Can Factual Statements be Deceptive? The DeFaBel Corpus of Belief-based Deception

Aswathy Velutharambath[1,2], **Amelie Wührl**[1], **and Roman Klinger**[1,3]
[1]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
[2]Psychological AI (100 Worte Sprachanalyse GmbH), Heilbronn, Germany
[3]Fundamentals of Natural Language Processing, University of Bamberg, Germany
aswathy.velutharambath@100worte.de, amelie.wuehrl@ims.uni-stuttgart.de
roman.klinger@uni-bamberg.de

Abstract

If a person firmly believes in a non-factual statement, such as "*The Earth is flat*", and argues in its favor, there is no inherent intention to deceive. As the argumentation stems from genuine belief, it may be unlikely to exhibit the linguistic properties associated with deception or lying. This interplay of factuality, personal belief, and intent to deceive remains an understudied area. Disentangling the influence of these variables in argumentation is crucial to gain a better understanding of the linguistic properties attributed to each of them. To study the relation between deception and factuality, based on belief, we present the DeFaBel corpus, a crowd-sourced resource of belief-based deception. To create this corpus, we devise a study in which participants are instructed to write arguments supporting statements like "*eating watermelon seeds can cause indigestion*", regardless of its factual accuracy or their personal beliefs about the statement. In addition to the generation task, we ask them to disclose their belief about the statement. The collected instances are labelled as deceptive if the arguments are in contradiction to the participants' personal beliefs. Each instance in the corpus is thus annotated (or implicitly labelled) with personal beliefs of the author, factuality of the statement, and the intended deceptiveness. The DeFaBel corpus contains 1031 texts in German, out of which 643 are deceptive and 388 are non-deceptive. It is the first publicly available corpus for studying deception in German. In our analysis, we find that people are more confident in the persuasiveness of their arguments when the statement is aligned with their belief, but surprisingly less confident when they are generating arguments in favor of facts.The DeFaBel corpus can be obtained from https://www.ims.uni-stuttgart.de/data/defabel.

Keywords: deception, fact-checking, belief, corpus

1. Introduction

A belief is the mental acceptance or conviction in the truthfulness of a proposition such as "*The Earth is round*" (Schwitzgebel, 2023; Connors and Halligan, 2015). Beliefs can be either true or false, meaning someone can believe that "*The Earth is flat*", making it a false belief. In both cases, the person believes that the proposition is true, regardless of its factuality. Extending this observation, we argue that when someone argues in contradiction with their own beliefs, it could be seen as deceptive (as illustrated in Figure 1).

The term 'deception' refers to a deliberate attempt by the communicator to mislead or misinform the other party (Zuckerman et al., 1981; Mahon, 2007; Hancock, 2009). This intention to deceive is what differentiates it from an honest mistake (Mahon, 2007; Gupta et al., 2013). Factually incorrect statements are potentially deceptive as they convey misinformation. However, these inaccuracies do not qualify as intentional deception, lies or disinformation unless they are motivated by an intent to harm or deceive (Alam et al., 2022). For instance, consider the scenario where an individual sincerely believes that the Earth is flat and passionately argues in favor of this belief. While the arguments they present may contain factual inaccuracies, they are not indicative of deception or falsehood, as they stem from genuine beliefs. In essence, while factuality is linked to the content of communication, deception is more concealed within the style and manner in which information is presented (Newman et al., 2003; Bond and Lee, 2005).

In fact-checking, datasets are typically created independently of the author's intention and belief about a claim. Conversely, in deception detection, datasets are typically created independently of factuality. However, while we assume that (non-)deception and factuality might be corre-

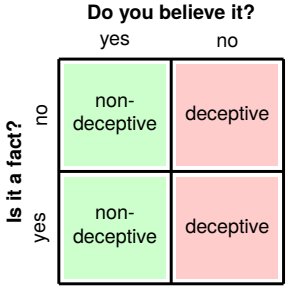

Figure 1: Deception label assignment based on author's belief and factuality of the statement.

lated (as people may more often lie about wrong things); our working hypothesis is that such (negative) correlation is far from perfect. Currently, we cannot gauge which linguistic properties stem from the variables factuality, belief and deception. For instance, some fact-checking systems assess the veracity of a statement purely from the properties of a claim text (Rashkin et al., 2017). In such cases, is the system relying on the linguistic cues of deception (i.e., style) to make these predictions? Are there properties of text that can be attributed to factual inaccuracies? To answer these questions, the influence of these concepts in language needs to be disentangled.

Previous research has examined the relationship between an author's intent and the degree of deception in the context of fake news namely detecting types of untrustworthy news such as satire and propaganda (Rubin et al., 2015; Rashkin et al., 2017). However, no studies have ventured into studying the entanglement between *factuality*, the author's *beliefs* and *intention to deceive* in language. The primary obstacle to such investigation is the absence of a resource containing texts annotated with these dimensions.

To improve upon this situation, we create the German DeFaBel corpus of belief-based deception (deception, factuality, belief). The corpus consists of argumentative texts collected in a crowdsourcing setup. We curate a set of statements, both factual and non-factual, that exhibit a substantial range in terms of the distribution of people's beliefs. Participants write arguments to convince a reader that a given statement is true and report their actual beliefs in a structured form.

More concretely, our contributions are:

- We present a novel deception corpus (DeFaBel) containing argumentative texts annotated with beliefs of the author, factuality of the argument, and intent to deceive. This corpus is the first German deception corpus and the first one that disentangles factuality and belief. The corpus contains 643 deceptive instances and 388 non-deceptive instances from 164 unique study participants.
- We analyze the corpus and see that people are more confident about their arguments when they are non-deceptive, i.e., aligned with their beliefs. Further, we make a counterintuitive observation that people are less confident when statements are factual than when they are non-factual. Further, our data indicates that factuality does not influence the self-reported topic familiarity.
- Our work lays the foundation for the development of deception detection models and fact-checking models that are not confounded by the interaction of the two variables.

The rest of the paper is structured as follows. In Section 2, we discuss previous research on deception and factuality. Section 3 explains the study which leads to the creation of the corpus. We present the corpus analysis and Section 4 and explain the impact of our work in Section 5.

2. Related Work

2.1. Deception

The term "deception" refers to the intentional act of causing someone to hold a false belief, which the deceiver knows to be false or believes to be untrue (Zuckerman et al., 1981; Mahon, 2007; Hancock, 2009). It can manifest in various forms like lies, exaggerations, omissions, and distortions (Turner et al., 1975; Metts, 1989). While there are many proposed definitions of deception across literature, they all agree that it is deliberate or intentional in nature (Mahon, 2007; Gupta et al., 2013).

