

# Manual for the Semantic Annotation of Noun Classes with ‘Split Stimulus’ PPs

Johanna M. Poppek, Simon Masloch

June 2022

## 1 Introduction

This manual was used to annotate semantic classes for the internal argument of the preposition and the subject with so-called ‘split stimuli’ in German (on a subset of the GerEO resource, (Poppek et al., 2022)). ‘Split stimulus’ refers to a phenomenon where with a psych-verb, the semantic stimulus of the experiencer’s emotion appears to be ‘split up’ between an NP and another phrase (usually a PP) (see Engelberg, 2015).

- (1) Ein Hosenanzug beispielsweise, Streifen und Karos gemixt, imponiert mit  
a pantsuit for.example, stripes and checks mixed, impresses with  
seiner klaren Linienführung [...]  
its clear lines  
‘For example, a pantsuit where stripes and checks are mixed, impresses with  
its clear lines [...]

For more details, please see our paper, (Poppek et al., in preparation), where we describe the annotation procedure and report the results. The annotated data is available via [https://github.com/Linguistic-Data-Science-Lab/German\\_EO\\_verbs/tree/main/split\\_stimuli](https://github.com/Linguistic-Data-Science-Lab/German_EO_verbs/tree/main/split_stimuli).

In order to investigate differences in the semantic constraints the relevant prepositions impose on their internal and external argument, we annotated all<sup>1</sup> sentences from GerEO (Poppek et al., 2022) in which the verb occurred in either the transitive or the intransitive (object-drop) pattern with the semantic classes of these arguments. These patterns were chosen because they are the ones in which the stimulus is the subject. We used an adapted version of the noun classes used by the CISLEX project (Langer, 2009) since it provides us with a rather large-grained semantic classification for the nominal domain. The sample does not contain occurrences with a PP that we regarded as clearly instrumental (this mostly applies to examples containing *mit* ‘with’ and *durch* ‘through’) since these are also not annotated as stimulus PPs in the original GerEO database. In total, about 780 sentences received the additional annotations.

---

<sup>1</sup>One sentence containing *von* ‘by’ was excluded. In it, the verb was embedded under causative *lassen* ‘let’. This is definitely a totally different construction. Another sentence was excluded because of a sentence tokenisation error that did not affect the original GerEO annotation, but made annotation here impossible.

Of course, the sentence will not necessarily contain only one stimulus PP (cf. (2)), and the internal argument of the preposition may contain some kind of coordination, in which case the semantic class of more than one noun has to be annotated (as in (3)). Additionally, even the same noun may have more than one plausible reading (in (4), the decision may be an event/action or (more likely here) the abstract result of it).

- (2) Die gemischten Meerfische vom Grill [...] fielen weniger durch ihren  
the mixed sea.fish from.the grill struck less through their  
Geschmack auf als dadurch, dass sie lauwarm serviert wurden [...] taste  
PTKV than that.through that they lukewarm served were  
‘The grilled mixed sea fish [...] stood out less for their taste than for the fact that they were served lukewarm [...]’
- (3) [...] der knapp 18jährige Seoane verblüffte lange mit stupender Zuspiel-  
the barely 18.year.old Seoane amazed long with amazing passing  
und Ballsicherheit [...] and ball.security  
‘[...] barely 18-year-old Seoane amazed [everyone] for a long time with his amazing passing and ball security [...]’
- (4) [...] in der Schiedsrichter und [...] Linienrichter lange Zeit mit  
in which referee and assistant.referee long time with  
unverständlichen Entscheiden [...] die Spieler beider Mannschaften  
incomprehensible decisions the players both.GEN teams.GEN  
verärgerten.  
annoyed  
‘in which the referee and [...] the assistant referees annoyed the players of both teams for a long time with incomprehensible decisions [...]’

## 2 Annotation process

The GerEO-sample was annotated by two native speakers of German with extensive knowledge of the project. It was evenly split among the annotators in a first stage and annotated following the present manual. Since we did not consider calculating a classic inter-annotator-agreement sensible for a dataset with many (overlapping or similar) categories, we decided on an adjudication stage instead, where every example was double-checked by the second annotator. In case of controversial annotations, the final annotation was agreed upon during the discussion.

## 3 Columns

In addition to the columns inherited from GerEO, we use the following ones:

***sem\_class\_subj*** The semantic class of the subject

***sem\_class\_PP\_int*** The semantic class of the head noun of the internal argument of the PP (in case it's a coordination: annotations for each head noun separated by slashes)

***interpr\_sem\_class\_PP\_int*** If the internal argument of the PP receives a reading different from the one annotated in *sem\_class\_PP\_int* in the context of the sentence, annotate it here. If not, leave blank.

***det\_PP\_int*** Is the determiner of the PP's internal argument somehow special? Annotate *no\_det* if the determiner is left out (not to be used with mass nouns, proper names, plural NPs etc., where no determiner would be expected anyway), *det\_coref\_subj* if the determiner is a possessive pronoun coreferent with the subject. Otherwise leave empty.

