A corpus of metaphors as register markers

Markus Egg and Valia Kordoni
Humboldt-Universität zu Berlin
Unter den Linden 6, 10117 Berlin
{markus.egg, evangelia.kordoni}@hu-berlin.de

Abstract

The paper presents our work on corpus annotation for metaphor in German. Metaphors denote entities that are similar to their literal referent, e.g., when Licht ‘light’ is used in the sense of ‘hope’. We are interested in the relation between metaphor and register, hence, the corpus includes material from different registers. We focussed on metaphors that can serve as register markers and can also be reliably indentified for annotation. Our results show huge differences between registers in metaphor usage, which we interpret in terms of specific properties of the registers.

1 Introduction

This paper presents ongoing work on annotating a German corpus for metaphor. We are interested in metaphors as register markers, therefore, the corpus includes material from a number of different registers. We annotate all the metaphors in the corpus but nevertheless put emphasis on a subgroup of metaphors which we believe can function as register markers.

The paper is structured as follows. After outlining the underlying theoretical concepts of metaphor and register and reviewing previous work, we introduce the corpus. Then we present the annotation results, which show huge differences in metaphor usage between the different registers in the corpus. These differences are then correlated with specific properties of registers.

2 Theoretical background

In this section, we introduce the two phenomena of metaphor and register, and the way in which they are related.

2.1 Metaphor

Metaphors involve a semantic shift of an expression in context. They refer to an entity that is similar to the referent of the literal interpretation of the metaphor. Theories reconstruct this similarity in different ways (for an overview see e.g. Ritchie, 2013). E.g., vorbeirasen ‘rush by’ in the temporal sense in (1) is metaphorical and shares with the literal, spatial interpretation the notion of a very fast development:

(1) das letzte Jahr ist nur so vorbei gerast ‘the last year has rushed by’

Metaphors can be assigned a degree of conventionalisation, from innovative to fully conventionalised. We distinguish conventionalised and non-conventionalised metaphors (see Section 3 for details), e.g., the metaphor in (1) is non-conventionalised.

A small number of metaphors like Blumen ‘flowers’ in (2) is signalled openly by ‘metaphor flags’, among them wie ‘like’ or praktisch ‘in effect’, but most metaphors are not.

(2) Wir sind wie Blumen praktisch, geerdet.
   ‘In effect, we are like flowers, earthed.’

In ‘extended metaphor’ (or ‘metaphor chains’), several metaphors in a discourse are based on the same kind of similarity (Reijnierse et al., 2020). E.g., once the word Licht ‘light’ is introduced as a metaphor for hope, other words related to light like anzünden ‘enkindle’ or Kerze ‘candle’ can emerge as metaphors for hope-related phenomena, too (as ‘introduce hope’ and ‘source of hope’, respectively).

Finally, ‘potential metaphor’ combines tokens of an expression with basic and metaphorical senses in the same discourse. E.g., in one of our texts the term dunkel ‘dark’ is used in its basic sense ‘without physical light’ before it is used metaphorically in the sense of ‘bad’. Potential metaphors typically participate in extended metaphor structures.

(3) in dem Dunkel, in dem Wurzelbereich bei
dem Weizen
‘in the dark zone, in the rhizosphere of the wheat’

(4) die dunkle Erde elterlicher Übermüdung
‘the dark soil of parental fatigue’

2.2 Register

Register refers to the influence of situational and functional context on intra-individual linguistic variation (Biber and Conrad, 2009). Systemic-Functional Linguistics (SFL) decomposes register into field, tenor, and mode (Halliday and Hasan, 1985). ‘Field’ refers to the nature of a linguistic interaction, including its subject matter and its purpose. ‘Tenor’ targets the participants, in particular, their statuses and social relationships. ‘Mode’ is about the role of language in the interaction, e.g., whether it is oral or literal, or a monologue or a dialogue.

When metaphors are alternatives to reference via literal expressions, they are optional ways of referring to an entity. This allows intra-individual variation in establishing reference to be influenced by and to influence the situational and functional embedding of a discourse, viz., register. Thus, metaphors can contribute to establishing a specific register or indicate compliance with it.

This relation of metaphors and registers is due to the fact that the function of a metaphor depends on the discourse it is part of (Goatly, 2011). For instance, the function of metaphors can be influenced by the relations between the interlocutors, in that peers strive to build and maintain rapport, whereas experts want to offer explanations to non-experts. Such differences can result in different realisations of the metaphors. For example, Deignan et al. (2013) report that metaphors in the form of a simile (‘A is like B’) are more likely in expert-non-expert communication than in exchanges between peers.