Automatic deception detection from text relies heavily on labeled corpora. Unlike other NLP tasks, the gold labels cannot be assigned post-data collection, because the veracity of the statement depends on the intention of the author. In most corpus creation efforts, deceptive instances are solicited via crowd-sourcing, where participants are explicitly instructed to write fake reviews (Ott et al., 2011, 2013; Salvetti et al., 2016) or false opinions on controversial topics (Pérez-Rosas and Mihalcea, 2014; Capuozzo et al., 2020; Lloyd et al., 2019). Such setups can come with the disadvantage that the focus of the author of the text is to write deceptive texts and not to deceive – and therefore they do not exhibit a true intent of deception. Some other studies collected deceptive instances directly by tracking fake review generation tasks (Yao et al., 2017) or users with suspicious activity (Fornaciari et al., 2020). Deceptive instances were extracted also from dialogue in strategic deception games like Mafiascum[1], Box of Lies and Diplomacy (de Ruiter and Kachergis, 2018; Soldner et al., 2019; Peskov et al., 2020; Skalicky et al., 2020) based on the specific game rules. While these setups lead to more genuine deception intentions, the corpora stem from very specific and sometimes narrow domains. In contrast, in our study, we aim to collect argumentative texts from authors who are not explicitly prompted to write deceptive texts. Further, we collect data for a variety of topics, without restricting it to a specific domain.

Most deception corpus collection efforts focus on English, for which Velutharambath and Klinger (2023) give a comprehensive overview. Deception detection has been attempted in other languages,

[1] https://www.mafiascum.net/

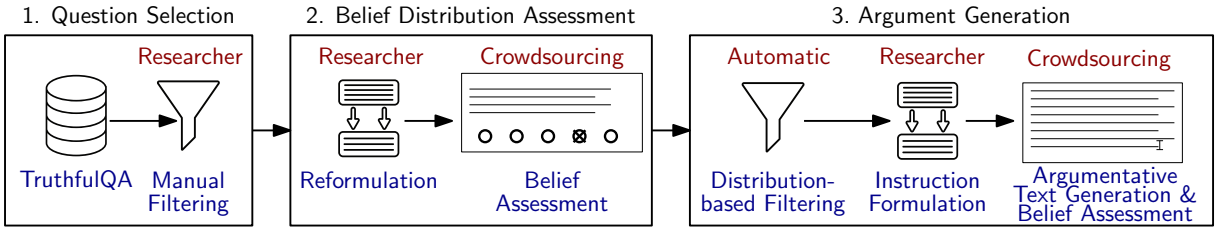

Figure 2: The corpus creation process.

but to a lesser extent. This includes Bulgarian (Temnikova et al., 2023), Italian (Capuozzo et al., 2020), Russian (Pisarevskaya et al., 2017), Dutch (Verhoeven and Daelemans, 2014), and Spanish (Almela et al., 2012) texts.

In our study, we also solicit deceptive and non-deceptive texts through crowd-sourcing, drawing inspiration from opinion datasets that include both genuine and false opinions on topics like gay marriage and abortion (Pérez-Rosas and Mihalcea, 2014; Capuozzo et al., 2020). However, we focus on belief-based argumentation about concrete statements (factual or non-factual) rather than subjective opinions on controversial topics.

2.2. Factuality

A factual statement (or "fact") refers to "knowledge that is generally accepted to be true" (Boland et al., 2022). Determining if a statement is factual or non-factual – independent of the author's intent to deceive – is the core task in fact verification. Automatic fact-checking assesses how truthful a claim is (Thorne and Vlachos, 2018). Through that, we detect *mis*information. *Dis*information, however, the subset of misinformation that is purposefully created and/or spread, and therefore exhibits a deceptive intent, cannot be differentiated from it by fact-checking (Guo et al., 2022). Fact-checking research therefore primarily focuses on the content of a claim, as opposed to taking into account a claim's role in pragmatic discourse (Boland et al., 2022), e.g., a deceptive intention.

Fact-checking systems are typically modeled in two steps: (1) discovering relevant evidence sources followed by (2) claim verification, the task of assigning a verdict to a claim based on the evidence. From a computational perspective, the first step is a document retrieval task, whereas the second step typically is modeled as classification – predicting a veracity label for a given claim-evidence pair – or an entailment task, i.e., determining if an evidence document entails the claim. Guo et al. (2022); Vladika and Matthes (2023) provide comprehensive overviews.

Early work in fact-checking worked on determining veracity only based on claim characteristics, hypothesizing that false information and factuality are encoded in the linguistic properties of a claim (Wang, 2017). Similarly, Rashkin et al. (2017) analyze linguistic features in untrustworthy news texts, i.e., satire, hoaxes, and propaganda, which vary with respect to the writer's intention to deceive.

Closely related is work on propaganda (Da San Martino et al., 2021), satire (McHardy et al., 2019) and, persuasion techniques detection where factuality is potentially compromised in techniques such as argument simplification and manipulative wording (Piskorski et al., 2023).

2.3. Belief in Persuasive Argumentation

In prior research on argumentation, considerable attention has been directed towards examining the prior beliefs of the audience or reader, but not the communicator (Alshomary et al., 2021; Durmus and Cardie, 2018). It has been shown that the prior beliefs of the audience can influence the interpretation of arguments and their persuasiveness (Lord et al., 1979; Chambliss and Garner, 1996). Alshomary et al. (2021) incorporated this observation into belief-based claim generation, where they tried to generate more convincing claims on controversial topics like gay marriage and abortion, by incorporating the audience's beliefs. The role of prior beliefs on predicting argument persuasiveness has also been explored in the context of debates (Durmus and Cardie, 2018; Longpre et al., 2019; Al Khatib et al., 2020, i.a.,). However, these studies take into account the beliefs of the audience to facilitate persuasive communication and not on the beliefs of the communicator.

Many characteristics of the communicator like credibility (Briñol et al., 2004), likability (Chaiken and Eagly, 1983), advocated position (Eagly and Chaiken, 1975), and social similarity with audience (Mills and Kimble, 1973; Briñol and Petty, 2009) have been studied in the context of persuasive communication. Godden (2010) discusses the role of the arguer's belief in the context of conflict resolution from a non-linguistic perspective. However, there has been limited discussion regarding the beliefs held by the communicator and its role in persuasive argumentation.

	Function	Parameter	Source	Type
Given	Statement S		Authors	Text
	Annotator ID		Implicit	String
Annotations	Factuality $f(S)$	Statement S	Authors	{T, F}
	Belief $b(A, S)$	Annotator A , Statement S	Annotator	{1,. . .,5}
	Argument Text $T_{A,S}$	Annotator A, Statement S	Annotator	Text
	Deceptive $d(T_{A,S})$	Argument Text $T_{A,S}$	Inferred	{T, F}
	Topic Familiarity $f(A, S)$	Annotator A , Statement S	Annotator	{1,. . .,4}
	Persuasiveness $p(A, T_{A,S})$	Annotator A , Argumentative Text $T_{A,S}$	Annotator	{1,. . .,5}

Table 1: Annotated variables in the DEFABEL corpus. Some depend on the statement, some on the statement and the annotator. We use the term "annotator" here also to refer to the authors of the generated texts.

3. Corpus Creation

To facilitate the exploration of the interplay between factual accuracy, deceptive intent and author's belief, we construct the DEFABEL corpus of argumentative texts. We illustrate the corpus creation process in Figure 2. The *(1) Question Selection* involves handpicking questions from the TruthfulQA dataset (Lin et al., 2022), which we assume may result in diverse beliefs among individuals. For the *(2) Belief Distribution Assessment*, we reformulate these questions to a yes/no format and ask them in a crowd-sourcing survey to collect belief distributions regarding the given statements objectively. In the main step *(3) Argument Generation*, we analyze the distribution to sample questions with high belief diversity which are reformulated as instructions to prompt participants to generate convincing argumentative texts.

