Annotation Scheme for Morpheme Annotation based on Semantürk

PART 1 Guide and Methodology

Content Table

Introduction

This annotation guide has been developed as part of the doctoral research on morphosemantics of Turkish nominal derivation[1]. The thesis aims to model the semantic structure of Turkish nominal derivation by linking morphological and semantic levels of representation. Within this framework, an ontology of derivational meaning called *Semantürk* has been designed. *Semantürk* is an adaptation and extension of the *Démonette-2* ontology, initially developed for French derivational nouns (Huguin et al., 2022, 2023; Namer et al., 2023), to a finer linguistic scale, adapting it to the annotation of morphological units smaller than words.

The top-level semantic categories used in this guide originate from the *Demonette-2* annotation guide (Huguin et al., 2022). They have been translated into English and, where necessary, adapted to the derivational annotation of Turkish morphemes.

This document (Part 1) presents the methodological component of the guide. It outlines the annotation workflow, principles, and conventions designed to ensure

[1] The guide is written in English to support international collaboration and to ensure compatibility with multilingual projects based on Démonette and related frameworks.

coherence and reproducibility in the semantic annotation of Turkish nominal morphemes.

The descriptive component, which provides detailed definitions of the semantic categories, examples, and disambiguation guidelines, is presented separately in a dedicated document entitled *Part 2: Semantic Categories*.

Semantic annotation at the level of derivational morphemes raises both theoretical and practical challenges. Like lexical units, morphemes often display polysemy and context-dependence, which require flexible yet principled annotation strategies. Moreover, the interlinguistic adaptation of an ontology initially designed for French to a typologically distinct language such as Turkish involves reconsidering the scope and hierarchy of semantic categories.

This work therefore stands at the intersection of linguistic modelling and resource development. It aims not only to produce an annotated dataset for Turkish but also to provide insights into how ontological representations of meaning can be adapted across languages and morphological systems.

1. Introduction to Semantürk Description of the ontology

The *Semantürk* ontology has been developed as a semantic adaptation of Démonette-2 (Huguin et al., 2022; Namer et al., 2023) to model derivational meanings in Turkish. Its structure follows the overall architecture of Démonette-2[2], which provides a linguistically grounded and hierarchically organised ontology for nominal annotation in French.

The base hierarchy of simple categories in Démonette-2 has been largely preserved in Semantürk. However, additional categories (both simple and complex) have been introduced to account for semantic distinctions required by Turkish derivational morphology, particularly when suffixes encode meanings not directly represented in the French system. In addition, the complex categories have undergone a major conceptual redefinition: we reframed those around composition rather than polysemy as explained in the following subsections.

Semantürk thus combines cross-linguistic continuity with language-specific adaptation, making it suitable for both descriptive and comparative morphosemantic analyses.

1.1. Hierarchy

The annotation scheme used in this guide is based on the semantic hierarchy defined in the *Semantürk* ontology. This hierarchy comprises:

- 62 simple categories, representing core conceptual domains; among which 30 from *Démonette-2.*

[2] *https://www.demonext.xyz/wp-content/uploads/2022/08/Guide_Demonext-semantique.pdf*

- 29 complex categories, formed by combining two simple categories through specific semantic operators.

Each category and subcategory is precisely defined in the ontology, often accompanied by comments and examples in order to facilitate interpretation.

At the top level, Semantürk maintains two broad domains inherited from Démonette-2:

1. ENTITY, which groups concrete and abstract nouns referring to objects, materials, places, persons, or conceptual entities;
2. SITUATION, which groups eventive and process-related nouns such as actions and states.

Both domains are hierarchically subdivided into finer categories, allowing the ontology to capture derivational nuances at different levels of abstraction.

1.2. Granularity and Ontological Design

The choice of an ontological framework is motivated by several key properties.

- Extensibility:

The ontology is designed to be expanded with new categories or relations as linguistic analysis progresses in future projects. This makes it adaptable to language-specific needs while maintaining internal coherence.

- Hierarchical transparency:

The hierarchical structure is explicitly visible, enabling annotators to understand how a specific category relates to higher-level conceptual classes. This transparency contributes directly to the consistency and interpretability of morpheme-level annotation.

- Controlled granularity:

Superclasses in Semantürk are not defined as the simple sum of their subclasses; rather, they represent broader conceptual groupings. This non-exhaustive structure allows annotation to occur at different levels of precision, depending on the clarity of the data or the needs of the analysis.

