

Entities

This chapter describes the different entity types of the CASTEMO data model, and their recommended usage.

- [Actions](#)
- [Concepts](#)
- [Attributes of entities](#)
- [Statements](#)

Actions

Actions (or more fully, Action types) represent **individual semantically disambiguated verbs**. They are **lemma-meaning units**, i.e. one meaning of a specific lemma corresponds to one Action. Thus, we can have many Actions labelled "to see": one for "be able of sight, not blind", one for "perceive with sight", one for "go to meet somebody".

Three valencies: Entity type valency, grammatical valency, and semantic valency

Introduction

Actions acquire three kinds of valencies for any actant slot (subject, object 1, object 2; the data model is potentially extensible further, beyond trivalent verbs):

1. **entity type valency**, which defines which entity type (Person, Concept, etc.) is allowed in the given actant slot;
2. **morphosyntactic valency**, which is a free text field defining the **prepositions and grammatical cases**, but uses a formalized notation (grammatical cases are noted with numbers, prepositions are in quote marks "", alternative is marked with a pipe "|"); and
3. **semantic valency**, i.e. what kind of role the entity occupying the given actant slot has by implication (e.g., the subject of the Action "to travel" would have the semantic valency C "traveller").

The main benefits from valencies are that they:

1. guide users in their **choice of the correct Action** (or creating a new one if none among the existing fits the meaning and syntactic structure);
2. help users with **validity of data in actant slots**;
3. allow InkVisitor to deploy **data validation features**;
4. facilitate **machine understanding of text**, allowing semantic disambiguation of verbs based on their morphosyntactic valency (recognized by dependency parsing), and optionally, entity type valency (recognized e.g. through named entity recognition).

Morphosyntactic valency notation for Latin

In the field marking morphosyntactic valency, we use the following **abbreviations and signs**:

- **Numbers 1-6**: cases. E.g. "1" means nominative, "6" means ablative.
- **Pipe sign ("|")**: denotes the logical "OR", i.e. marks alternative morphosyntactic valencies.
- **Plus sign ("+")**: denotes concatenation, e.g. "de" + 6 means: "with preposition *de* and ablative case".

- **Words in quote marks ""**: denote the actual words used in this valency, e.g. prepositions in this valency.
- **inf**: infinitive.
- **4inf**: accusative with infinitive.

E.g., `4 | 4inf | "quod"` means that in this actant slot, this verb can only take either an accusative, or a sentence rendered as accusative with infinitive, or a clause starting with "quod".

Recommended standards for a finalized (**approved**) action

Before assigning an Action the **approved** status, it should meet the following standards:

- Its **meaning is described** in the **detail** field. (You will benefit from the use of printed or online dictionaries or LLMs.)
- It has the **Action/Event Equivalent relation filled in** with a Concept which has its meaning defined in its own "detail" field.
- It has **full information on the three valencies** for each actant slot (including the explicit declaration of **empty** in the entity type valency, if no entity is allowed in that slot).
- It has a **reference to an external lemma collection ID** (in DISSINET, we use the LiLa Lemma Collection).
- If you **have found a corresponding meaning among WordNet synsets**:
 - It has a **Reference to the corresponding WordNet synset**.
 - Its definition in the **detail** field **takes the WordNet definition into account**.
- If you **haven't found a corresponding meaning among WordNet synsets**:
 - You have **defined the meaning** yourself or based on dictionaries.
 - If there is any **synset in WordNet which is a superclass** of this (more specific) meaning, then **an Action corresponding to the WordNet meaning is created** (if Latin WordNet has it, then in Latin; if not, then in English), **described, has a Reference to the WordNet synset, and it forms the Superclass** of this more specific Action you are working on.
- There is **no remaining error message from InkVisitor validation**.
- All of this has been **checked**, i.e. it is not just a first draft of the Action that you still plan to come back to.

For something to be aligned with a synset definition in WordNet, it is *not* required that you accept its hypernyms or synonyms, just the definition needs to match.