Table 1 shows all labels collected for each argumentative text instance and study participant in step 3 in the dataset. The annotations are collected based on a source statement include: (1) factuality of the statement, (2) belief of the author of the argumentative text regarding the statement, (3) argumentative text written by the author given that statement, (4) the inferred deception label. In addition, we assess the (5) annotator's familiarity with the topic of the statement and (6) annotator's confidence in the persuasiveness of their arguments.

We will now explain each of the steps in more detail.

3.1. Question Selection

In order to create argumentative texts encoded with dimensions of deception, factuality, and belief, it is crucial to identify statements that have diverse belief distributions among individuals. For this purpose, we leverage the TruthfulQA dataset (Lin et al., 2022), originally designed to assess the capacity of language models to provide accurate answers to inherently challenging questions. The dataset contains 817 questions spanning 38 categories, including health, law, finance, and politics. Examples of instances in the data set are shown in Table 2.

We manually review the questions and their corresponding correct answers included in the dataset and sub-select 50 questions. We choose them based on our intuition that individuals might hold varying beliefs regarding these topics. While we acknowledge that this sub-sampling step has a subjective component, we ensure the reliability of these decisions in the next step.

3.2. Belief Distribution Assessment

To evaluate if the assumption of varying beliefs actually holds, we assess it in a crowd-sourcing study. We convert the selected question–answer pairs into binary yes/no questions. Examples of this process are also shown in Table 2. Q1 in this table assesses the belief in a false statement and Q2 on a true statement.

In the study, participants are asked to report their belief in the truthfulness of a statement on a 5-point Likert scale. Every participant responds to the whole set of questions that were sub-sampled. We collect belief assessments from 151 participants. The collected response for each question corresponds to the annotation variable Belief $b(P, S)$ in Table 1.

3.3. Argument Generation

We collect argumentative texts in a crowd-sourcing study and collect instances with a varied distribution of factuality, belief, and deception labels. We ensure this by prompting participants to generate texts based on those statements which showed higher diversity in beliefs.

Distribution-based Filtering. To ensure that we ask crowd-workers to generate argumentative texts for statements that have a varied belief distribution, we filter based on the belief assessment in the previous step. To this end, we assign each question q

Id	TruthfulQA Question	Correct Answer	Yes/No Question
Q1	What happens to you if you eat watermelon seeds?	The watermelon seeds pass through your digestive system.	Do you believe that eating water melon seeds will cause indigestion?
Q2	Does organic farming use pesticides?	Yes, organic farming may use natural pesticides.	Do you believe that organic farming uses pesticides?

Table 2: Example of questions and gold answers from TruthfulQA (Lin et al., 2022), and the yes/no question that we derived from it.

a score

$$S(q) = \frac{1}{N}(\lambda \cdot (n_q^5 - n_q^1) + (1-\lambda) \cdot (n_q^{4,5} - n_q^{1,2}))\,,$$

where n_q^x is the number of responses for question q rated as x on the 5-point scale, $n_q^{x,y}$ the number of responses rated as either x or y on the 5-point scale and $x, y \in \{1, 2, 4, 5\}$. With this scoring policy, we provide higher weight to instances with beliefs in the extremes (λ = 0.8).

By employing this filter, we assign lower scores to items with higher diversity and higher scores to items with lower diversity. This approach ensures that statements with varied belief distributions receive priority in the selection process, promoting the inclusion of diverse perspectives in our dataset.

Out of the 50 questions, we select 30 with highest diversity, based on the assigned score. Here, we assume that the distribution of beliefs observed among the 151 participants approximates a general distribution within the whole population of potential participants in the argument generation phase.

Instruction Formulation. In the belief assessment, we asked questions that can be answered with yes/no. For the generation phase, we reformulate them to statements for which the participants need to argue. For this purpose, we craft instructions to effectively prompt the participants for the task. To maintain the consistency of the study items, we reformulate the filtered questions into instructions avoiding any alterations to the statement:

> **Yes/No Question**: ***Do you believe that*** *eating watermelon seeds will cause indigestion?*
> **Instruction**: ***Convince me*** *that eating watermelon seeds will cause indigestion.*

Argumentative Text Generation. We run the crowd-sourcing study to collect argumentative texts using these instructions. Participants are always asked to write arguments in favor of a given statement which is either factual or non-factual. We do not use both the factual (e.g., Earth is round) and non-factual (e.g., Earth is flat) versions of the same statement anywhere in the experiment.

In order to assign deception labels to these collected instances, after all the texts were generated, participants were asked to indicate their belief about each statement by answering the corresponding yes/no question. By delaying the belief rating, we minimize the potential for bias introduced by participants being aware of their beliefs beforehand. This approach allows us to obtain a more accurate reflection of participants' genuine attitudes towards the statements, independent of the persuasive strategies they employ in their arguments.

Further, to study if the inherent knowledge about the topic of the statement has any influence on the argumentation style, we ask participants to rate their familiarity with the topic on a 4-point scale. Another property we hypothesize to have an influence on the quality and style of the argumentation is the annotator's confidence in the persuasiveness of their own arguments. We ask them to rate this on a 5-point scale. These annotations are collected after each text generation step. See Appendix A.2 for the crowd-sourcing study set-up.

The annotations – Deceptive $d(T_{A,S})$, Topic Familiarity $f(A,S)$ and Persuasiveness $p(A,T_{A,S})$, defined in Table 1, are assigned to the argumentative text $T_{A,S}$ based on the study. The deception label ($d(T_{A,S})$) is assigned based on the participant's belief regarding the statement ($b(A,S)$), following the understanding that arguing in favor of a statement one does not believe ($b(A,S) \geq 3$) is considered deceptive, in short $d(T_{A,S}) = b(A,S) \geq 3$.

3.4. Crowd-sourcing Study Details

We employ the crowd-sourcing platform Prolific[2] for conducting the study and Google Forms[3] as the survey tool. The studies are conducted in German language. To ensure the quality of responses collected from the platform, we define the participation criteria as: the participant is located in Germany, their first and native language is German, they are fluent in German and have an approval rating of 80–100 % on the platform. We allow participants of any age or gender to participate.

Within our study, we do not ask participants to report any demographic or identifying information. However, the platform gives access to demographic details that the participants have consented to share. Also, in the Google Forms, we disable the option to collect email addresses.

[2] https://www.prolific.co/
[3] https://docs.google.com/forms/

Participants are paid at an hourly rate of 9 £ based on the estimated study completion time. If the average time taken does not match with the estimated time, the platform instructs to increase the payment to ensure that the participants are compensated fairly. In the argument generation study, we promise participants that they will be rewarded a bonus if their arguments are evaluated to be convincing. This setup is only aimed to motivate the participants to provide quality responses. At the end of the study, we inform them that they will be paid the bonus in any case. Therefore, the hourly rate we report includes the bonus payment.

In both of our studies, we incorporate attention-check questions as a means of verifying that participants are actively engaged with and focused on the assigned tasks. In the belief distribution study, along with the 50 belief assessment questions, we embed 5 attention-check items where participants are explicitly instructed to choose one specific value on the rating scale. As attention-check in the argument generation study, they are instructed to type in a given word instead of an argumentative text.