***Comment\_PP*** Personal comments of the annotators

## 4 General

- All annotations lowercase and English, spaces replaced by underscores (\_)
- If the noun is ambiguous, annotate the possible readings separated by commas (no spaces). It is not necessary to list every reading the noun may be coerced into: annotate only the readings that are really salient in context.
- Annotations for coordinated NPs (internal argument of PP is coordination) are separated by slashes (no spaces: /).
- Slashes are also used in the determiner column, but not if the coordination applies below the determiner:

(5) mit ihrem ländlichen Charme und dem «unverdorbenen» Charakter  
with their rural charm and the untainted character  
*det\_coref\_subj*/ (note the slash at the end)

(6) mit ihrem ländlichen Charme und «unverdorbenen» Charakter  
with their rural charm and untainted character  
*det\_coref\_subj* (no slash)

- If there are multiple stimulus PPs, their annotations are separated by a plus sign (+), no spaces. This also applies to the determiner column.

## 5 Semantic Classes

We follow the CISLEX Project's classification (Langer, 2009, p. 81) with some modifications, namely the addition of *abstract\_obj*, *anim\_proxy*, *event(action)* and *ambiguous*. For a description of the other classes, please refer to that work (see also Table 1). In general, we agree with Langer that 'abstract object' is not a very useful category, but

Klassenbezeichnung	Hyperonym/ Beschreibung	Typische Kontexte/ lin- guistische Eigensch.
Lebewesen	Lebewesen	<i>X lebt/stirbt/vergeht</i>
Menschen	Alle Bezeichnungen für Menschen	<i>X lacht/weint</i>
Tiere	Tiere	<i>A tötet X</i>
Pflanzen	Pflanzen	<i>X wächst, wird gepflanzt</i>
Konkreta(Objekte)	Gegenstände	<i>X ist groß/klein/schwer...</i>
Stoffbezeichnungen	Bezeichnungen für Stoffe	<i>ein A aus X, ein Kilo X (X im Singular)</i>
Ereignisse	Ereignis	<i>X beginnt /endet, entspre- chende verbale Umschrei- bung möglich</i>
Zustände	Zustand	<i>X herrscht, A verharrt im X</i>
Eigenschaften	Eigenschaft	<i>X ist eine Eigenschaft</i>
Zeitbezeichnungen	Zeitpunkte/räume	<i>an X, ein X lang</i>
Formen	Abstrakte Begriffe für Formen	<i>ein X an A, ein X-förmiges A (wobei A ein Gegenstand ist)</i>
Lokativa	Ortsbezeichnungen	<i>in/an/auf X</i>
Diversa	Einzelklassen	—

Table 1: Overview of the CISLEX categories taken from (Langer, 2009, p. 81)

in the absence of a more appropriate classification, it is still more informative than *misc*. Thus, we use the following classes:

**organism (Lebewesen)** Only used if *human*, *animal*, and *plant* are not appropriate.

**human (Mensch)**

**animal (Tiere)**

**plant (Pflanzen)**

**anim\_proxy** Institutions, groups etc. (cf. our manual for the GerEO annotation, Masloch et al., 2021)

**concrete\_obj (Konkreta)** All concrete objects

**abstract\_obj** Non-concrete objects that don't fit into one of the other categories.

**material (Stoffbezeichnung)** prototypical mass terms with substance/material reading

**event (Ereignisse)** Event without overt/covert agents

**event(action)** Action

**state (Zustand)**

**property (Eigenschaft)**

**temporal (Zeitbezeichnungen)**

**form (Form)**

**locative (Lokativa)** Places etc. Since the test Langer proposes is not very useful, we use the following one: A noun can be considered locative if the NP it is heading may be the subject of *liegen* ‘to lie’ in its reading ‘be situated’.

**misc (Diversa)**

**ambiguous** Cannot be known from context given. *misc* is used for nouns that don’t fit into the other categories, *ambiguous* for nouns and pronouns that cannot be assigned a category within the context given.

## 6 *Comment\_PP*

This column contains additional comments of the annotators. During annotation, we noticed that many abstract objects were (parts of) pieces of art or related to utterances, so we commented (*part of*) *piece of art* and *utterance*.

## References

- Engelberg, S. (2015). Gespaltene Stimuli bei Psych-Verben: Kombinatorische Muster-suche in Korpora zur Ermittlung von Argumentstrukturverteilungen. In S. Engelberg, M. Meliss, K. Proost & E. Winkler (Eds.), *Argumentstruktur zwischen Valenz und Konstruktion* (pp. 469–492). Narr Francke Attempto.
- Langer, S. (2009). *Selektionsklassen und Hyponymie im Lexikon: Semantische Klassifizierung von Nomina für das elektronische Wörterbuch CISLEX* (PhD Thesis). LMU München.
- Masloch, S., Poppek, J. M., Robrecht, A. & Kiss, T. (2021). Syntactic pattern distribution analysis of experiencer-object psych verbs: An annotation manual. *Studies in Linguistics and Linguistic Data Science*, 4, i–51. [https://ldsl.rub.de/media/pages/research/resources/sllds/a3914a6ac2-1640260430/sllds\\_4.pdf](https://ldsl.rub.de/media/pages/research/resources/sllds/a3914a6ac2-1640260430/sllds_4.pdf)
- Poppek, J. M., Masloch, S. & Kiss, T. (in preparation). A corpus-based perspective on ‘split stimuli’ in German.
- Poppek, J. M., Masloch, S. & Kiss, T. (2022). GerEO: A large-scale resource on the syntactic distribution of German experiencer-object verbs. *Proceedings of the Language Resources and Evaluation Conference*, 3391–3397. <https://aclanthology.org/2022.lrec-1.362>