2.3 Which metaphors for register?

The perspective on metaphor as register marker (or as a marker for other phenomena) raises issues with the state of the art in metaphor annotation, as it was established by Steen et al. (2010) and introduced into computational approaches to metaphor by Shutova and Teufel (2010) and Shutova et al. (2013): In these approaches, all metaphorical expressions are annotated, irrespective of their degree of conventionalisation.

To be able to function as a marker for register, however, metaphors must be free choices in the linguistic system, whose optional use can then be reused to mark a specific register. In contrast, any metaphor whose use is necessitated by the language system cannot be employed for the purpose of register.

For example, in the description of temporal constellations it is often not possible not to use highly conventionalised spatial metaphors, e.g., to express the fact that one time span is located before or inside another one. I.e., these interpretations of prepositions belong to the lexicon as parts of polysemous sense structures. Since they are not created by a productive metaphorical interpretation and are obligatory irrespective of register, they cannot function as metaphorical register markers.

Steen (2015) comes to similar conclusions about highly conventionalised metaphors and focuses on ‘deliberate’ metaphors, i.e., those that are intended to be recognised as such by the recipient. We believe that this is the group of metaphors that is also relevant for the relation between metaphor and register.

However, deliberate metaphor is hard to define in formal terms (see e.g. Krennmayr, 2011 and Reijnierse et al., 2018), which raises doubts as to whether it can annotated with sufficient accuracy. Therefore we based our conclusions predominantly on deliberate metaphors that are recognisable with high accuracy in our corpus, viz., those with a metaphor flag, and non-conventional, extended, and potential metaphor.

3 Previous work

The interdependence between metaphor and register has been investigated for specific registers, e.g., academic discourse (Littlemore, 2001; Herrmann, 2015; Beger, 2015), fiction (Dorst, 2015), newspapers (Krennmayr, 2011) or educational discourse (Cameron, 2003). Functions of metaphors were correlated with SFL features of metaphors (Goatly, 2011; Steen et al., 2010). E.g., the latter claim that metaphor is used in informational registers (like news, fiction, or academic discourse) to express content to a much larger extent than in conversation. Berber Sardinha (2015) investigates the influence of metaphor-related features on register variation.

The group of Gerald Steen created and annotated the VU Amsterdam Metaphor Corpus (187,000...
words from the British National Corpus) with the four registers academic discourse, newspaper texts, fiction, and conversations (Steen et al., 2010).

Shutova and Teufel (2010) and Shutova et al. (2013) annotated a corpus of 13,700 words according to whether the words were used metaphorically or literally. They report different frequencies of metaphors for specific registers, in particular, a very low frequency of metaphor in spoken language. Bizzoni and Lappin (2018) compiled a corpus of 200 sets of metaphorical sentences and potential paraphrases (rated for their aptness). Zayed et al. (2020) created a corpus of 1,500 metaphorical verbs with a direct object.

Steen et al. (2010) developed detailed guidelines for the annotation of their corpus (later adapted to German in Herrmann et al., 2019). They define the context-based sense of an expression as a metaphor if it differs from another, more ‘basic’ sense of the expression (e.g., one which is more concrete or related to bodily action). These senses must be similar but not subsumable under a common hypernym, like in the case of the contextual temporal sense of vorbeirassen ‘rush by’ in (1). Senses are defined by suitable dictionaries; if both senses appear in the dictionaries, the metaphor counts as conventionalised, if only the basic sense does, it is regarded as non-conventional.

<table>
<thead>
<tr>
<th>subcorpus</th>
<th>hierarchical/equal</th>
<th>distant/close</th>
<th>oral/literal</th>
<th>dialogue/monologue</th>
</tr>
</thead>
<tbody>
<tr>
<td>speeches</td>
<td>E</td>
<td>D</td>
<td>L</td>
<td>M</td>
</tr>
<tr>
<td>sermons</td>
<td>H</td>
<td>C</td>
<td>L</td>
<td>M</td>
</tr>
<tr>
<td>commentaries</td>
<td>H</td>
<td>D</td>
<td>L</td>
<td>M</td>
</tr>
<tr>
<td>light fiction</td>
<td>E</td>
<td>C</td>
<td>L</td>
<td>M</td>
</tr>
<tr>
<td>debates</td>
<td>E</td>
<td>D</td>
<td>O</td>
<td>D</td>
</tr>
</tbody>
</table>

Table 1: SFL register properties of the subcorpora

Table 1 shows the distribution of SFL register properties in the corpus. We vary two dimensions of tenor, viz., hierarchy vs. equality and distance vs. closeness, and the two mode dimensions of dialogue vs. monologue and of spoken vs. written register. Following Koch and Oesterreicher’s (1994) distinction of conceptual literality vs. orality, speeches and sermons are classified as literal (they are prepared and fixed in advance), despite their oral presentation.