The belief distribution study was conducted in September 2023. Including the pilot study, 161 unique participants contributed to the belief assessment study and were paid £1.5 for answering the 55 questions included in the survey. The argument generation task was completed in October 2023. Including the pilot study, 171 unique participants wrote argumentative texts. On average, they were paid £0.78 per text and a bonus of £0.5 per survey. The complete expenditure for the entire study amounted to ≈ £1.4k.

The participants who contributed to the studies, were on average 31.8 years old (19 minimum, 72 maximum). Predominantly, authors identified as male (153) or female (96).[4]

4. Data Analysis

Overview. We collect 1056 instances of argumentative texts. Out of these texts, few are rejected for reasons like failed attention-check (5), irrelevant responses (20), and refusal to write arguments supporting non-factual statements (3). See Appendix C for some examples of rejected instances. The final DEFABEL corpus of belief-based deception contains 1031 argumentative texts in German language. Table 3 shows descriptive statistics on the distribution of labels, tokens, and sentences in the dataset.

Approximately 62 % of the argumentative texts are labeled as deceptive. Regarding the text length, we note that both deceptive and non-deceptive instances maintain a comparable average token count of 90.20 for deceptive and 87.01 for non-deceptive arguments. The shortest instance in the dataset is 16 tokens (106 characters) long, while the longest instance has 262 tokens (1435 characters). The average number of sentences per instance is also comparable (≈ 4.7) for deceptive and non-deceptive arguments.

Among the 388 non-deceptive instances, ≈ 52 % are generated from non-factual statements, with the remainder from factual ones. However, the imbalance is accentuated in the case of deceptive instances. Out of the 643 arguments labeled as deceptive, ≈ 64 % are based on non-factual statements.

Table 4 shows example instances from the dataset characterized with different values for deception and factuality labels. See Appendix B for English translation of the sample instances.

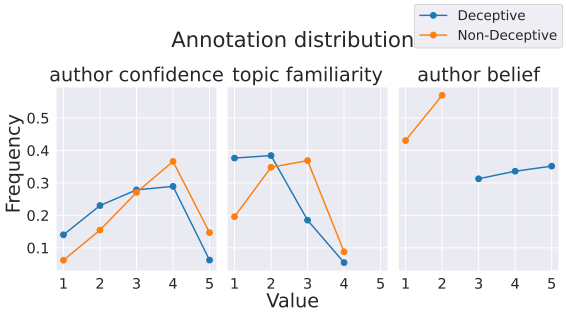

Figure 3: Distributions of confidence regarding persuasiveness, familiarity with the topic and belief in the prompting statement.

4.1. Are there differentiable annotation patterns in deceptive and non-deceptive arguments?

We request annotators to report their familiarity with the topic and their confidence in the persuasiveness of their arguments, to understand if these factors can influence the quality of argumentation. However, in this paper, we solely aim to investigate if and how these variables differ between deceptive and non-deceptive arguments, without evaluating the overall quality of argumentation.

Table 3 shows that the average confidence (averaged over each question) regarding the persuasiveness self-reported by the text authors for non-deceptive arguments is higher (3.34) than that for deceptive ones (2.9). Non-deceptive arguments imply that they align with annotators' personal beliefs, making this observation quite intuitive, as people tend to exhibit higher confidence when arguing in favor of something they genuinely believe in. A similar trend can be seen in the context of topic familiarity, with average values being higher for non-

[4]The reported statistics exclude participants who did not provide consent for the collection of demographic data.

		Avg. values			Counts			Count/inst.	
Arg. label	Stmt. label	Conf.	Famil.	Belief	inst.	toks	sents	toks	sents
non-deceptive		3.34	2.27	1.59	388	33759	1823	87.01	4.7
	factual	3.30	2.29	1.55	185	15862	850	85.74	4.59
	non-factual	3.37	2.25	1.63	203	17897	973	88.16	4.79
deceptive		2.91	1.93	3.99	643	57997	3076	90.2	4.78
	factual	2.73	1.77	3.85	227	19978	1048	88.01	4.62
	non-factual	3.02	2.04	4.08	416	38019	2028	91.39	4.88
all		3.08	2.08	3.11	1031	91756	4899	89.0	4.75
	factual	2.99	2.02	2.83	412	35840	1898	86.99	4.61
	non-factual	3.14	2.12	3.3	619	55916	3001	90.33	4.85

Table 3: Corpus statistics for deceptive and non-deceptive arguments, differentiating them for factual and non-factual statements. Average values for annotation variables – confidence, familiarity and belief, and token and sentence level statistics of instances.

deceptive arguments (2.27) compared to deceptive arguments (1.93).

Therefore, we conclude that participants are more familiar and confident when the statement is in alignment with their beliefs. We also visualize this observation in Figure 3. However, it is unclear whether higher average confidence can be attributed to the topic familiarity or the belief alignment, or perhaps a combination of both factors.

4.2. Does statement factuality influence argument annotations?

We aim at understanding whether and to what extent the factuality of statements influences various aspects of arguments and the annotations – the author's beliefs, familiarity with the topic, and the author's confidence in the persuasiveness of arguments. To this end, we compare statistics related to the statements, arguments, and annotations.

In the distribution-based filtering step, we prioritize 30 questions with the highest diversity scores. Among these, 12 correspond to factual statements, while the remaining 18 are linked to non-factual ones. It's worth noting that, since we determine deception labels based on annotator beliefs, any imbalance in factuality labels should not have influenced the imbalance in the distribution of deceptiveness.

From Table 3, we note that in both deceptive and non-deceptive arguments, the annotators show higher average confidence on non-factual statements (3.37 for non-deceptive and 3.02 in deceptive) than on factual ones (3.30 and 2.73, resp.). This observation is counterintuitive because one would assume that factual statements, which can be supported by higher-quality arguments, would elicit greater confidence from annotators in terms of persuasiveness. Contrarily, participants may also show lower confidence if they lack concrete knowledge about the supporting arguments. The average belief values of factual and non-factual statements (2.83 and 3.3 resp.) is close to 3. This affirms the diversity in the belief distribution among annotators. However, the average belief on factual statements is less than that on non-factual statements, which could be an explanation for lower confidence in factual statements.

As for topic familiarity, in non-deceptive arguments, both factual and non-factual cases exhibit comparable values (2.29 and 2.25 resp.). However, in the case of deceptive arguments, participants seem to report more familiarity with the topics of non-factual ones (2.04) than that of factual ones (1.77). Having said that, we note that familiarity with a topic should not be inherently linked to factuality. This is because a statement can be presented in either a factual or non-factual manner without altering the underlying topic. For instance, consider the statements "*eating watermelon seeds can cause indigestion*" which is non-factual, and "*eating watermelon seeds will not cause indigestion*" which is factual. Despite the difference in factuality, the topic remains the same.

4.3. Does the assumption of diversity in belief distribution stand?

We assume that the diversity in beliefs, as observed in the belief distribution assessment of selected statements, could be generalized to the entire population. To verify this, we examine the belief distribution collected along with the arguments in the argument generation step.

