

Decide on the focus and extent of annotation

Any semantic annotation, as comprehensive as it might be, always has a **purpose**, that is, is connected to a set of research questions currently pursued or considered as relevant in future work. Therefore, at the early stages of annotation, you will need to **make choices** about the **focus and extent** of your full-text annotation:

1. What entity types are relevant? Are you interested in capturing people? Places? Dates? Concepts?
2. Do you need only to annotate entities, or you need to put them in clauses, statements?

1. What entities?

A first entity that needs to be put in the text is obviously the **Territory you are annotating**; and, if any subdivisions (into chapters in a book, individual documents within a notarial register etc.) can be relevant (usually they can), also the **Subterritories**, i.e. Territory entities which are children of the first-level Territory.

If you do only annotation without CASTEMO Statements, it is usually a good idea in most projects to annotate **people (Persons) and places (Locations)**. These are the obvious candidates to include. However, if you are doing CASTEMO and anchoring Statements in full-text documents, you might not be interested in doing the further work of also delimiting the text spans for people and places.

Generally, if you want to understand **meaning**, you will not be able to miss **Concepts** from annotation.

If you are making CASTEMO Statements, there are several reasons why to also anchor them in full-text:

1. **Up-to-date text span**, which will get updated even when you change the text (e.g. correct errors).
2. **Use as training data for generative LLMs**, where a workflow can use your existing annotation to produce new, machine-generated annotation, check for inconsistencies etc.

2. Only collecting entities, or building clauses?

If your annotation has a clear focus, such as "this text is only on Hussites and I only annotate hate speech concepts related to them", then you might be fine with only annotating the Concepts used

to denigrate them, without doing anything else. However, if you require more analytical opportunities, you will generally need to distinguish **to whom such Concepts are applied, and in what way**. If this is the case, you will generally need to use **Statements**. Statements are the way to distinguish e.g. "Hussites - are" from "Hussites - are similar to", and also to record, if needed, reported speech and other devices constituting narrative perspective.

Thus, shortly: **if Concepts relate to more entities than one, you need Statements**, and thus, a CASTEMO approach, actually modelling clauses as statements (whether summarized - SumCASTEMO, or selective - SelCASTEMO, or full sequential - FullCASTEMO).

Revision #4

Created 12 July 2025 11:13:45 by David Zbírál

Updated 12 July 2025 13:20:02 by David Zbírál