Guidelines for the Annotation of Deliberate Linguistic Metaphor

Stefanie Dipper and **Adam Roussel** and **Alexandra Wiemann**
and **Won Kim** and **Tra-My Nguyen**
Department of Linguistics
Fakultät für Philologie
Ruhr-Universität Bochum
firstname.lastname@ruhr-uni-bochum.de

Abstract

This paper presents guidelines for the annotation of deliberate linguistic metaphor. Expressions that contribute to the same metaphorical image are annotated as a chain along with a semantically contrasting expression of the target domain, which helps to make the domain contrast inherent to metaphor more explicit. So far, a corpus of ten TEDx talks with a total of ca. 20k tokens has been annotated according to these guidelines. 1.35% of the tokens are deliberate metaphorical expressions according to our guidelines, which shows that our guidelines successfully identify a significantly higher proportion of deliberate metaphorical expressions than previous studies.

1 Introduction

In conceptual metaphor theory (Lakoff and Johnson, 1980, CMT), the idea of a conceptual metaphors refers to the understanding of one conceptual domain in terms of another. This involves taking an expression from a literal, usually more concrete, source domain and transferring it onto a target domain in order to shape our understanding of this target domain concept in some way. This cross-domain mapping effects a transfer of properties of the source domain to the target domain, as the source domain is reinterpreted.

Such conceptual metaphors can be implemented in any of a number of ways, but one common medium for conceptual metaphors is language. Linguistic metaphor is often associated with certain properties: there is usually some kind of semantic mismatch between certain words in a sentence, which triggers the reinterpretation of the metaphorically-used words. According to Hanks (2013), such mismatches, which he calls 'exploitations', stem from a deliberate departure from an established pattern of normal word use. For instance, in example (1), the subject *Bodenschätze* 'natural resources' (lit. 'ground-treasure'), is normally used with container expressions referring to soil or huge shipping containers, so referring to people's minds as containers deviates from the norm. As a consequence, *Bodenschätze* is reinterpreted as the valuable content of minds, such as intelligence or creativity.

(1) Das kann sich ein Land, dessen Bodenschätze in den Köpfen unserer Bevölkerung stecken, nicht leisten.
'A country whose natural resources are in the minds of our population cannot afford this.'

There is a related notion that metaphoric expressions can be observed to stand out in their immediate context, that it will be surprising to find language pertaining to product packaging in the context of a poetry slam for instance, as in example (2), and this element of surprise can also trigger the reinterpretation of expressions that are intended metaphorically.

(2) Du bist so vakuumverpackt, so in deiner Komfortzone versackt.
'You are so vacuum-packed, so stuck in your comfort zone.'

In order to learn more about the linguistic dimensions of metaphor and the relationship between linguistic metaphors and their context, we annotate whole texts and will eventually expand our corpus to encompass a variety of text genres.

Previous annotation efforts that have covered the annotation of complete texts, most notably the VUA Metaphor Corpus (Steen et al., 2010), often used guidelines oriented broadly towards the annotation of all kinds of metaphor, and accordingly their datasets consist mostly of conventionalized metaphors, of which speakers are mostly unaware and which don't serve a particular discourse-communicative purpose. In contrast, our guidelines

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 53–58
June 21, 2024 ©2024 Association for Computational Linguistics

are focussed more squarely on *deliberate metaphors* in the sense of Steen (2008), which play an important role in a discourse and of which speakers and listeners are likely aware.

The contributions of this paper are: (i) annotation guidelines for identifying deliberate metaphor; (ii) an annotated corpus of TEDx talks with 20k tokens, which is made freely available.[1]

2 Related work

The first work on the annotation of metaphors in texts comes from an interdisciplinary group of researchers who define a Metaphor Identification Procedure (MIP) to recognize metaphorically used expressions in texts (Pragglejaz Group, 2007). The MIPVU guidelines went beyond MIP by also taking into account explicit comparisons or similes (Steen et al., 2010). In both approaches, the annotator must first determine the contextual meaning of a word, i.e. the current meaning in the text, and then use a reference lexicon to check whether there is a 'more basic' literal meaning (e.g. a more concrete meaning). If the contextual meaning is in contrast to the literal meaning, but is at the same time in some way similar and can be understood in comparison to it, the word is labeled as 'MRW' (metaphor-related word). The guidelines are designed as to identify all metaphors, including conventionalized ones.