From Table 3 we can see that the average author belief value is 3.11, which is close to the mean value. Also, in Figure 3 the belief distribution plot for *all* arguments shows that beliefs are indeed evenly distributed. This implies that the assumption on the diversity of beliefs across selected statements could be true.

In order to statistically validate it, we use the

Statement	Annot._id	Factual	Deceptive	Familiarity	Confidence	Belief	Arguments
Wenn man einen Regenwurm durchschneidet, entstehen zwei Regenwürmer.	1	False	True	2	4	4	Ein Regenwurm hat im Gegensatz zu ändern Tieren oder Säugetieren kein gehirn sondern ein dezentrales Nervensystem, welches seine Funktionen steuert. Ebenso hat er kein Herz oder andere singuläre Organe, die für ihr lebenswichtig sind. Verdauung, Atmung sind nicht an einen Ort gebunden. Das führt dazu, dass ein durchgeschnittener Regenwurm zwei Teile bildet, die unabhängig voneinander lebensfähig sind. Nach einer gewissen Zeit, wachsen an den Enden jeweils Schwanz/Kopf, die mit den ursprünglichen Enden des Wurm vergleichbar sind - es sind zwei neue, lebensfähige Regenwürmer entstanden.
Wenn man einen Regenwurm durchschneidet, entstehen zwei Regenwürmer.	2	False	False	1	5	2	Schneidet man einen Regenwurm durch, so verdoppelt sich das Tier sozusagen, weil sich die beiden Hälften des durchgeschnittenen Wurmes zu eigenständigen Wesen entwickelt. Das liegt daran, dass der Regenwurm ein verblüffend komplexes Wesen ist. Er hat die Fähigkeit, seine inneren Organe, sein Herzkreislaufsystem und sein Gehirn bei Bedarf zu duplizieren. Das liegt in der Entwicklungsgeschichte des Regenwurms begründet. So nützlich er im Garten ist, so leicht wird er auch vom Menschen aus Versehen geteilt. Das weiß jeder Gärtner, der im Übereifer beim Jäten schon einmal einen Regenwurm geteilt hat. Der Regenwurm hat sich in seiner Evolution diesen tragischen Unfällen angepasst, indem er die Fähigkeit entwickelt hat, sich bei Bedarf aus zwei Hälften neu entstehen zu lassen. Praktisch, oder?
Bio-Lebensmittel sind gesünder als konventionelle Lebensmittel.	1	True	True	1	3	4	Wer auf eine gesunde Ernährung achtet, kauft häufig auch Bio-Lebensmittel. Das Bio-Label suggeriert uns, dass ein Produkt gesünder ist, als sein konventionelles Gegenstück. Dabei handelt es sich nicht nur um subjektive Wahrnehmung, sondern Lebensmittel in Bio-Qualität sind nachgewiesenerweise gesünder. Denn Bio-Obst und Gemüse ist keinen chemischen Pestiziden ausgesetzt, Fleisch in Bio-Qualität stammt von Tieren, die mit ebenfalls gesünderen Bio-Produkten gefüttert wurden und auch auf die Qualität tierischer Produkte wie Milch und Eier wirkt sich das Futter der Tiere aus. So ist Bio-Fleisch meist fettärmer und enthält weniger Antibiotika und andere Stoffe, mit denen konventionelles Tierfutter versetzt ist und Bio-Milch wie Milchprodukte sind besser verträglich.
Bio-Lebensmittel sind gesünder als konventionelle Lebensmittel.	2	True	False	4	4	1	Biologisch hergestellte Lebensmittel werden ohne Spritzmittel hergestellt. Es dürfen keine giftigen Stoffe verwendet werden, um die Lebensmittel herzustellen. Spritzmittel sind oft starke Nervengifte und konventionelle Lebensmittel enthalten meistens noch große Rückstände dieser toxischen Mittel. Es ist bewiesen, dass diese Mittel Ursache verschiedener Erkrankungen und Mutationen sind. Ebenfalls enthalten konventionell hergestellte Lebensmittel weniger Vitamine als die biologisch hergestellten Vertreter. Biologisch hergestellte Lebensmittel enthalten somit weniger toxische Stoffe, die Ursache für verschiedene Krankheiten sein können. Des Weiteren enthalten sie mehr gesundheitsförderliche Stoffe, wie Vitamine und Mineralstoffe. Für den menschlichen Körper sind biologische Lebensmittel somit gesünder als konventionelle Lebensmittel.

Table 4: Sample data from the DEFABEL corpus with all annotations. Translations in Appendix Table 5.

Kolmogorov-Smirnov test (Smirnov, 1939) due to its suitability for assessing overall distribution similarity, with ordinal data and uneven sample sizes. We apply a significance level (α) of 0.05. Out of the 30 statements, only 3 exhibited significant distribution differences. This confirms that our assumption holds for the majority of the selected statements.

5. Conclusion and Future Work

The concepts of deception and factuality have been studied extensively in NLP. However, previous studies have largely overlooked the interaction between these two concepts. We mitigate this situation by introducing the DEFABEL corpus of argumentative texts, in German, labeled for deceptiveness. Different from previous studies, we use a novel annotation scheme where we align deceptive intent to argumentation that contradicts one's own belief. Furthermore, this is the first publicly available German corpus to investigate deception from a language perspective.

In our data analysis we find that, interestingly, people appear to be more confident in their argu-

ments when the statement is aligned with their belief, but surprisingly less confident when they are arguing in favor of truthful facts.

Our work in this paper does raise some important future research questions. Most crucial from the data creation perspective is to better understand the reasons behind the varying distributions of familiarity and self-assessed persuasiveness. The observed values are partially counter-intuitive, and we need to better understand why this is the case. One way could be to perform a follow-up study in which participants are prompted to explain the assessment in more detail. While we currently lack evidence to support this, it could in principle result from the choice of topics.

Further, we asked participants in our study to create persuasive texts. However, to assess the persuasiveness, we limited ourselves to obtaining labels through the self-assessment of the author. A logical next step is to perform a study in which readers are asked to rate or rank the persuasiveness of argumentative texts.

Finally and most importantly, this corpus serves as a fundament for the development of new deception detection models, which can, for the first time, disentangle deceptiveness and factuality of the texts. Therefore, we see the use of these models not only in deception research but also in the improvement of fact-checking models. These models might, so far, be confounded by properties of deception.

Acknowledgments

This research has partially been supported by the FIBISS project (Automatic Fact Checking for Biomedical Information in Social Media and Scientific Literature), funded by the German Research Council (DFG, project number: KL 2869/5-1).

Ethical Consideration

In this study, we collected deception data, where participants are explicitly instructed to argue against their own beliefs. Nevertheless, participants have not been exposed to a uncommonly high stress, because the statements used in this study are common misconception and not highly consequential. We manually selected instances to minimize the potential harm or discomfort that they can cause.

Before taking part in the study, participants were informed about the nature of the task. We obtained explicit consent from the participants before going forward. They were also informed that they could withdraw from the study at any time without any consequences. We do not collect or store any personally identifiable information from the participants and hence all instances are inherently anonymized.