For example, when a morpheme's meaning does not match with any specific categories at the end of an ontological branch, annotators may select a higher-level category (e.g. AGENT, ARTIFACT), whereas clear and specific cases can be annotated with finer subcategories (e.g. HABIT_PERSON, SHAPE). In other words, when the morpheme's contribution is underspecified in the available evidence, the annotator should select the closest parent rather than forcing a fine-grained choice.

Such flexibility in granularity ensures that *Semantürk* remains both linguistically accurate and operationally efficient for morphological annotation tasks.

1.3. Simple Classes and Complex Classes

In the annotation framework, a simple category corresponds to a single semantic class capable of representing the meaning of a morpheme on its own. For instance, the

suffix *-lİk* in *güzel-lik* (“beauty”) as a noun can be directly annotated with the simple category COGNITION.

However, derivational morphemes in Turkish often display a certain degree of polysemy, as a single suffix can take on multiple meanings depending on its base and context (Lieber, 2004). When the meaning contributed by the morpheme is not entirely dependent on the semantics of the base, the suffix may correspond to several possible simple categories rather than a unique semantic category. For example, the suffix *-lİk* in *sarı-lık* (from the base *sarı* “yellow”) can yield distinct interpretations:

- *sarı-lık* meaning “yellowness”
- *sarı-lık* meaning “jaundice”

In such cases, the morpheme is annotated with multiple possible simple categories, separated by a semicolon (COGNITION; DISEASE), indicating that the suffix exhibits lexical polysemy across derivations rather than within a single composite meaning.

Unlike *Démonette-2*, where complex categories were primarily motivated by lexical polysemy (tested through co-predication at the word level), *Semantürk* defines them in terms of compositionality at the morpheme level. In other words, the goal is not to capture distinct lexical senses but to model how different semantic values are combined within a single derivational process.

Accordingly, *Semantürk* distinguishes three types of complex categories, each corresponding to a specific type of semantic composition and represented with different operators. These complex categories are defined within the categories they combine:

(a) 10 hybrid categories (operator ‘+’)

(b) 4 fused categories (operator ‘*’)

(c) 15 distributive categories (operator ‘x’)

(a) Hybrid categories (+)

Hybrid categories are created using the + operator. They represent additive compositionality, where two distinct semantic values are expressed jointly and transparently within a single morpheme. Each component retains its individual interpretability, but both are activated together in the derivation.

For example, the suffix *-Cİ* in *ekmek-çi* (“bread seller, maker / bakery owner”) combines the meanings AGENT and POSSESSOR, yielding the hybrid category AGENT+POSSESSOR. Similarly, *-DAŞ* in *meslek-taş* (“colleague”) expresses both PERSON and FELLOWSHIP yielding the hybrid category PERSON+FELLOWSHIP. Such combinations result in cumulative meaning, where the morpheme encodes more than one semantic value simultaneously.

Hybrid categories correspond to cumulative compositional values, meaning that both semantic components are clearly expressed and can be paraphrased separately as well as jointly. This cumulative configuration may occasionally give rise to controlled polysemy,

since one or both values may be contextually activated. For instance, in AGENT+POSSESSOR, the suffix *-Cİ* may express only the agentive value (*one who sells or makes X*), only the possessive value (*one who owns or runs X*), or both simultaneously. However, this polysemy is not systematic. In contrast, in PERSON+FELLOWSHIP, as observed with the suffix *-DAş* in *meslek-taş* ("colleague"), both semantic values are consistently and jointly expressed.

Thus, hybrid categories describe compositional meanings that combine distinct semantic values, sometimes allowing for variable activation of each value, yet always within a coherent and predictable derivational pattern.

(b) Fused categories (*)

Fused categories employ the * operator. They represent integrative compositionality, in which two semantic values are inseparably blended into a single meaning that cannot be decomposed or paraphrased independently. In *Semantürk*, fused categories are limited to combinations where the first component is a subclass of GRADABLE_QUANTITY (DIMINUTION in *Semantürk*).

For instance, DIMINUTION*ARTIFACT denotes "a small object of type X" (*tarih-çe* "short booklet on history"), a meaning where the scalar and the nominal components are jointly realised and conceptually inseparable. Unlike hybrid categories, fused ones do not encode two parallel facets, but a single blended semantic unit emerging from the fusion of both components.

