Manual of the semantic annotation

The table of semantic categories as presented in the following has been developed in the PhraseoRom project. It aims at annotating recurrent lexico-syntactic trees (RLTs). There are numerous semantic classifications which have inspired us to create our own list of semantic categories, e.g. Fellbaum (1998), Mahlberg (2007), Rayson (2008) or McCarthy et al. (2015). None of these classifications were completely suited for our categorisation due to the complexity of our study. Thus, it was necessary to develop a system of classifications that allows us to annotate non-phraseological multi-word expressions semantically.

1. List of semantic categories

dimensions	values
action	undefined
	movement
	travel
	other
communication	undefined
	verbal
	physical
state	undefined
	animate
	inanimate
	abstract
cognition	
qualia	undefined
	sensation
	perception
	affect
time	
place	
event	
other	*RLT which do not correspond to any of the categories*
-	*noise (RLT without verbs, traces of copyright)*

Vannina Goossens, Université Grenoble Alpes
Susanne Dyka, Friedrich-Alexander Universität Erlangen-Nürnberg

2. General principles of the categorization

The semantic categorisation is applied to recurrent lexico-syntactic trees (RLTs) extracted from the corpus. These are statistically significant co-occurrences which consist of lexical units in a specific syntactic pattern. Only RLTs containing a verb were analyzed. The annotation has been done in a spreadsheet.

2.1.   Format of annotation

- lower-case letters
- no spaces

2.2.   Dimensions and values

- The semantic categorisation comprises eight dimensions, some of them with sub-categories ("values").
- Additionally there is a category "other" for cases which cannot be categorized and a category "-" for cases which do not match our criteria and therefore constitute noise. When patterns are not correctly annotated and, for example, do not contain a verb, the pattern was marked "-".
- Dimensions which allow different values (action, state, communication and qualia) have one general value in common: "undefined". "Undefined" was chosen in cases where it was impossible to differentiate between the different values of a dimension or to allow the accumulation of dimensions (see below for examples).

2.3.   In case of two dimensions

In cases where RLTs were more complex a combination of two dimensions and the possible values was used. This is represented in the table with two columns for dimensions and values which were combined using either "AND" or "OR". The following cases describe this:

1. If a single semantic category does not entirely cover the meaning of an RLT, the two categories are linked by **AND**:
 – *you are in love*: `state:animate` AND `qualia:affect`
 – *acknowledge with a nod*: `communcation:physical` AND `action:movement`
 – *was in the sitting room*: `state:undefined` AND `place`

→ **order of the dimensions**:
- if the verb carries one dimension of the meaning and the other dimension is realised by a complement (for example), **the verb determines** the choice of the first category:
  follow into the kitchen = `action:travel` AND `place`
- if an RLT can be analysed in two different ways, the literal meaning determines the choice of the first category:
  take her in his arms = `action:movement` AND `communication:physical`

2. In case of polysemy, the different senses of the RLT are realised by OR:
 – *curling up on the sofa*: `state:animate` OR `action:movement`
 – *Going the wrong way :* `action:travel` OR `qualia:sensation`
→ **order of the dimensions**: the first category shows the most frequent meaning.

3. If both polysemy and complex meaning apply, the indication of polysemy takes precedence: the first semantic category then is given for each meaning.

Vannina Goossens, Université Grenoble Alpes
Susanne Dyka, Friedrich-Alexander Universität Erlangen-Nürnberg

→ **order of the dimensions:** the same rule as for polysemy alone was applied (= first category shows the most frequent meaning).

3. Definition of the dimensions and their values

3.1. Action

Dimension	Value	Example
action	undefined	
	movement	*open his eyes, wiped his brow, hold out a hand, held up a hand, cupped her breasts, held her in his arms, slapped on the back;* <u>movements with parts of the face</u>: *rolled his eyes* (+communication)
	travel	*follow into the kitchen, walk up/down the aisle, walk along/on/to the beach, walk into the bedroom, walk through/into cabin, walk to the center, walked in the direction, take a pace*
	other	*accept the offer, blow out the candle, broke the connection* (AND communication), *close the curtains, close the door, make a cup of coffee, have a drink, pick up the phone, get a good look*

This dimension comprises all actions executed by an agent in a dynamic situation.

special cases:
- actions without an agent (*the door closed, the phone rang*) and weather phenomena (*it was raining*) are classified as `event`.
- actions of a body part (i.e. feelings that 'involve' organs = where the body part is the 'agent', e.g. *my heart is beating*) are classified as `event` AND `qualia:affect` or `qualia:sensation` (or `qualia:undefined` if it unites sensation and affect)