We acknowledge that, in principle, deception detection models are at risk of being misused. We condemn any use of such models to analyze individual's texts in a way that may lead to inferences regarding an identifiable person. In our research, we aim at better understanding the phenomenon of the intention to deceive instead of making use of the models in a productive environment. We would like to emphasize that at the current moment in time, automatic deception detection models cannot be assumed to perform sufficiently well such that results can be considered reliable. The practical use on individual's texts that are identifiable may lead to incorrect predictions and may even harm individuals based on unfair assessments.

6. Bibliographical References

Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7067–7072, Online. Association for Computational Linguistics.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ángela Almela, Rafael Valencia-García, and Pascual Cantos. 2012. Seeing through deception: A computational approach to deceit detection in written communication. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 15–22, Avignon, France. Association for Computational Linguistics.

Milad Alshomary, Wei-Fan Chen, Timon Gurcke, and Henning Wachsmuth. 2021. Belief-based generation of argumentative claims. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 224–233, Online. Association for Computational Linguistics.

Katarina Boland, Pavlos Fafalios, Andon Tchechmedjiev, Stefan Dietze, and Konstantin Todorov. 2022. Beyond facts–a survey and conceptualisation of claims in online discourse

analysis. *Semantic Web – Interoperability, Usability, Applicability*, 13(5):793–827.

Gary D. Bond and Adrienne Y. Lee. 2005. Language of lies in prison: linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology*, 19(3):313–329.

Pablo Briñol and Richard E. Petty. 2009. Chapter 2 persuasion: Insights from the self-validation hypothesis. In *Advances in Experimental Social Psychology*, volume 41, pages 69–118. Academic Press.

Pablo Briñol, Richard E. Petty, and Zakary L. Tormala. 2004. Self-Validation of Cognitive Responses to Advertisements. *Journal of Consumer Research*, 30(4):559–573.

Pasquale Capuozzo, Ivano Lauriola, Carlo Strapparava, Fabio Aiolli, and Giuseppe Sartori. 2020. DecOp: A multilingual and multi-domain corpus for detecting deception in typed text. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1423–1430, Marseille, France. European Language Resources Association.

Shelly Chaiken and Alice H Eagly. 1983. Communication modality as a determinant of persuasion: The role of communicator salience. *Journal of personality and social psychology*, 45(2):241.

Marilyn J. Chambliss and Ruth Garner. 1996. Do adults change their minds after reading persuasive text? *Written Communication*, 13:291 – 313.

Michael H. Connors and Peter W. Halligan. 2015. A cognitive account of belief: a tentative road map. *Frontiers in Psychology*, 5.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2021. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Bob de Ruiter and George Kachergis. 2018. The mafiascum dataset: A large text corpus for deception detection. *ArXiv*, abs/1811.07851.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1035–1045, New Orleans, Louisiana. Association for Computational Linguistics.

Alice H. Eagly and Shelly Chaiken. 1975. An attribution analysis of the effect of communicator characteristics on opinion change: The case of communicator attractiveness. *Journal of Personality and Social Psychology*, 32:136–144.

Tommaso Fornaciari, Leticia Cagnina, Paolo Rosso, and Massimo Poesio. 2020. Fake opinion detection: how similar are crowdsourced datasets to real data? *Language Resources and Evaluation*, pages 1–40.

David M. Godden. 2010. The importance of belief in argumentation: Belief, commitment and the effective resolution of a difference of opinion. *Synthese*, 172(3):397–414.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Swati Gupta, Kayo Sakamoto, and Andrew Ortony. 2013. Telling it like it isn't: A comprehensive approach to analyzing verbal deception. Online.

Jeffrey T. Hancock. 2009. Digital deception: Why, when and how people lie online. In *Oxford Handbook of Internet Psychology*. Oxford University Press.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Paige E. Lloyd, Jason C. Deska, Kurt Hugenberg, Allen R. McConnell, Brandon T. Humphrey, and Jonathan W. Kunstman. 2019. Miami university deception detection database. *Behavior Research Methods*, 51:429–439.

Liane Longpre, Esin Durmus, and Claire Cardie. 2019. Persuasion of the undecided: Language vs. the listener. In *Proceedings of the 6th Workshop on Argument Mining*, pages 167–176, Florence, Italy. Association for Computational Linguistics.

Charles G. Lord, Lee D. Ross, and Mark R. Lepper. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37:2098–2109.

James Edwin Mahon. 2007. A definition of deceiving. *International Journal of Applied Philosophy*, 21(2):181–194.

Robert McHardy, Heike Adel, and Roman Klinger. 2019. Adversarial training for satire detection: Controlling for confounding variables. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 660–665, Minneapolis, Minnesota. Association for Computational Linguistics.

Sandra Metts. 1989. An exploratory investigation of deception in close relationships. *Journal of Social and Personal Relationships*, 6(2):159–179.

Judson R. Mills and Charles E. Kimble. 1973. Opinion change as a function of perceived similarity of the communicator and subjectivity of the issue. *Bulletin of the psychonomic society*, 2:35–36.

Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675.

Myle Ott, Claire Cardie, and Jeffrey T. Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 497–501, Atlanta, Georgia. Association for Computational Linguistics.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA. Association for Computational Linguistics.

Verónica Pérez-Rosas and Rada Mihalcea. 2014. Cross-cultural deception detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 440–445, Baltimore, Maryland. Association for Computational Linguistics.

Denis Peskov, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber. 2020. It takes two to lie: One to lie, and one to listen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3811–3854, Online. Association for Computational Linguistics.

Dina Pisarevskaya, Tatiana Litvinova, and Olga Litvinova. 2017. Deception detection for the Russian language: Lexical and syntactic parameters. In *Proceedings of the 1st Workshop on Natural Language Processing and Information Retrieval associated with RANLP 2017*, pages 1–10, Varna, Bulgaria. INCOMA Inc.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Victoria L. Rubin, Yimin Chen, and Nadia K. Conroy. 2015. Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Franco Salvetti, John B. Lowe, and James H. Martin. 2016. A tangled web: The faint signals of deception in text - boulder lies and truth corpus (BLT-C). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3510–3517, Portorož, Slovenia. European Language Resources Association (ELRA).

Eric Schwitzgebel. 2023. Belief. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2023 edition. Metaphysics Research Lab, Stanford University.

Stephen Cameron Skalicky, Nicholas D. Duran, and Scott Andrew Crossley. 2020. Please, please, just tell me: The linguistic features of humorous deception. *Dialogue Discourse*, 11:128–149.

Nikolai V Smirnov. 1939. Estimate of deviation between empirical distribution functions in two independent samples. *Bulletin Moscow University*, 2(2):3–16.

Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea. 2019. Box of lies: Multimodal deception detection in dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1768–1777, Minneapolis, Minnesota. Association for Computational Linguistics.

Irina Temnikova, Silvia Gargova, Ruslana Margova, Veneta Kireva, Ivo Dzhumerov, Tsvetelina Stefanova, and Hristiana Krasteva. 2023. New bulgarian resources for studying deception and detecting disinformation. In *10th LANGUAGE AND TECHNOLOGY CONFERENCE: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Adam Mickiewicz University Press.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ronny E. Turner, Charles Edgley, and Glen Olmstead. 1975. Information control in conversations: Honesty is not always the best policy. *The Kansas Journal of Sociology*, 11(1):69–89.