Steen et al. (2010) annotated the VUAMC (VU Amsterdam Metaphor Corpus) according to MIPVU. The corpus contains 190k words and consists of fragments from four registers of the BNC-Baby corpus (academic texts, conversation, fiction, and news texts). 86% of the words are clearly non-metaphorical and 13% are clear MRWs, and 1% are borderline cases. The highest proportion of MRWs is found among prepositions. In different studies on inter-annotator agreement (IAA), Steen et al. (2010) achieved Fleiss' κ between 0.70 and 0.96 (with texts in English and Dutch).

Deliberate metaphor The DMIP guidelines (Deliberate Metaphor Identification Procedure) aim at excluding dead and conventionalized metaphor (Reijnierse et al., 2018). Deliberate metaphors are those that are intentionally used as metaphor and draw attention to the cross-domain mapping, as opposed to conventionalized metaphors where no such processes take place. According to the DMIP guidelines, only *potentially* deliberate metaphors can be identified sensibly. Rather than providing detailed and specific criteria for the identification of deliberate metaphor, Reijnierse et al. (2018, p. 137) give the following instruction: "Determine whether the source domain of the MRW is part of the referential meaning of the utterance in which the MRW is used." However, they mention some typical indicators of deliberate metaphor, including novel metaphor and extended metaphor, consisting of multiple words that relate to the same metaphor, as well as direct metaphor, signaled by lexical cues such as *as* or *like*, or topic-triggered metaphor, where lexis related to the overall topic of the text is used metaphorically.

The DMIP guidelines have been tested on pre-marked MRWs of a set of selected VUAMC sentences, resulting in Cohen's κ between 0.70 and 0.73 (with 129 and 130 pre-marked MRWs from VUAMC, respectively). In the two datasets, 11.6% and 9.2% of the MRWs are annotated as deliberate.[2] The size of the data sets is not specified in the paper, though. Since around 11.1% of all tokens in VUAMC are MRWs, it can be estimated that deliberate metaphor accounts for approximately 1.2% of all tokens.

Beigman Klebanov and Flor (2013) present an annotation protocol for the identification of "metaphorical expressions that are noticeable and support the author's argumentative moves" (p. 15). The guidelines do not specify detailed criteria for identification, but rather describe metaphors in general terms: "Generally speaking, a metaphor is a linguistic expression whereby something is compared to something else that it is clearly literally not, in order to make a point." (p. 14). A total of 116 test-taker essays, discussing the role of electronic media for communication, are annotated with 55k tokens ($\kappa = .575$). On average, the two student annotators marked 4.86% of all tokens as metaphorical according to the guidelines; the union set, which serves to account for the fact that disagreement is often due to attention slips (Beigman Klebanov et al., 2008), comprises 6.83% of all tokens. The evaluation shows that verbs in particular are used metaphorically disproportionately often.

Novel metaphor Do Dinh et al. (2018) investigate novel metaphors (which constitutes a subset of deliberate metaphor). Their work is based on the VUAMC. For all content-word MRWs (i.e. excluding auxiliaries and prepositions), they an-

[1] https://gitlab.rub.de/comphist/figlang2024

[2] The annotated MRWs are freely available at https://osf.io/c8bxs.

notate whether the metaphor is novel, i.e. non-conventionalized. Crowd workers receive random samples with four MRWs each and annotate which of these is the most novel and which is the most conventionalized (no IAA calculable). The proportion of novel metaphors (353) of all tokens (240k) ranges from 0.04–0.26% across the four registers.

Parde and Nielsen (2018) also investigate novel metaphor and annotate MRWs from the VUAMC, similar to Do Dinh et al. (2018). However, the crowd workers only annotate selected word pairs that consist of content words (or a personal pronoun), at least one of which is an MRW and which are syntactically linked. The annotations consist of gradual scores, from 0 'not metaphoric' to 1 for 'low metaphor novelty' up to 3 for 'high metaphor novelty'. IAA was calculated between trained annotators with κ scores of 0.435, and, with relaxed constraints, 0.897 (on 3k instances). In total, the corpus contains more than 18k annotated word pairs, however, the exact proportion of novel metaphor (with scores 2 or 3) is not specified in the paper.[3]

Alnajjar et al. (2022) annotate metaphors in 27 YouTube videos of the start-up domain. The criteria for annotation are kept very simple: A word is considered a metaphor if its meaning is not literal, if the meaning is not listed in the lexicon (i.e. it is not a conventionalized metaphor), or if it is not meant sincerely but sarcastically. However, if the metaphor includes several words, it is considered an idiom and annotated, even if it is conventionalized (e.g. *give it a shot*). The two expert annotators annotate both vehicle (the metaphorical expression from the source domain) and tenor (the expression from the target domain) – the criteria for tenor, however, remain unclear, as these are typically interpreted literally. No IAA is reported. In total, 672 metaphorical tokens have been annotated, among them 45% novel metaphors, which roughly seem to correspond to 0.23% of all tokens.