Recommended linkage to external lemma and meaning banks

- **Link each Action through a Reference to at least one external lemma bank.** A major lemma bank is still **WordNet for the given language**. For Latin, we use the [LiLa Lemma Collection](#) in DISSINET.
- **Link each Action through a Reference to at least one external bank of meanings/senses.** A major meaning bank is still **WordNet synsets**.

- Such linkages are important for the **interoperability of your data**, and giving it meaning curated by bigger projects and infrastructures.
- **Reference** is a pair composed of a Resource entity representing the given resource, and its part (typically unique identifier).
- If the lemma or meaning is **not found** in the lemma or meaning bank you are normally using, it is useful to know that you checked: in such a case, **add the Reference, but put Value "NA" as the Reference part**.
- Also **DISSINET Database (DDB) and MedHate database (MDB) are providers of identifiers (UUIDs) that you can link to**. If upon its creation you request to have your CASTEMO database pre-populated with some entities from another deploy of InkVisitor, the UUIDs will be the same and the references will already be there.

Concepts

Concepts represent, alongside Action types, another **generic entity type**, which holds the data semantically together. Concepts are **lemma-meaning units**. That is, for **any distinct lemma and meaning, you create a new Concept**. This proliferation of Concepts is made manageable through various [Relations](#), such as **synonym, superclass** (~ hypernym, genus proximum) etc., which place any Concept in a robust web of semantic relations.

One major function of Concepts is to serve as **Property Type** in [Properties](#), which is a flexible yet reliable way of creating connections between entities.

CASTEMO knowledge graphs are **multi-lingual**; thus, Concepts from different languages can coexist, but are defined by semantic relations.

While CASTEMO knowledge graphs are multi-lingual, we recommend choosing **one analytical language** for higher levels of the conceptual taxonomies.

Attributes of entities

Any entity type has some **internal Attributes**, which allow to characterize the entity. The InkVisitor interface guides users as to what attributes are expected for a given entity type.

A first and obvious Attribute is **label**, that is, the name of the entity. The name can change; it is the identifier (UUID) rather than the label which holds the semantic identity. However, **label changes** need to be done with consideration, as they influence the uses of this entity in already existing data.

The **detail** Attribute should be used to **define** the entity. This is especially important for Actions and Concepts, which keep the data semantically together.

Another attribute used in all entities is **label language**, which defines in what language the label is written.

Languages not available in the interface (including old development phases of modern languages) can be added upon request.

To further extend the possibilities of finding an entity rather than create duplicates e.g. for mere **orthographic variants** (e.g. democratisation vs. democratization), there are **alternative labels**, which means other labels than the main one, highlighted in label.

Alternative labels should *not* be used for labels in another language (create the entity in that language and link it with a SYN Relation to this one instead).

In Concepts and Actions, alternative labels should *not* be used for different lemmas.

Technically, the main label and alternative labels are one single field; the first position in this field is held by the main label. This is how the application ensures non-duplication between main label and alternative labels.

Apart from such internal Attributes, any entity can enter in **three types of connections to different entities**: Properties, Relations, and References.

Statements

Structure and purpose

Statements model the syntactic structure and semantics of clauses. They have a **quadruple structure** with **action slot** and **three actant slots: subject, actant1, and actant2**. The semantic core is the **action** slot, which holds the **predicate** of the clause, and is linked to actants as defined by the **syntactic valency** of the Action type used, up to trivalent verbs (for instance, 'Peter received a gift from Elisabeth'). Any Action Type can define through its **syntactic valency** that an actant is required, optional, or forbidden (required empty).

While InkVisitor can be used for classical entity-relationship modelling done in any database, the CASTEMO data collection workflow is specific in its focus on **modelling textual clauses** through statements. In this sense, it is a **statement-based data collection workflow**. This allows texts to be comprehensively modelled, either selectively (based on what is relevant for specific research) or in entirety. CASTEMO is intended to **keep the order of appearance and contextual embeddedness of information** in the text.

Action attributes

Actant attributes

Pseudo-actant