3.1.1. Action:undefined

This category comprises actions, but it cannot be decided which of the values can be applied. For example:
- *made in the kitchen:* other or travel

3.1.2. Action:movement

This category comprises actions that imply some sort of movement without changing place horizontally (vertical movement):
- body movements (with the entire body)
    - changing of a sitting/standing/lying position to another one: *sit down, crouching, standing up, drop to his knees, got to his feet*
    - towards another person (AND `communication:physique`): *pull a hug, give a hug, give a kiss*
- movement of a body part: *hold out a hand, held up a hand, cupped her breasts, drummed his fingers, grabbed the phone*

Vannina Goossens, Université Grenoble Alpes
Susanne Dyka, Friedrich-Alexander Universität Erlangen-Nürnberg

- movement of the face/part of the face: *open his eyes, wiped his brow, smile* (but *his smile disappeared* = `event`)*, grit his teeth*

special cases:
- **actions with a body part as agent**: *her cheeks flushed* → `event`
- **actions which involve 'look' and lead to a perception** are not considered movements of a body part and are therefore classified as `action:other`.
- **verbs denoting movement or a state** (*sitting at a table* which could be "seated at a table" = state or "sit down at the table" = movement):
    - if the form of the RLT does not allow us to determine if it is movement or state (because there is no further indication of aspect or preposition), we classify it as both with action as the first and state as the second category: e.g*. sit on the sofa*: `action:movement` OR `state` (**but**: *was sitting in a chair* = `state:animate` (the aspect indicates a state) or *sit down on the sofa* = `action:movement`)

3.1.3. Action:travel

This involves movement from one point A to a point B, even if the two points are not mentioned explicitly (horizontal movement): *walk up/down the aisle, walk along/on/to the beach, walk into the bedroom, burst into the room, go to church, leave the apartment*
If point A and/or point B are expressed the RLT is analysed as place in the second category.

3.1.4. Action:other

Actions which do not indicate vertical nor horizontal movement. This category includes a variety of all sorts of actions: *accept the offer, blow out the candle, broke the connection* (AND communication), *close the curtains, close the door, make a cup of coffee, have a drink, pick up the phone, avoid like the plague, try to avoid*

special cases:
- actions which lead to a perception: *look at the camera, look down at his feet, stare into the fire, get a good look*: `action:other`
- RLTs which involve a movement but do not mention it explicitly: *raise a glass, feel in his pocket*: `action:other`(OR `qualia:perception` in the case of *feel his pocket*)

3.2.   Communication

Dimension	Value	Examples
communication	undefined	*accept the offer*
	verbal	*have a chat, have a conversation, say the boy/to the barman/the captain/ the captain said/ say with a chuckle/with dignity/with emphasis, add in a tone, call for help, cleared his throat*
	physical	*give a smile, blow/give a kiss, show*

Vannina Goossens, Université Grenoble Alpes
Susanne Dyka, Friedrich-Alexander Universität Erlangen-Nürnberg

		the way (OR communication:verbale)

This category comprises RLTs that express either the intention or the actual act of sharing information, feelings, thoughts etc. with another person and in various ways.

special cases:
Some verbs do not allow to differentiate between communication:physical and communication:verbal: *show the way* → `communication:verbal` OR `communication:physical`.

3.2.1. Communication:undefined

This category allows us to accumulate several dimensions as some RLTs might express both forms of communication and another semantic dimension. Thus, the RLT *show the way* is analysed as `communication:verbal` OR `communication:physical` as this RLT does not have a second semantic dimension. But for *offered a cigeratte* this analysis is not possible as another dimension (`action:other` AND `communication:physical` AND `communication:verbal`) would be needed. The value `undefined` allows us in this case to keep both dimensions (→ `action:other` AND `communication:undefined`).

3.2.2. Communication:verbal

Communication with the help of the articulatory organs
- with propositional content (intentional utterance, with communicational intent): *have a chat, have a conversation, say the boy/to the barman/the captain/ the captain said/ say with a chuckle/with dignity/with emphasis, add in a tone, call for help, cleared his throat, left a message*
- without propositional content (e.g. unchecked physical reactions): *gave out a harsh shout, give a cry, give a scream, gave a snort* (AND `qualia:affect`)
- indication of the quality of the voice: *he lowered his voice*

special cases:
- **animal sounds**: *the dog barked* → `communication:verbal`
- **laughter, tears and mockery:** *had a laugh, he was joking, made a joke*
- **absence of communication:** *I wasn't talking* → `communication:verbal`
- **promises and oaths:** → `communication:verbal`
- **without agent:** *a voice said, got a call (*AND `communication:verbal`)→ `event`
- *finish the call*: `action:autre` AND `communication:verbal`