Resources for German To the best of our knowledge, there are no annotated texts for German available. Herrmann et al. (2019) adapt MIPVU to German. They calculate IAA for set of 559 sentences, obtaining Fleiss' κ = 0.71. The analyzed corpus of 20k sentences is not available.

Egg and Kordoni (2022, 2023) also adopt the MIPVU guidelines, but extend them to include the annotation of elements in the context of the metaphorical expression that trigger the metaphorical meaning, which they call 'background'. They also determine the conventionality of an MRW: An MRW is conventionalized if its meaning is listed in the lexicon. Using INCEpTION (Klie et al., 2018), they annotate a corpus with five different registers, which should ultimately contain 150k words. In Egg and Kordoni (2023) an IAA of Krippendorff's α = 0.89 is reported, but it is unclear on which data this was calculated. In their data, the conventionalized MRWs have a proportion of 4–15% and the non-conventionalized MRWs of 0.01–0.29% (again, the size of the underlying data is unclear). The guidelines and the corpus are not yet available.

3 Guidelines

We are interested in deliberate metaphor in German-language data. In most studies, deliberate MRWs represent a very small proportion of all tokens, less than 0.3%. The study by Beigman Klebanov and Flor (2013) clearly deviates from this with proportions of 4.86 and 6.83%, but it is unclear whether this is due, for example, to the open guidelines or to the text type or to the fact that the texts come from learners.

Our aim is to produce guidelines with specific criteria, offering supportive guidance for the annotators, so that the proportion of overlooked cases due to attention slips is minimized and we are able to identify more instances of deliberate MRWs than has been the case in previous studies. Our criteria, detailed in the following, are based on those for deliberate MRWs in Reijnierse et al. (2018).

Deliberate An MRW is considered deliberate if the metaphorical image is new or if the MRW used for an known metaphorical image is unusual and innovative. Alternatively, the MRW can be deliberate because it is marked in some way, e.g. if it occurs in a construction that is normally used in the active voice but now occurs in the passive voice, if the MRW is typographically emphasized, e.g., by italics or quotation marks, or if it stands out because it also appears in the title of the text.

For instance, example (3) contains a well-known metaphor, *ein Strauß an Forderungen* 'a bouquet of demands'. However, this established metaphor is expanded and modified by the adjective *bunt* 'colorful' and the verb *binden* 'to bind', so we consider it a deliberate metaphor.

[3] The data are available at https://computerscience.engineering.unt.edu/labs/hilt/resources.

(3) einen Strauß bunter Forderungen binden
'tying a bouquet of colorful demands'

The label **grey area** is used when an MRW shows characteristics of both deliberate and conventionalized metaphors.

Revitalized A subset of conventionalized expressions is also relevant here, namely revitalized usages: A conventionalized MRW can appear in a new light in a particular context, e.g. when a deliberate MRW that refers to the same image occurs in the immediate vicinity, so that the conventionalized expression could plausibly have been chosen deliberately rather than arbitrarily or a listener might plausibly perceive it in this way. The otherwise conventionalized expression is thus considered 're-vived' or revitalized.

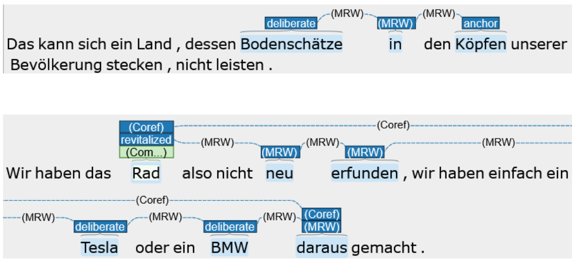
Anchor Usually, the annotation process begins when an annotator, in the course of reading through a text, notices some unusual or conspicuous combination of words, which impression is often the result of a domain clash or a kind of semantic incompatibility between them. One of the words, corresponding to the source domain, then needs to be re-interpreted metaphorically, while the other, corresponding to the target domain, is taken literally. We label this second expression the 'anchor', as this is the expression that 'anchors' the metaphorical image in reality. In example (3) above, the anchor is *Forderungen* 'demands', because this is the expression that is intended literally – the statement is ultimately really about 'demands' of some kind and not flowers.