<table>
<thead>
<tr>
<th>subcorpus</th>
<th>reference</th>
<th>persuasion</th>
</tr>
</thead>
<tbody>
<tr>
<td>speeches</td>
<td>+</td>
<td>o</td>
</tr>
<tr>
<td>sermons</td>
<td>o</td>
<td>+</td>
</tr>
<tr>
<td>commentaries</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>light fiction</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>debates</td>
<td>o</td>
<td>+</td>
</tr>
</tbody>
</table>

Table 2: Biber dimension properties of the subcorpora

The subcorpora also represent the variation we expected along two important Biber (2009) dimensions (Table 2). For ‘situation-dependent vs. elaborated reference’ (how dependent is reference on the situational context), we expect that commentaries and speeches relate to concrete extralinguistic situations and individuals, whereas debates and sermons are more abstract deliberations, and fiction is highly detached from reality. Thus, the anticipated level of situation dependence for reference is low for fiction, medium for debates and sermons, and high for commentaries and parliamentary speeches. For ‘overt expression of persuasion’, the expected level is high for debates, sermons, and commentaries, moderate for speeches (whose influence on actual decision making in politics is usually quite low), and low for fiction.

4 Our approach

4.1 The corpus

To investigate the relation of metaphor and register, we have compiled a corpus that integrates a wide range of register variation. Its five parts (of eventually 30,000 words each) are parliament speeches from the German Parlamentsreden-Korpus (Blaette, 2017), news commentaries (the Potsdam Commentary Corpus; Stede, 2004), sermons, light fiction (written by amateurs for their peers), and debates from competitions of the organisation ‘Jugend debattiert’ (Kemmann, 2013).
Krippendorff’s (2011) alpha emerged as .89. The annotation includes a layer of syntactic structure, derived by the Stanza package (Qi et al., 2020), to allow the identification of syntactic constellations for analyses of their metaphorical potential in future work.

To distinguish degrees of conventionalisation of the metaphors, we also fell back on suitable lexical resources, in our case, the Duden dictionary and the Digitales Wörterbuch der deutschen Sprache:\footnote{www.duden.de and www.dwds.de}.

When the context-based sense of the expression qualifies as metaphorical according to definition of Steen et al. (2010) (see Section 3), we check if it is listed in at least one of the lexical resources along with the basic sense of the expression. If yes, the metaphor is classified as conventionalised, otherwise, we assume that it is non-conventionalised, like (1).

We created guidelines for the annotation, starting out from the guidelines of Steen et al. (2010) and Herrmann et al. (2019). The corpus and the guidelines will be made available to the research community after their finalisation. See Egg and Kordoni (2022) for a more detailed description of the guidelines.

## 5 Results

The results of our annotation are summarised in Table 3 (percentages are calculated for word tokens\footnote{Extended and potential metaphors as a whole are counted only once. The participating metaphors are then counted separately as conventionalised or non-conventionalised metaphors.}), showing clear differences in metaphor usage between the different registers. First, the level of conventionalised metaphors is high for speeches, medium for sermons, commentaries, and debates, and low for fiction. Also, metaphor flags are extremely rare in general, which parallels the results of Steen et al. (2010).

Potential metaphor is restricted almost exclusively to sermons. As soon as we omit sermons from consideration in the evaluation of our corpus, potential metaphor does not exhibit a correlation to register anymore (p = .33). Consequently, we omit it from further investigations into interdependencies between metaphor types and register properties.

Non-conventionalised and extended metaphors pattern similarly, occurring mostly in sermons and commentaries. We argue that this is due to the fact that these registers are highly persuasive. This correlation is less visible in the debates, which we put down to the time pressure of oral discourse, a conflicting factor impeding the creation of these types of metaphor. Table 4 summarises the counts of highly persuasive against the other registers and shows that the correlations are significant for these two types of metaphor.

Next, we investigated a potential interdependence between metaphoricity and the distinction in oral and literal discourse (summarised in Table 5). Our results first show that oral and literal discourse do not differ significantly for conventionalised and non-conventionalised metaphor. What is more, our oral register did not exhibit a significantly lower