(c) Distributive categories (x)

Distributive categories use the x operator. They represent relational compositionality, in which a semantic value (restricted to combinations with GROUP or PART in the current version) is distributed across another entity type. For instance:

- GROUPxPLANT can be glossed as group of plants;
- PARTxBODY can be paraphrased as part of body.

These categories capture the structural relations that emerge when the morpheme expresses a collective or partitive relationship between entities.

1.4. Description of categories

Each class described in this guide follows a consistent descriptive format designed to facilitate navigation within the ontology and ensure transparency between categories inherited from Démonette-2 and those newly introduced in Semantürk.

The same structure is applied to both simple and complex categories, with minor adjustments depending on the type of category.

Path

Indicates the position of the category within the ontological hierarchy, providing contextual information about its relationship to parent categories.

For complex categories, the path is provided for both component categories to show how they combine within the hierarchy.

Definition

Provides a clear and concise description of the semantic class.

Definitions may consist of a translation or reformulation of the *Démonette-2* definition when applicable, or a new or refined definition when the *Démonette-2* version was absent or insufficiently explicit (particularly for higher-level or newly introduced categories). *Definitions are provided only for simple and hybrid categories.*

Source

Indicates whether the category originates from Démonette-2 or whether it was newly created or adapted specifically for Semantürk.

Examples

Examples are given in alphabetical order, without implying frequency or representativeness.

They include French examples (fr) taken from the *Démonette-2* annotation guide, and Turkish examples (tr), where the morpheme is visually indicated in bold, e.g. dava-**lı**.

The examples are illustrative only; the list is not exhaustive. Their purpose is to clarify the semantic category and to show how Turkish derivational morphology can encode particular meanings through suffixation.

Notes

Provide additional comments or clarifications to help delimit the scope of the category, especially by contrast or negation with related categories.

Complex categories or Categories

Complex categories are provided only for simple categories. They list the complex categories in which the described simple category appears.

Categories are provided only for complex categories. They list the component simple categories that combine to form the complex category.

2. Annotation

2.1. Corpus

You are provided with a corpus of 100 derived nouns, each accompanied by its definition and morphological process. Your task is to annotate the morphemes using the predefined semantic categories provided in the annotation guide. As shown in Table 1 each entry includes:

- The morpheme (suffix or prefix)
- The base word
- The derived noun (word in context)

- The definition of the base word
- The definition of the derived noun
- The English gloss or translation

Table 1: Example of an entry of the corpus

Morpheme	Base noun	Derived Noun	Base definition	Derivative Definition	English gloss
-Cİ	balık	balıkçı	Omurgalılardan tatlı ve tuzlu sularda yaşayan, […] yumurta ile çoğalan kemikli hayvanların genel adı.	1. Balık tutan kimse. 2. Balık satıcısı. 3. Balıkla ilgilenen.	fisher

Annotators are required to fill in the fields Semantic category, Confidence, and Comments. The Semantic category cell should contain the most appropriate semantic label(s) from the ontology. The Confidence cell is to grade the choice of the chosen semantic category (from lowest 1 to surest 3). The Comments field may be used to clarify the reasoning behind the choice, mention uncertainty, or note any contextual difficulty as illustrated in Table 2 below.

Table 2: Example of annotation for *balıkçı*

Semantic category	Confidence (1-3)	Comments
Agent; Habit_Person	3	Can be both, depends on the context.

2.2. Rules

Remember that the annotation concerns the suffix, not the whole word.

Your task is to identify the semantic contribution of the suffix to the derived noun, in other words, the meaning added to the base:

$$Meaning(derived\ noun) = Meaning(base) + \boldsymbol{Meaning(suffix)}$$

For instance, the analysis of the suffix in the derivative *içki-ci* ("alcoholic") could yield:

$$AFFECTED\_PERSON(içkici) = SUBSTANCE(içki) + \boldsymbol{DEPENDENT\_PERSON(-ci)}$$

The definitions provided for the base and the derived noun are only interpretative clues to help infer the contribution of the suffix. The goal is not to assign a semantic category to the entire derivative, but solely to the morpheme's contribution only.

- **Granularity**

Always select the most specific semantic label available in the ontology, that is, the lowest-level category that matches the morpheme's meaning.