3.2.3. Communication:physical

Establishing non-verbal contact with the objective of sharing information, feelings, thoughts etc. with another person: *acknowledge with a nod, take her in his arms*

special cases:

Vannina Goossens, Université Grenoble Alpes
Susanne Dyka, Friedrich-Alexander Universität Erlangen-Nürnberg

- for verbs which indicate a body/face movement, the category communication is the second category, e.g. *blow/give a kiss* = `action:movement` AND `communication:physical`
- *raise a glass* → in those cases no physical communication or movement are implied and therefore it is analysed as `action:other`

3.3. State

Dimension	Value	Examples
state	undefined	*was in the corridor, was in the house*
	animate	*be in love, was wearing a dress/shirt, I am tired, I am not interested, was on his knees*
	inanimate	*the words were out, 's the money, The body lay*
	abstract	*it was in the world, there were in the sky, there was in the car*

States of animate, inanimate and abstract subjects. This category comprises RLTs which denote different lengths of periods. A differentiation is sometimes made between short period states (stage-level: the dog is small -> it will grow) and more permanent states (individual-level: the dog is small and the elephant is big).

special cases:
- Changes of states (e.g. *fall asleep, wake up*) are categorized as `event`.
- States that have a component of cognitive, sensory or affective experience will be classified separately as `cognition` and `qualia` (affect, sensation, perception).
- Verbs which denote a state or a movement (*sit* vs *sit down*) → cf special cases in action:movement for annotation rules
- If RLTs contain more than one verb it is possible to accumulate actions and states in the analysis even if this seems contradictory. For example, *sit in silence* is annotated as `state:animate` AND `communication:verbal`. To be consistent, we follow the rules of annotation as described before: in particular the absence of communication or movement is annotated in the same way as its presence.

3.3.1. State:undefined

RLTs missing qualifying elements which would allow to determine if the subject is animate or inanimate cannot to be categorised with a specific value, e.g. *stood in the hall* (in this case it is not clear if the subject is a person or an object), *float in the air*

3.3.2. State:animate

States of animate subjects: physical states (*I am tired*) and psychological characteristics (*he was charming*).

Vannina Goossens, Université Grenoble Alpes
Susanne Dyka, Friedrich-Alexander Universität Erlangen-Nürnberg

3.3.3. State:inanimate

States of inanimate subjects: *The body lay*

3.3.4. State:abstract

States of abstract subjects: *It's a trap*

3.4. Cognition

Dimension	Value	Examples
cognition		*get the message, get the impression, have a dream*

RLTs which describe the process of acquisition or the loss of knowledge: *learn, understand, think, remember, forget, dream, imagine ...*

special cases:
− The dimension `qualia:perception` also allows to classify intellectual phenomena but it focuses on information perceived through the senses: *perceive, feel*

3.5. Qualia

Dimension	Value	Examples
qualia	undefined	*like the smell*
	sensation	*had a heart attack,*
	perception	*breathing the scent, caught a whiff, hear in the kitchen,*
	affect	*felt a rush/a stab/a tingling/a surge, he was in a mood, I'm shocked*

Qualia are about perception and more generally about sensorial experiences: sensations, perceptions and affects.

special cases :
− **actions of a body part**: `event` AND `qualia:affect` or `qualia:sensation` or `qualia:undefined` if affect or sensation are present.
− **tiredness, sexual desire and nausea:** `qualia:sensation` AND `qualia:affect`

Vannina Goossens, Université Grenoble Alpes
Susanne Dyka, Friedrich-Alexander Universität Erlangen-Nürnberg

3.5.1. Qualia:undefined

3.5.2. Qualia:sensation

Physical sensation perceived by the five senses: *had a heart attack*

3.5.3. Qualia:perception

Mental representation of a sensation, organisation of sensory impressions. Activity by which someone experiences objects or properties in his/her environment and this "experiencing" is based on information perceived by the senses: *inhale the scent*, *caught a whiff*

special cases:
- Some RLTs denoting perceptions are similar to cognition, but can be distinguished from cognition in the absence of mental/cognitive processing: *feel a presence*
- perceptions following an **action** are classified as `action:other`: *breathing the scent*, *look at* (vs. *see*), *listen to* (vs. *hear*).