<table>
<thead>
<tr>
<th>subcorpus</th>
<th>metaphor flags***</th>
<th>conventionalised metaphor***</th>
<th>non-conventionalised metaphor*</th>
<th>extended metaphor***</th>
<th>potential metaphor***</th>
</tr>
</thead>
<tbody>
<tr>
<td>speeches</td>
<td>0%</td>
<td>15.13%</td>
<td>.12%</td>
<td>.01%</td>
<td>0%</td>
</tr>
<tr>
<td>sermons</td>
<td>.02%</td>
<td>10.14%</td>
<td>.24%</td>
<td>.29%</td>
<td>.27%</td>
</tr>
<tr>
<td>commentaries</td>
<td>.05%</td>
<td>11.40%</td>
<td>.26%</td>
<td>.12%</td>
<td>.01%</td>
</tr>
<tr>
<td>light fiction</td>
<td>.04%</td>
<td>4.06%</td>
<td>.14%</td>
<td>.04%</td>
<td>0%</td>
</tr>
<tr>
<td>debates</td>
<td>0%</td>
<td>10.38%</td>
<td>.15%</td>
<td>.09%</td>
<td>0%</td>
</tr>
</tbody>
</table>

* = significant at p < .05; ** = significant at p < .01; *** = significant at p < .0001

Table 3: Metaphor counts for the subcorpora

<table>
<thead>
<tr>
<th>subcorpus</th>
<th>metaphor flags</th>
<th>conventionalised metaphor***</th>
<th>non-conventionalised metaphor*</th>
<th>extended metaphor***</th>
</tr>
</thead>
<tbody>
<tr>
<td>highly persuasive</td>
<td>.03%</td>
<td>10.88%</td>
<td>.23%</td>
<td>15%</td>
</tr>
<tr>
<td>medium or not persuasive</td>
<td>.02%</td>
<td>9.91%</td>
<td>.14%</td>
<td>.02%</td>
</tr>
</tbody>
</table>

Table 4: Metaphor counts for highly persuasive subcorpora
score for extended metaphor as well, even though this kind of metaphor is non-local in that it is based on more than one expression in the discourse, as two or more expressions have to share the same kind of metaphorical similarity. In contrast, conventionalised and non-conventionalised metaphor are local in that they are based on single expressions.

There is a difference between the extended metaphors in debates and in the other registers, however, which has to do with the fact that debates are dialogues that consist of comparatively short turns of different speakers: We found that many extended metaphors are the result of the collaboration of different speakers, in that one speaker introduces a metaphor with a specific kind of similarity and other speakers subsequently pick up this metaphor or use metaphors that exhibit the same kind of similarity.

To sum up, our results for the oral register of debates thus suggest that previous very low metaphoricity scores for oral discourse as in Steen et al. (2010) might not be related to orality in general but to the conversational nature of their data, which calls for further investigation of differences within oral registers.

As for individual registers, our data first suggest a mixed pattern for fiction, like in Reijnierse et al. (2019) in that it is low on conventionalised metaphors but occupies a middle position w.r.t. non-conventionalised and extended metaphor. At the same time, the register in the corpus that conveys the highest degree of register marking are sermons: they exhibit a high degree of non-conventional metaphors, also, extended and potential metaphor emerge as clear register markers for sermons.

6 Conclusion

We presented current work on a German corpus with different registers, which is annotated for metaphors. Future work will use the corpus to investigate the metaphoric potential of specific syntactic constellations (like verb-object and adjective-noun) and include metonymy as another register-sensitive phenomenon (Deignan et al., 2013; Littlemore, 2015).

Also, our results suggest that further research on oral registers is called for to delimit the actual interdependence between metaphor and the distinction between oral and literal registers. As a first step in this direction, we will include TEDx talks into our corpus, which complement the debates in that they are also oral but at the same time monologic and not persuasive. Other registers we plan to look into are sales talks and classroom interactions.

Limitations

In our study, we have argued for a correlation between forms of metaphor (non-conventional, extended, and potential) and persuasiveness. However, we are at this stage not yet in a position to rule out the competing hypothesis that there is a relation between metaphor and tenor in that metaphor correlates with a hierarchical difference between the interlocutors. Table 6 shows that this hypothesis would be significant for the corpus in its present form, which shows that the inclusion of further registers into the corpus is needed in order to distinguish between the competing hypotheses.

Ethics Statement

We took great care in the compilation of the corpus to include only material that can be published in
this form in order to be able to make the corpus available to the scientific community.

The debates consist of material produced by minors (16-18 years). In the corpus, we anonymised the names of the debaters throughout as ‘Speaker 1-4’. At the same time, we transcribed only debates that had already been made public on the Youtube channel of ‘Jugend debattiert’ (URL) in order to include only material whose publication had already been accepted by the respective speakers. At the same time, we contacted the spokesperson of ‘Jugend debattiert’ and got his consent on our activities as far as they include the debates.

The other material is either taken from already licenced corpora (for the parliament speeches and the commentaries) or has an appropriate CC license. Still, we contacted the authors to inform them about our project and to confirm their willingness to have their material included in our corpus.

Acknowledgment

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1412, 416591334.

References