In addition, we mark **flags** (Steen et al., 2010) indicating a comparison, e.g. expressions such as *wie* 'like' or *sozusagen* 'so to speak'.

MRW chains A metaphorical image is often verbalized by several MRWs and enriched with details. All MRWs that contribute to the same metaphorical image are annotated together and linked as a chain annotation, that is, an unordered set of token spans.

Of these MRW expressions, one can often be considered central, insofar as it best characterizes or names the metaphorical image. In example (3) above, *Strauß* 'bouquet' is the central expression, and *binden* 'tie' and *bunt* 'colorful' also contribute to the metaphorical image.

This central expression is the one that is given a specific label in the annotation that characterizes the whole metaphorical instance, while all of the other MRW expressions in the chain are only marked with the general label 'MRW'. Such specific labels are 'deliberate', 'grey area', 'revitalized', and 'extended'.


Das kann sich ein Land , dessen Bodenschätze in den Köpfen unserer Bevölkerung stecken , nicht leisten .

Wir haben das Rad also nicht neu erfunden , wir haben einfach ein Tesla oder ein BMW daraus gemacht .

Figure 1: Metaphor annotations in INCEpTION for examples (1) and (4).

Locality principle As a general rule, though not a strict requirement, the anchor should be determined in such a way that there is a direct syntactic dependency relation between the anchor and the central MRW, e.g. an MRW verb with its subject as the anchor, or an MRW noun with its modifier as the anchor. Very often a suitable anchor is easily found among the syntactically close expressions, since this direct relation is what allows the two expressions to better clash semantically.

Due to this close syntactic relationship between the MRWs and the anchor, an MRW chain usually only involves one clause or at most one sentence.[4]

Extended If a metaphorical image extends over several sentences, e.g. because it is introduced and then elaborated in subsequent sentences, we annotate the 'local' chains in each sentence individually. This can lead to there being no clear anchor in these subsequent sentences, therefore, in such cases, the MRWs may be annotated without an anchor. The otherwise deliberate MRW is then labeled 'extended'.

Examples (1) from above and (4) are two examples from our corpus. Figure 1 shows the annotation of these examples in INCEpTION.

(4) Wir haben das Rad also nicht neu erfunden, wir haben einfach ein Tesla oder ein BMW daraus gemacht.
'So we haven't reinvented the wheel, we've simply made a Tesla or a BMW out of it.'

[4] If a chain contains a pronoun, the pronoun is additionally linked to its antecedent via a coreference link. Such a chain is not extended to multiple sentences.

In example (1) there is a clear semantic clash between *Bodenschätze* 'natural resources' (= metaphorical, meaning 'intelligence', 'creativity', etc.) and *Köpfen* 'heads, minds' (= literal). Normally we only annotate nouns, verbs, and adjectives for metaphoricity. In this case, however, the preposition *in* 'in' plays an important role, so it is also annotated as MRW and included in the chain.

Example (4) contains what would ordinarily be considered a conventionalized metaphor: *das Rad neu erfinden* 'reinvent the wheel'. The second clause takes up part of the conventionalized image through the pronoun *daraus* 'out of it', which refers to *Rad* 'wheel' (see the coreference link in Fig. 1), and then elaborates upon this image, thereby revitalizing it. There is no clear clash in either clause and thus no anchor. However, the wider context makes it clear that *wir* 'we', the speakers, do not work in the automotive industry and are not talking about actually producing vehicles of any kind.

4 Data and results

Corpus The current corpus consists of the transcriptions of a total of ten TEDx Talks which were given in German on a range of different topics. Four of the texts have been doubly annotated and curated (see below). The texts are subject to licenses that permit free redistribution.[5] The corpus contains 20k tokens (averaging 1979.4 ±481.7 tokens per document). 1.35% of the tokens are deliberate metaphorical expressions, which shows that our guidelines successfully identify a significantly higher proportion of deliberate MRWs than previous studies. Of course, we cannot say what part the text type – TEDx Talks – has in this. Future work with annotations of other text types will have to show this.

Table 1 shows the distribution of the different types of MRWs. The numbers indicate the total number of chains per label, where a chain is categorized according to the label of its 'central MRW', such as 'deliberate', as well as the total number of tokens (including anchors) in each kind of chain.

