Annotation Guidelines for the Feature-based Approach

In the following, we describe the category inventory which should be considered for the annotation of utterances. (Each utterance comprises exactly one sentence.) The categories are **not** mutually exclusively. That is, for each utterance more than one category may apply.

A prerequisite of all categories is that the overall meaning of the sentence has to be <u>negative</u> (or at least is meant to be perceived as negative). Sarcastic sentences may also be considered negative if their intended meaning is negative. Utterances that do not fulfil this requirement should not be considered for any of the following categories.

Many of the of the utterances are creative paraphrases of simple cue phrases. (For instance, for the cue phrases *You are arrogant* we have the creative paraphrase *You should compete in the ego Olympics*.) For some of the following categories the set of those cue phrases needs to be considered. Therefore, they are also provided for this annotation task.

Negated Antonyms of (Explicitly) Abusive Words

We are looking for sentences in which there is a negated antonym of an explicitly abusive word, for example, *You are <u>not clever</u>* (*clever* is an antonym of the explicitly abusive word *stupid*) or *You are <u>not humble</u>* (*humble* is an antonym of the explicitly abusive word *arrogant*). These understatements may be perceived as euphemistic.

In the following, we clarify what we understand by "negation", "antonym" and "explicitly abusive word":

- We use a wide concept of negation. That is, we do not only count "strict" negation words (e.g. *no*, *not*, *never*, *nothing*, *without* etc.) but also shifters (e.g. *lack of* X, *hardly*, *fail to X*, *violates X* etc.).
- You should not pay (too much) attention to the semantics of intensifiers. For instance, *You are not very bright* should be read as *You are not bright*. In principle, any formulation that is semantically similar to a negation would count (e.g. *You could be more intelligent==You are not intelligent*).
- Explicitly abusive words: You should not rely on just one specific lexicon containing such words since all existing lexicons have a limited coverage. However, if a word is also an any entry of [the lexicon of abusive words from Wiegand et al. (2018)]() then the word should definitely be counted as an explicitly abusive word; any content word from the "cue phrases" of the dataset euphemistic abuse counts as well.
- By antonyms we only mean **gradable antonyms**, i.e. word pairs whose meanings are opposite and which lie on a continuous spectrum, e.g. good/bad. So, we are not interested in other forms of antonyms, e.g. *come*/*go*.
- For establishing an antonym relation between two lexical units, again, you should not only rely one single resource as every existing resource will have a too limited coverage. Nevertheless, the following resources/methods may help you: WordNet (this resource includes the relation "antonym"); Google (given the word <WORD> you want to examine, just type in as a query "antonym of <WORD>" and you will typically get pointers to websites that list antonyms of <WORD>. If something is considered a "near-antonym", then you should also consider this as an antonym as well since we use a wide concept of antonyms.)

Opposing Sentiment

This category should be used for utterances in which we observe opposing sentiment. For the realization of individual polarities, following things are to be considered:

- By and large, we are aiming for explicit sentiment, i.e. sentiment that is expressed by polar expressions (e.g. *love*, *hate*, *nice*, *awful*, *beauty*, *horror* etc.)
- Explicit sentiment may be conveyed by adjectives, verbs, nouns and multi-word expressions (e.g. *kick the bucket*).
- We also allow implicit forms of sentiment if they express situations that are generally considered positive/negative. (For example, *being woken up in the middle of the night* should be considered a negative situation while *getting a job* should be considered a positive situation.) In terms of what situations count as "positive" or "negative", Western values should be assumed.

With regard to opposing sentiment, the following types of utterances should be considered:

- A positive assessment of something generally considered negative. For example, this could be a positive evaluation of a negative character trait. Such utterances are often perceived as contradictions, many of them are also ironic/sarcastic, e.g. *You are very good at disappointing people* or *You are the king of laziness*.
- The utterance describes the fondness of the target towards something generally considered negative (e.g. a negative event, activity, ideology). Examples are *You seem to like seeing people suffer in awkward situations* or *You love the totalitarian regimes on this globe*. Moreover, fondness of the target towards someone generally considered negative/evil (e.g. *Hitler*) should also be considered. For example, *You really adore Hitler*. In order to decide what is "good" and "evil", we assume Western values.
- Making someone generally considered evil content/pleased, e.g. *With your actions, you would make the devil smile*.

