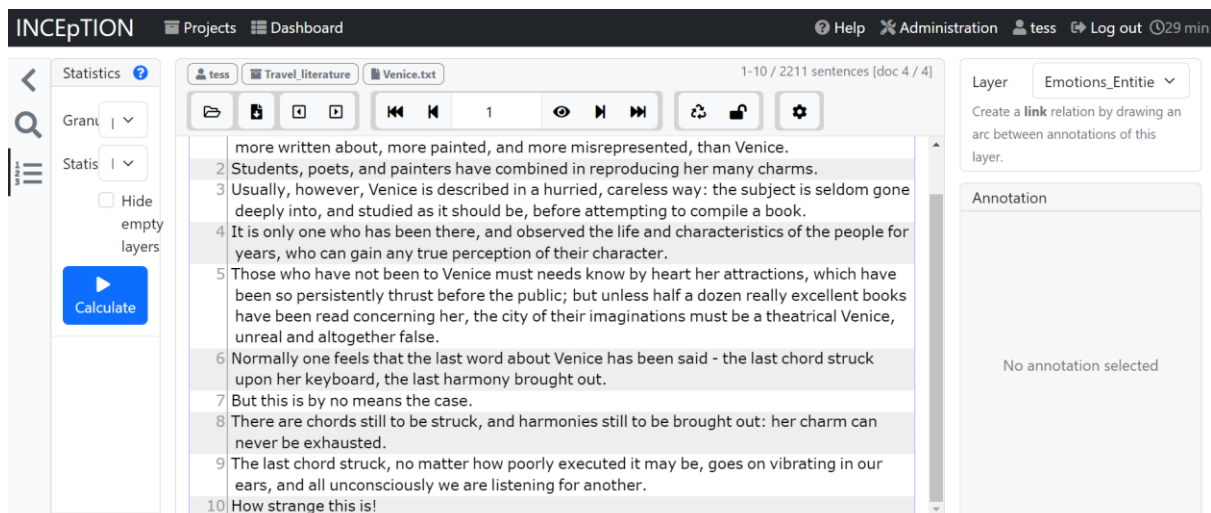
Upon selecting a word, a decision panel will pop un on the right, like below. Here, we selected a **location**, namely *Venice*.
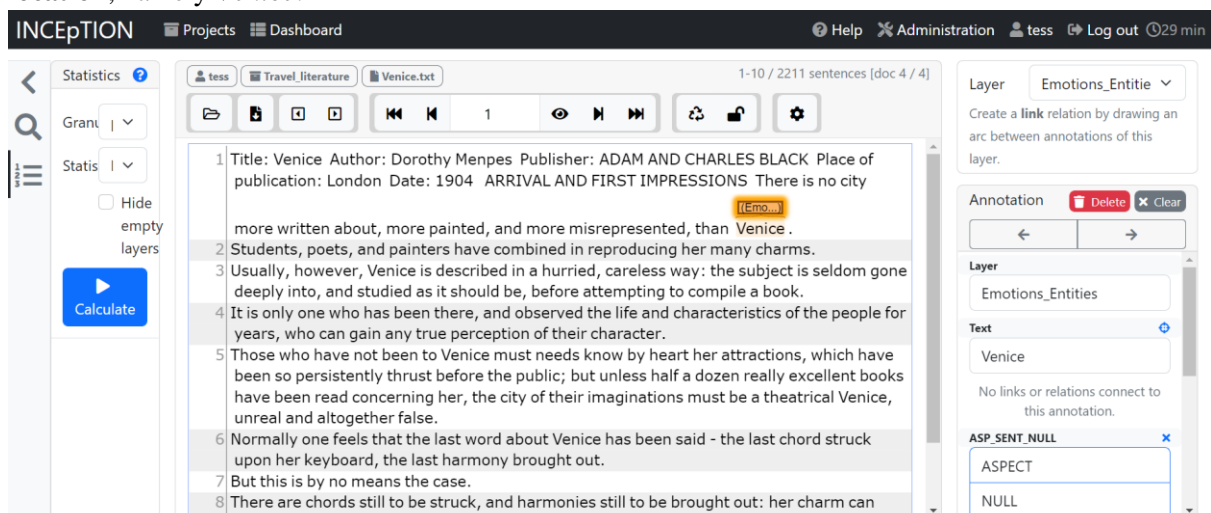
When labelling, we have to go through the following decision process:

1) Choose whether the word is an **ASPECT**, a **SENTIMENT** or **NULL**. This step is required, so you'll get an error message if you skip it.

In the case of an ASPECT or NULL:
1) Choose an "**entity category**" in the dropdown menu as shown below.

In the case of a **SENTIMENT linked to an aspect**:
1) Choose a "**sentiment category**" (1-5).
2) Link it to the entity the sentiment using "**link type**" "**sentiment_of**".

For each sentence, **also label the sentiment on the sentence-level**.

1) Choose the layer "**sentiment_sentence**".
2) Choose the adequate sentiment for the entire sentence in "sentiment_sentence".

Coreferences

1) Label the reference word (it/her/…) as an aspect or a sentiment, depending on what it refers to. (e.g.: *Venice* is an aspect, so *her* is also an aspect.).
2) Select the word you want to link to another word.
3) Drag the arrow from the reference word to the aspect/sentiment.
4) Use the "**refers_to**" label.

Annotation of irony

The sentence in the example is not ironic, since I couldn't find an example of one.

You don't have to follow these steps for every sentence.

1) Select the sentence.
2) Choose the layer **"irony_sentence".**
3) Label the sentence as "**ironic**".

Your annotation work is saved automatically on the server, and you can pick it back up anytime. Annotations can always be stacked. In other words, you can label a word using multiple labels.

Checklist for each text

- ☐ Upon opening a new text, check which text you are annotating in the **metadata folder**. The **name of the text** is the ID used in the **metadata Excel file.** Tess will send you the link to this file when it is ready.

- ☐ You are annotating excerpts of books, so it might not always be super clear what the text is about, or what the sentiments are. **Read a bit about the context of the text.** What is the setting? Who is the traveler? Don't overdo this, you don't need to become an expert.

- ☐ When in doubt about a part of the text, drop it in the shared troubleshooting Excel sheet. We'll go over them together during the next annotation lab. If it's urgent, please send me an e-mail.