Inter-annotator agreement Our validation corpus consists of four talks from the TEDx series. These texts were doubly annotated in their entirety according to our guidelines by two of the authors.

Type	# Chains	# Tokens
deliberate	85	264
extended	25	52
grey area	15	30
revitalized	20	46

Table 1: Distribution of different types of MRWs.

Our annotation scheme aims to capture more of the complexity of linguistic metaphor than previous annotation efforts, but the increased complexity of the annotation scheme brings with it both benefits and drawbacks. The information that is made available in the annotations is accordingly rich, but evaluating the reliability of the annotation effort becomes more difficult – in addition to the increased difficulty of the task itself.

To evaluate the reliability of the annotations, we employ the γ agreement measure (Mathet et al., 2015), specifically the implementation of Titeux and Riad (2021). This is a holistic agreement measure that determines the alignment between annotated units jointly with the measurement of disagreements in categorization.

We use a dissimilarity measure that takes into account the conceptual similarity between the category labels. For instance, metaphors that are labeled 'deliberate' can be considered more similar to those labeled 'grey area' than 'anchor'. As such, our dissimilarity measure will consider disagreement between 'deliberate' and 'grey area' to be less than between 'deliberate' and 'anchor'.

The γ statistic, calculated on these data with the parameters described above is 0.35, 0.43, 0.49 and 0.56 for each of the four evaluation texts, respectively. Especially considering the complexity of the phenomenon itself and the annotation scheme, these are promising results, which we expect could be improved in the future with further refinement of the annotation guidelines.

Acknowledgements

We are very grateful to the anonymous reviewers for their helpful and valuable comments.

Funded by/gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – SFB 1475 – Projektnummer 441126958.

[5] The TEDx Talks are part of this playlist: https://www.youtube.com/playlist?list=PLzPiBVgAHXijVDasy92X6lZkl0DvFgSEg, accessed 2024-02-26. Our annotations are based on the subtitles extracted from these videos.

References

Khalid Alnajjar, Mika Hämäläinen, and Shuo Zhang. 2022. Ring that bell: A corpus and method for multimodal metaphor detection in videos. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 24–33, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Beata Beigman Klebanov, Eyal Beigman, and Daniel Diermeier. 2008. Analyzing disagreements. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 2–7, Manchester, UK. Coling 2008 Organizing Committee.

Beata Beigman Klebanov and Michael Flor. 2013. Argumentation-relevant metaphors in test-taker essays. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 11–20, Atlanta, Georgia. Association for Computational Linguistics.

Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. 2018. Weeding out conventionalized metaphors: A corpus of novel metaphor annotations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1424, Brussels, Belgium. Association for Computational Linguistics.

Markus Egg and Valia Kordoni. 2022. Metaphor annotation for German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2556–2562, Marseille, France. European Language Resources Association.

Markus Egg and Valia Kordoni. 2023. A corpus of metaphors as register markers. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 220–226, Dubrovnik, Croatia. Association for Computational Linguistics.

Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations*. MIT Press.

Berenike Herrmann, Karola Woll, and Aletta Dorst. 2019. Linguistic metaphor identification in German. In Susan Nacey, Aletta Dorst, Tina Krennmayr, and Gudrun Reijnierse, editors, *Metaphor identification in multiple languages. MIPVU around the world*, pages 113–135. Benjamins, Amsterdam.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Natalie Parde and Rodney Nielsen. 2018. A corpus of metaphor novelty scores for syntactically-related word pairs. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22:1–39.

W. Gudrun Reijnierse, Christian Burgers, Tina Krennmayr, and Gerard J. Steen. 2018. DMIP: A method for identifying potentially deliberate metaphor in language use. *Corpus Pragmatics*, 2:129–147.

Gerard Steen. 2008. The paradox of metaphor: Why we need a three-dimensional model of metaphor. *Metaphor & Symbol*, 23(4):213–241.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification. From MIP to MIPVU*. Number 14 in Converging Evidence in Language and Communication Research. John Benjamins, Amsterdam.

Hadrien Titeux and Rachid Riad. 2021. pygamma-agreement: Gamma γ measure for inter/intra-annotator agreement in Python. *Journal of Open Source Software*, 6(62):2989.

A Sources of example sentences

The example sentences (1), (2) and (4) are taken from the following talks in the TEDx series:

- Example (1): *Schüler, Zukunft & Motivation*
- Example (2): *Vacuum-packed*
- Example (4): *Der Supermarkt der Zukunft*