Taboo Topics

For this category, we are looking for any utterance that refers to a taboo topic. As taboo topics we define:

- sex
- death and illness
- bodily functions/excretion (including lack of hygiene)
- physical/mental abnormity

We regard drug addiction and drug abuse as some form of illness. Therefore, this counts as a taboo topic.

If someone's clothing is commented on then this is **not** considered a taboo topic. However, if someone's appearance is commented on and this comment refers/alludes to some physical abnormity, then this should be considered a taboo topic.

Extreme

This category should be used for any utterance which can be considered as some form of extreme or absolute language. In the following, we list the following criteria that each qualify for this category (or not):

- Superlatives (e.g. *best*, *worst*, *most adorable* etc.) Notice that we are only interested in plain superlatives. Formulations like *not the best* or *not the worst* do not count.
- Similar to superlatives creative expressions such as *leader of stupidity* or *king of trouble* should be considered as instances extreme.
- Generalizations: they can be conveyed by certain adverbs (e.g. *never*, *always*, *not at all*). Note that any form of (scientific) definition does not qualify for a generalization.
- Please be careful with negations. You should only consider a negation as an instance of extreme if it reads as an exhaustive negation as in *No one would regard you as very attentive, You never listen to me*, *You are not listening at all*, *There isn't anything you pay attention to*, *My cat is devoid of attention.* A formulation such as *You do not listen* would not qualify as it mostly reads as *You do not listen properly*. Negation involving *any* may be difficult. For example, *With these measures I do not feel any safer* would not be considered extreme, however, *You have not considered any single idea that was brought forward by your colleagues* would.
- Any exaggeration (hyperbole), such as *You would fit well in a kindergarten* or *I would rather chew grass than write you a letter of recommendation*. Hint: watch out for expressions that have a property that can be placed at an extreme position of a scale, for instance *galaxy*, *atom* (scale: size); *feather*, *elephant* (scale: weight).
- Exuberant assessments/evaluations (e.g. *impressive*, *amazing*, *marvellous*, *superb*). Please note, however, that not every (unambiguously) positive assessment/evaluation counts as exuberant. For example, *I like X* or *This is great* should not be considered as exuberant. However, *I (really) love X* or *This is excellent* should.
- Our dataset contains many contradictions that are meant sarcastically. Please note that not all of these sentences count as an instance of this category. For instance, *You are good at avoiding work* should not be considered as an instance of extreme but *You are the greatest at avoiding work* should.
- An intensified (negative) expression does not suffice for this category, per se. For example, *You are very disappointed* should not be labelled as an instance of extreme. There are a few intensifiers, however, that would qualify, such as *particularly*, *especially*, *immensely*, *extremely*.
- Formulations matching the pattern *far from X* as in *far from easy* or *far from perfect* should not be considered as extreme as they often do not imply an end-of-scale value. For example, *far from perfect* does not imply *extremely bad*.
- Formulations matching the pattern *too X* should be considered as extreme as long as they are not negated (i.e. *not too X*).
- Expressions that imply an extreme negative assessment pragmatically, relying on world knowledge. These are often recognizable only by considering other possible/alternative propositions. Some of such cases are marked by expressions such as *even* or *let alone* but they need not be.

    - *I won't call you let alone visit you*.
        - If we had only an isolated *I won't call you* then for lack of context, we'll err on the side of interpreting it as non-extreme
    - *I don't even give her the time of day* (= I completely ignore her)
    - *I don't give him the time of day.*

        - *Not give X the time of day* is an idiom that conventionally expresses great disregard
    - *You should be behind bars*. (cf. *Lock her up!*)
        - Again, this makes sense only if we're imagining this being said in an everyday interaction rather than say in a court of law by a prosecutor or judge. But this is what we do: we adopt non-special contexts to the degree possible.
    - *I would rather not be seen with you.* (avoiding even being seen with somebody implies a very negative attitude)

- Interactions: The above features may not occur by themselves in atomic sentences but interact with each other or still other constructions. Co-occurrence with positive polarity expressions: we ignore those, even if discourse structure signals that they should be taken as more important than the negative expressions:

        - *Max is quite nice as a person but he's the worst player ever.* → extreme
        - *Max is the worst player ever but quite nice as a person.* → extreme

    - Co-occurrence with modal and epistemic markers
        - Epistemic and modal markers such as *obviously*, *clearly*, *surely*, *definitely* are not taken to contribute extremeness in the relevant sense: we focus on extreme evaluation or propositional content, not the attitude of the speaker towards the proposition:
        - *He is definitely/obviously/certainly impolite.* → not extreme
        - Similar to the above, claims about widely shared but weak attitudes do not count as extreme.
            - *Everybody knows that you're ...* → not extreme
            - *All the world has realized by now that you are …* → not extreme

    Continuing with the above logic, weak epistemic and modal markers such as may, possibly, probably do not detract from propositional extremeness.