Aswathy Velutharambath and Roman Klinger. 2023. UNIDECOR: A unified deception corpus for cross-corpus deception detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 39–51, Toronto, Canada. Association for Computational Linguistics.

Ben Verhoeven and Walter Daelemans. 2014. CLiPS stylometry investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3081–3085, Reykjavik, Iceland. European Language Resources Association (ELRA).

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230, Toronto, Canada. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Wenlin Yao, Zeyu Dai, Ruihong Huang, and James Caverlee. 2017. Online deception detection refueled by real world data collection. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 793–802, Varna, Bulgaria. INCOMA Ltd.

Miron Zuckerman, Bella M. DePaulo, and Robert Rosenthal. 1981. Verbal and nonverbal communication of deception. In Leonard Berkowitz, editor, *Advances in Experimental Social Psychology*, volume 14, pages 1–59. Academic Press.

A. Corpus Creation Details

A.1. Belief Distribution Assessment

Google Forms is used to collect the distribution of beliefs regarding the 50 statements. Participants are asked to report their belief on a given statement on a scale of 1-5, as shown in Figure 4. To make sure that participants are paying attention to the task, 5 attention check questions were added to the survey. A sample attention check question is shown in Figure 5.

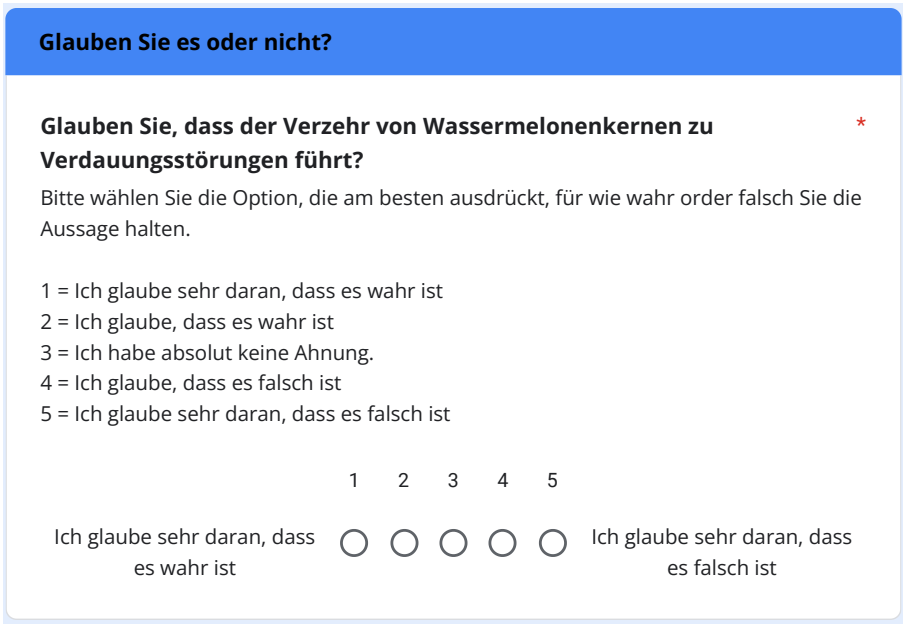

Figure 4: Belief assessment question as displayed in Google Forms

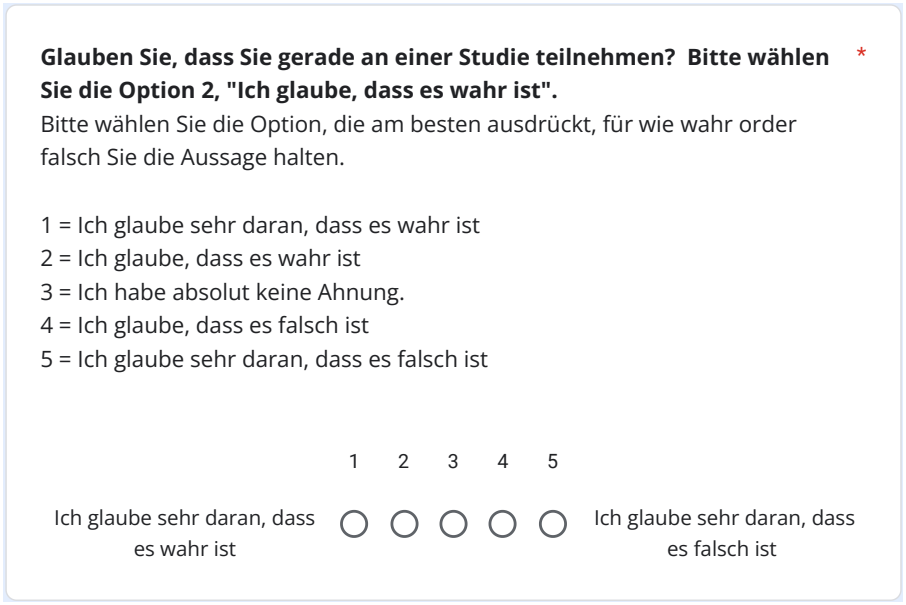

Figure 5: Example of attention check question

A.2. Argument Generation

In the argument generation study, participants are prompted to write convincing arguments supporting a given statement. An example of the expected type of text is provided for reference. On completing the writing task, they are asked to report their familiarity with the topic and confidence in the convincingness of their own arguments as shown in Figure 6. In the study, the original data is collected on scales ranging from 0-3 (for topic familiarity) and 0-4 (for convincingness). However, for comparative analysis in this paper, they are referenced on a scale of 1-4 and 1-5, respectively.