If there is no match among the lowest categories, it is acceptable to move up to a more general (parent) category rather than forcing artificial precision. Table 3 summarises the decision rules to understand at what level to annotate the morpheme.

Table 3: Decision rules for granularity selection

Situation	Decision
The morpheme's meaning clearly matches a fine-grained class	Choose the lowest-level (most specific) category.
None of the lower categories apply	Use the closest higher category that captures the general meaning.

- **Polysemy and compositionality**

A single morpheme may match different semantic categories depending on the base to which it is attached. Thus, a suffix such as *-lİk*, cannot be systematically annotated with the same category througout the corpus. Each occurrence must be evaluated independently, based on its base-derived pair.

When several categories seem to fit a morpheme in the same derived noun, a deeper semantic analysis is required. The potential categories must be evaluated according to whether the meanings are compatible or incompatible.

- Incompatible meanings

Hints at polyfunctional aspects of the morphemes, which means that it is polysemic. The meanings correspond to distinct, mutually exclusive derivational interpretations. Only one can be true for the given base+suffixed form.

Annotation rule: List simple categories alphabetically, separated by semicolons

Example: COGNITION; DISEASE for *-lİk* in *sarı-lık* ("yellowness" vs "jaundice").

- Compatible meanings

The morpheme contributes two semantic facets that co-occur within a signle derivational meaning. These facets must then be encoded with a complex category, following the compositional logic defined in *Semantürk*, see Section 1.3.

Final instructions

- Do not create new categories, including complex ones. If you feel that an existing label does not fit, list several possible categories separated by a semicolon (;) and explain your reasoning in *Comments*.
- If no category applies, leave the *Semantic category* cell empty and justify your decision in *Comments*.

By following this principle, you will help test whether the semantic categories of the ontology truly capture the meanings expressed by Turkish nominal morphemes.

2.3. Expected Format

When completing the annotation file, please follow the formatting conventions below carefully. Consistency in format and spelling is essential for accurate processing and comparison of annotations.

Column: Semantic Category

- Always use the exact labels provided in the ontology. Respect the case (uppercase/lowercase) and spelling of category names.
- When indicating multiple possible categories separate them with a semicolon (;) (e.g. AGENT; HABIT_PERSON) in alphabetical order. Do not use commas or slashes.
- For complex categories, make sure to use the correct operators: + for hybrid categories (e.g. AGENT+POSSESSOR), x for distributive categories (e.g. GROUPxPLANT), * for fused categories (e.g. DIMINUTION*ARTIFACT).
- Double-check spelling and operator placement before submitting.

Column: Confidence

Fill this column with a single number between 1 and 3, representing your level of confidence: 1 = uncertain or hesitant; 2 = fairly sure but with minor doubt; 3 = completely sure of the chosen category. Use only digits (1, 2, or 3), not words.

Column: Comments

You may write any relevant remarks that help interpret your decision, such as reasons for hesitation or ambiguity, interpretation difficulties, suggestions for missing or unclear categories, questions about the guidelines.

Writing in English is preferred, but feel free to use another language if it helps you express your thoughts more clearly.

File Naming and Submission

Before submitting your annotated file, please:

- Add your first name to the filename (e.g. Annotation_Semanturk_Name.xlsx).
- Keep the same format and structure as the received file.
- If you encounter technical problems, you may use a different format temporarily, but please inform the coordinator.

Bibliography

Huguin, M., Barque, L., Haas, P., Namer, F., & Tribout, D. (2022). *Guide d'annotation Demonext*. https://hal.science/hal-03638962

Huguin, M., Barque, L., Haas, P., & Tribout, D. (2023). Typage sémantique des noms dans la ressource morphologique Démonette. *Lexique*, *33*, 41–56. https://doi.org/10.54563/lexique.1086

Lieber, R. (2004). *Morphology and Lexical Semantics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511486296

Namer, F., Hathout, N., Amiot, D., Barque, L., Bonami, O., Boyé, G., Calderone, B., Cattini, J., Dal, G., Delaporte, A., Duboisdindien, G., Falaise, A., Grabar, N., Pauline, H., Henry, F., Huguin, M., Juniarta, N., Liégeois, L., Lignon, S., & Tribout, D. (2023). Démonette-2, a derivational database for French with broad lexical coverage and fine-grained morphological descriptions. *Lexique*, *33*, 6–40. https://doi.org/10.54563/lexique.1242