COLF Annotation Guidelines

October 25, 2018

Contents

1 The Annotation Task

The main goal of this annotation task is to annotate German verbal multiword expressions (VMWEs) - more specifically verbal idioms - and their literal counterpart. E.g. the expression *mit dem Feuer spielen* can be used idiomatically (meaning 'to be careless') or literally (meaning that someone plays with matches or something similar). So in essence it is a kind of word sense disambiguation (WSD) task. In our case, it is a lexical sample task which means that we have a pre-selected set of expressions whose instances will be mapped to a set of readings. For the expression *über Bord gehen*, for example, we extracted 115 instances which have to be annotated according to their readings. To do that for this and all the other expressions please follow the subsequent instructions.

2 Annotation Instructions

1. Identify the MWE in the sentence. The MWE to identify will be the name of the annotation file (which is also the name of the table sheet, i.e. every annotation file will have its own table sheet). Every annotation file will include instances and their contexts of only one

MWE type. E.g. the file for *mit dem Feuer spielen* will only contain instances of *mit dem Feuer spielen*.

2. Annotate the instances according to four labels: **literal**, **idiomatic**, **undecided** and **both**. In the following we will explain what those labels mean:

   - **literal**: We basically equate literality with compositionality. Compositionality denotes the property that the meaning of an expression is determined by the most basic meanings of its components without any form of figuration involved. Consider the sentence *Maria deckt ihre Karten auf* for example. In the literal sense, Maria actually has playing cards which she turns around. In the figurative (idiomatic) sense, Maria reveals her intentions.
   - **idiomatic**: According to Baldwin and Kim (2010) there are different forms of idiomaticity: lexical, syntactic, semantic, pragmatic and statistical. But a lot of times in the MWE literature idiomaticity is used as a synonym for what Baldwin and Kim would call semantic idiomaticity. We will adopt this usage, thus in the context of the annotation process, the label idiomatic means semantically idiomatic. Semantic idiomatictiy is closely related to compositionality, since it denotes the property of an expression that it is not possible to fully derive its meaning by considering the meanings of its components. So semantic idiomaticty describes a lack of compositionality. The MWEs which are the target of this annotation, verbal idioms, are usually a prime example of this lack of compositionality. E.g. you cannot derive the meaning 'to be careless' from the semantics of the components of *mit dem Feuer spielen*, even though the figurative process that underlies the formation of this meaning is rather transparent (someone who plays with fire is careless). A more opaque example would be *auf den Arm nehmen*. The derivation of the meaning 'to kid somebody' from the action of picking someone up is far less obvious. And the English expression *shoot the cat* ('vomit') is an example for a MWE where the figurative process behind the meaning is completely opaque.
   - **undecided**: This label is for cases in which it is not possible to decide whether the target expression is literal or idiomatic. E.g. this is notoriously difficult for the MWE *sich die Haare raufen*, because the gesture denoted by its literal reading ('to scuffle your

hair') is an expression of the feeling ('to be annoyed', 'to be worked up') denoted by the idiomatic expression, i.e. a person that is worked up tends to scuffle her/his hair. We will not include *sich die Haare raufen* in the corpus, but there will be instances of other MWEs where the context sometimes will not be informative enough to make a conclusive decision.

- **both**: While the label *undecided* means that there is only one possible reading, but it's not feasible to decice which, there is also the phenomenon of the two readings being activated at the same time: "Vom Sockel geholt und auf dem Rücken liegend, werde er gut sichtbar aufgebahrt der Betrachtung erst richtig zugänglich. [...] Zugleich werde Lenin der Charakter eines Heiligen genommen, und er komme unter das Volk, jedem unmittelbar zugänglich. "Wer möchte, könnte ihm den Kopf waschen, ihm mal auf den Zahn fühlen oder ihn gar auf den Arm nehmen, wobei sich feststellen ließe, daß das garnicht so leicht ist." Here the author plays deliberately with the ambiguity of different MWEs which makes a binary decision between literal and idiomatic impossible. But caution: If you consider the sentence *Nicht nur Houdini hielt während seiner Entfesslungsnummer unter Wasser den Atem an, sondern auch die Zuschauer, die diesem denkwürdigen Spektakel beiwohnten* you can see that it also contains both readings. But in contrast to the former example there is one important difference: The sentence contains an ellipsis. Although the surface form of the sentence only contains one instance of *den Atem anhalten* there are actually two. The first one has Houdini as its subject and is the literal version of the expression, and the second one has *die Zuschauer* as its subject and is ommited. In these cases the annotation should stay on the surface which means that only the "visible" instance should be annotated. So the Houdini example should be annotated with the label *literal* and not with the label *both*.

3. Please annotate whether you needed more context than one sentence to determine the reading of the expression. This is annotated in the +1 Sentence-column of the table sheets. If you only need one sentence, please choose "No" if you need more than one sentence, please choose "Yes".

IMPORTANT: Every annotator will annotate every file separately.

Thus you should not discuss your annotations with each other.

3 References

Baldwin, T., & Kim, S. N. (2010). Multiword expressions. Handbook of natural language processing, 2, 267-292.