        - *Max may be the most irritating guy I've ever met.* → extreme
        - *Kim probably bores the hell out of everybody she meets.* → extreme

    And likewise, strong attitudes held by few people count as extreme:

        - *Some people consider you the worst person in the world.* → extreme

    - Co-occurrence with frequency markers:

    evaluative expressions

        - *Max is always the worst player.* → extreme
        - *Max can be annoying sometimes* → not extreme
        - *Max is rarely (=not often) the worst player.* → not extreme

    addressee evaluated negatively for situations they are involved in

        - *You're always late.* → extreme
        - *You rarely disappoint.* → not extreme
        - *Sometimes/occassionally you irritate/annoy me.* → not extreme

Lexicalization: Derogatory Idioms

This category should be used for utterances in which certain idioms occur. We specifically look for any idiom that expresses a derogatory connotation towards the target (e.g. *You are not the sharpest knife in the drawer*). Further criteria are:

- The idiom should be some expression that is lexicalized (i.e. that could potentially be found in a dictionary). [Since dictionaries do not cover idioms very well, we do not insist that there is actually a publicly available dictionary that indeed includes an entry with the pertaining idiom.]

- The idiom should be decoding, that is, its semantics should not be explainable by the composition of its component words. For example, *You are Satan's favourite* would **not** count as an idiom.
- The idiom itself must be derogatory. It is insufficient if the derogatory meaning is conveyed by the interplay between the idiom and one of its arguments. For example, in *I dreamed of you 6 feet under* the derogatory nature does not stem from the idiom *6 feet under*.
- Many of those **derogatory idioms may be euphemistic**, e.g. *Were you shopping with a five-finger discount again? (five-finger discount=theft)*.
- If an idiom is a synonym or near-synonym to a predicate of one of our cue phrases, then this idiom should be considered derogatory (For example, the phrase *ladies of the night*, which is a euphemism for *prostitute*, should be considered derogatory because *whorey*, i.e. a near synonym to *prostitute*, is in one of our cue phrases *You are whorey*. We consider those predicates from the list of cue phrases as derogatory.) Please remember that a mere paraphrase of these predicates is insufficient to be labelled as idiom. The paraphrase must be lexicalized, decoding and have a derogatory connotation.
- We consider multiword expressions as idioms. However, we also consider hyphenated compounds (e.g. *sweet-talk*) and even closed compounds (e.g. *currymuncher*).
- Be careful with negations. They may be part of the idioms we are looking for (e.g. *not have all marbles*) but a negation with some multiword expression need not always be the type of idiom we are looking for (e.g. *Refinement is not your forte* – this sentence does not include an idiom that is in itself derogatory).

„Unusual"

This category should be used for any utterance which includes some unusual property/behaviour/situation that is related to the target. In the following, we clarify what we mean by "unusual":

- The target is attributed unusual properties or displays some unusual behaviour. This could be strange hobbies (e.g. staring at an empty wall for hours), preferences (e.g. fancy other people failing) or beliefs (e.g. believing in fairy tales).
- The target causes unusual situations/events (e.g. *You will cause anyone to scream and run away*) or unusual behaviour on the part of the speaker (e.g. *You will make me want to do very nasty things*).
- Negative behaviour/properties do(es) not count as unusual behaviour/properties per se. For example, many people are not fond of tidying up. Neither are people always considerate. If a sentence simply states such negative behaviour, then this does not qualify for this category. The degree to which someone displays negative behaviour/properties, however, may decide whether the sentence is considered to be unusual or not. For instance, *You don't like tidying up* is not particularly unusual. However, the alternative formulation *You have never seen a feather duster in your life, have you.* suggests an unusually strong reluctance towards tidying up coupled with the allegation of being out of touch with everyday life.
- Unusual situations/behaviour etc. sometimes coincide with unusual wording/formulations, e.g. *Remind me not to give you these information* (one would expect something like: *Remind me to give you these information*) or *You inspire people to do the wrong thing* (one would expect something like: *You inspire people to do the right thing*).

- The usage of imagery can be clue for this category: *Your brain bid farewell to you a long time ago* or *You wouldn't know honesty if it passed you in the street.*