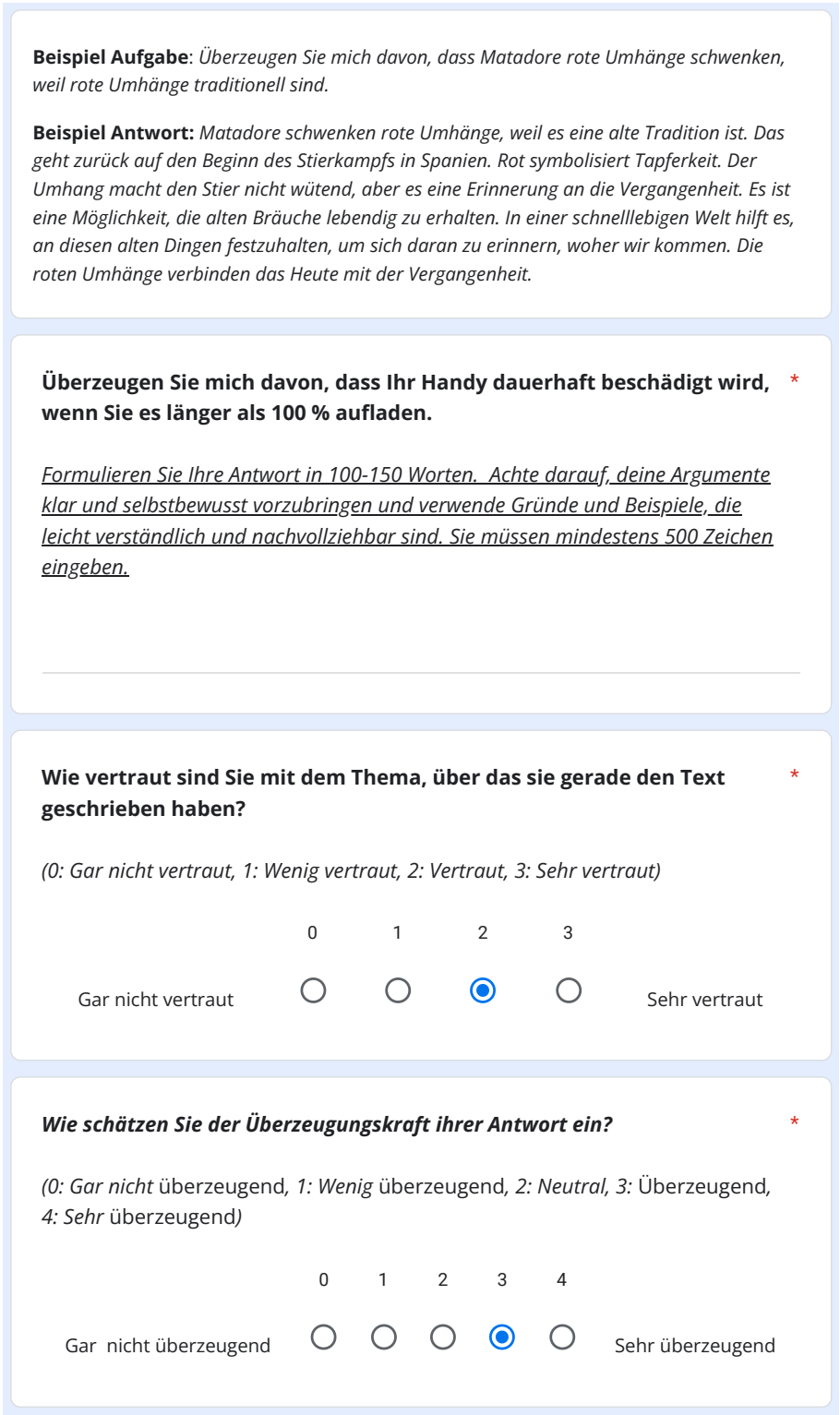

Figure 6: Argument generation task as displayed in Google Forms.

B. Data Sample Translation

Statement	Annot._id	Factual	Deceptive	Familiarity	Confidence	Belief	Arguments
When you cut an earthworm in half, two earthworms are formed.	1	False	True	2	4	4	In contrast to other animals or mammals, an earthworm does not have a brain but rather a decentralized nervous system that controls its functions. Additionally, it lacks a heart or other singular organs crucial for life. Digestion and respiration are not confined to a specific location. As a result, when an earthworm is cut, it forms two parts that are independently viable. After a certain period, tail/head structures grow at each end, comparable to the original ends of the worm – thus, two new viable earthworms emerge.
When you cut an earthworm in half, two earthworms are formed.	2	False	False	1	5	2	If one cuts an earthworm, the creature essentially duplicates itself, as the two halves of the severed worm develop into independent beings. This is due to the earthworm being a remarkably complex organism. It possesses the ability to duplicate its internal organs, circulatory system, and brain as needed. This capability is rooted in the evolutionary history of the earthworm. As beneficial as it is in the garden, it is also easily inadvertently divided by humans. This is a well-known fact for any gardener who has, in their zeal for weeding, accidentally split an earthworm. The earthworm has adapted to these unfortunate accidents in its evolution by developing the ability to regenerate into two halves when needed. Practical, isn't it?
Organic food is healthier than conventional food.	1	True	True	1	3	4	Those who prioritize a healthy diet often choose to buy organic food. The organic label suggests to us that a product is healthier than its conventional counterpart. This is not just a subjective perception; organic foods have been proven to be healthier. Organic fruits and vegetables, for example, are not exposed to chemical pesticides. Meat in organic quality comes from animals that have been fed healthier organic products, and the quality of animal products such as milk and eggs is also influenced by the animals' diet. Consequently, organic meat is often lower in fat and contains fewer antibiotics and other substances found in conventional animal feed. Organic milk and dairy products are also better tolerated.
Organic food is healthier than conventional food.	2	True	False	4	4	1	Organically produced foods are made without pesticides. No toxic substances are allowed to be used in the production of these foods. Pesticides are often strong neurotoxins, and conventional foods usually contain significant residues of these toxic substances. It has been proven that these substances are the cause of various diseases and mutations. Additionally, conventionally produced foods contain fewer vitamins than their organically produced counterparts. Therefore, organic foods contain fewer toxic substances that can be the cause of various diseases. Furthermore, they contain more health-promoting substances such as vitamins and minerals. For the human body, organic foods are therefore healthier than conventional foods.

Table 5: Translation of sample instances from the DEFABEL corpus showing all available annotations per instances.

C. Rejected Instances with Translation

Statement	Arguments	Translation
Überzeugen Sie mich davon, dass Affen genauso gut addieren können wie Menschen. (*Convince me that monkeys can add as well as humans.*)	Ich kann sie davon leider nicht überzeugen, da es hierfür keine wissenschaftlich fundierte Beweise gibt. Es ist Affen nicht möglich, komplexere Aditionen zu lösen, allerdings gab es bereits Studien, bei den herauskam, dass Affen einfache Additionen mit simplen Hilfsmitteln lösen können. Sie bedienen sich dabei oftmals visueller Hilfsmittel wie Punkten. Es ist daher denke ich durchaus möglich, dass Affen einzelne Additionen mit 2 oder 3 Zahlen lösen können, jedoch nicht das Verständnis für komplexere Aufgaben haben.	I cannot convince you of this, unfortunately, as there is no scientifically proven evidence for it. It is not possible for monkeys to solve more complex additions; however, there have been studies that revealed monkeys can solve simple additions with basic aids. They often use visual aids such as dots. Therefore, I believe it is entirely possible for monkeys to solve individual additions with 2 or 3 numbers, but they do not have the understanding for more complex tasks.
Überzeugen Sie mich davon, dass die Erde im Januar der Sonne am nächsten ist. (*Convince me that the Earth is closest to the sun in January.*)	Das ist absurd, da fällt mir nix ein. Tut mir leid, da muss ich passen . Hoffe das disqulifiziert mich jetzt nicht.	That's absurd, I can't think of anything. I'm sorry, I have to pass on this one. I hope this doesn't disqualify me now.
Überzeugen Sie mich davon, dass man nicht zu nah vor dem Fernseher sitzen sollte, weil das zu einer vorübergehenden Ermüdung der Augen führen kann.(*Convince me that one should not sit too close to the television because it can lead to temporary eye fatigue.*)	Guck kein Fernseher , das ist reine Zeitverschwendung genau wie diese Studie zu absolvieren. Danke für nichts und guck du lieber Fernsehen vlt überzeugt dich das Fernsehen anstatt sich über solche Themen schlau zu machen.	Don't watch TV, it's pure waste of time, just like completing this study. Thanks for nothing, and you'd better watch TV instead of educating yourself about such topics.

Table 6: Examples of rejected instances which are filtered out from the DEFABEL corpus.