3.5.4. Qualia:affect

Feelings and emotions: psychological states which vary in their intensity and are generally marked as positive or negative that are experienced by an animate subject (being), that are connected with a source (a cause or an object), that are inserted in a temporal frame and that cannot be observed: *give a toss, be in love*

special cases:

3.6. Place

Dimension	Value	Examples
place		*was at the gates, was in the sitting room, was in town, gone into the house*

Limited to place or location, objects are excluded (*armchair)*. Doors and windows are considered places if they are combined with a motion verb and function as indication of location: *and walked to the window* (`action:travel` AND `place`) but not *watched through the window* (`action:other`).

special cases:
- If place is not really indicated by the verb but additionally, this will be categorised in a second dimension: *gone into the house* = `action:travel` AND `place`

3.7. Time

Vannina Goossens, Université Grenoble Alpes
Susanne Dyka, Friedrich-Alexander Universität Erlangen-Nürnberg

Dimension	Value	Examples
time		*spend a week, spent the rest of his life,*

This dimension classifies RLTs which comprise temporal information.

special cases:
- Weather phenomena with a temporal component (e.g. *the sun is rising*) are classified first as `event` or `state:inanimate`, then as `time`.

3.8. Event

Dimension	Value	Examples
event		*filled the air, the lights flickered, the alarm went off, the lights went out, the door opened*
		his eyes fixed, get a call, (the words/smell) hung in the air, a shiver ran
		the sun rises

A dynamic process with a non-agentive dimension which can indicate a change of state is characterised as event (Talmy 1998, 2000).

- **"actions" without agents**: *the lights flickered, (the words/smell) hung in the air*, *the door opened*
- **non-static weather phenomena**: *the sun rises*. Static weather phenomena are categorised as `state:inanimate`.
- **verbs which imply the passing of time but without agent:** If the event affects the subject, a double categorisation is possible with `qualia` as second category.
- **changes of states:** *dozed off, fall asleep*, *gone to sleep*, *woke in the morning* (AND `temps`).
- **actions of a body part**: *his eyes fixed*, *her cheeks flushed*.
- **actions which imply the feelings of an organ**:

special cases:
- *get a call*: event AND communication:verbal

3.9. Other

This mixed-bag category contains RLTs which cannot be attributed to one of the other categories. For example, RLTs with "be" and "have" can be found there. The RLT does not provide enough information to determine the semantic dimension of the verb: *the truth is, the driver was, the crew were ...* which could be categorised as state, action, qualia …

Vannina Goossens, Université Grenoble Alpes
Susanne Dyka, Friedrich-Alexander Universität Erlangen-Nürnberg

3.10. Diverse cases

aspect and modals: *was trying to persuade*
→ are categorised according to the verb: *persuade* = `communication:undefined`
→ neither aspect nor modality is indicated in the categorisation

4. References

Baudet S. (1990). "Représentations cognitives d'état, d'événement et d'action". *Langages*, 100, 45-64.

Carlson, G. N. (1977). *Reference to Kinds in English*. Ph.d. dissertation, University of Massachussets, Amherst. Published by Garland Press, 1980.

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge.

François, J. (1990). "Classement sémantique des prédications et méthode psycholinguistique d'analyse propositionnelle". *Langages*, 100, 13-32.

Mahlberg, M. (2007). "Corpus stylistics: bridging the gap between linguistics and literary studies". *In*: Michael Hoey, Michaela Mahlberg, Michael Stubbs, and Wolfgang Teubert, eds. *Text, Discourse and Corpora. Theory and Analysis. With an introduction by John Sinclair*. London: Continuum, 219–247.

Martin, F. (2008). "Deux types de stage-level predicates". *Langages*, 169, 111–128.

McCarthy, D., Kilgarriff, A., Jakubicek, M., Reddy, S. (2015). "Semantic word sketches". *Corpus Linguistics, 2015*.

Oxford Learner's Dictionaries [OALD online]: https://www.oxfordlearnersdictionaries.com

Rayson, P. (2008). "From key words to key semantic domains". *International Journal of Corpus Linguistics*, 13:4, 519-549.

Talmy, L. (2000). Toward a Cognitive Semantics. Vol. 1. *Concept structuring systems*. Cambridge, MA (The MIT Press).

Talmy, L. (1988a). Force Dynamics in language and cognition. In *Cognitive Science*, 12, 1, 49-100

Vendler, Z. (1957). "Verbs and Times". *The Philosophical review,* 66:2, 143-160.

Vendler, Z. (1967). *Linguistics in philosophy*. Ithaca: Cornell University Press, 97 121.

Vannina Goossens, Université Grenoble Alpes
Susanne Dyka, Friedrich-Alexander Universität Erlangen-Nürnberg