

Crowdsourcing and Validating Event-focused Emotion Corpora for German and English

Enrica Troiano, Sebastian Padó and Roman Klinger

Institut für Maschinelle Sprachverarbeitung

University of Stuttgart, Germany

{firstname.lastname}@ims.uni-stuttgart.de

Abstract

Sentiment analysis has a range of corpora available across multiple languages. For emotion analysis, the situation is more limited, which hinders potential research on cross-lingual modeling and the development of predictive models for other languages. In this paper, we fill this gap for German by constructing deISEAR, a corpus designed in analogy to the well-established English ISEAR emotion dataset. Motivated by Scherer’s appraisal theory, we implement a crowdsourcing experiment which consists of two steps. In step 1, participants create descriptions of emotional events for a given emotion. In step 2, five annotators assess the emotion expressed by the texts. We show that transferring an emotion classification model from the original English ISEAR to the German crowdsourced deISEAR via machine translation does not, on average, cause a performance drop.

1 Introduction

Feeling emotions is a central part of the “human condition” (Russell, 1945). While existing studies on automatic recognition of emotions in text have achieved promising results (Pool and Nisim (2016); Mohammad (2011), *i.a.*), we see two main shortcomings. First, there is shortage of resources for non-English languages, with few exceptions, like Chinese (Li et al., 2017; Odbal and Wang, 2014; Yuan et al., 2002). This hampers the data-driven modeling of emotion recognition that has unfolded, *e.g.*, for the related task of sentiment analysis. Second, emotions can be expressed in language with a wide variety of linguistic devices, from direct mentions (*e.g.*, “I’m angry”) to evocative images (*e.g.*, “He was petrified”) or prosody. Computational emotion recognition on English has mostly focused on explicit emotion expressions. Often, however, emotions are merely

inferable from world knowledge and experience. For instance, “*I finally found love*” presumably depicts a joyful circumstance, while fear probably ensued when “*She heard a sinister sound*”. Attention to such *event-related emotions* is arguably important for wide-coverage emotion recognition and has motivated shared tasks (Klinger et al., 2018), structured resources (Balahur et al., 2011) and dedicated studies such as the “International Survey on Emotion Antecedents and Reactions” (ISEAR, Scherer and Wallbott, 1994). ISEAR, as one outcome, provides a corpus of English descriptions of emotional events for 7 emotions (anger, disgust, fear, guilt, joy, shame, sadness). Informants were asked in a classroom setting to describe emotional situations they experienced. This focus on private perspectives on events sets ISEAR apart. Even though from psychology, it is now established in natural language processing as a textual source of emotional events.

With this paper, we publish and analyze deISEAR, a German corpus of emotional event descriptions, and its English companion enISEAR, each containing 1001 instances. We move beyond the original ISEAR in two respects. (i), we move from on-site annotation to a two-step crowdsourcing procedure involving description generation and intersubjective interpretation; (ii), we analyze cross-lingual differences including a modelling experiment. Our corpus, available at <https://www.ims.uni-stuttgart.de/data/emotion>, supports the development of emotion classification models in German and English including multilingual aspects.

2 Previous Work

For the related but structurally simpler task of sentiment analysis, resources have been created in many languages. For German, this includes dictionaries

(Ruppenhofer et al., 2017, *i.a.*), corpora of newspaper comments (Schabus et al., 2017) and reviews (Klinger and Cimiano, 2014; Ruppenhofer et al., 2014; Boland et al., 2013). Nevertheless, the resource situation leaves much to be desired. The situation is even more difficult for emotion analysis. Emotion annotation is slower and more subjective (Schuff et al., 2017). Further, there is less agreement on the set of classes to use, stemming from alternative psychological theories. These include, *e.g.*, discrete classes vs. multiple continuous dimensions (Buechel and Hahn, 2016). Resources developed by one strand of research can be unusable for the other (Bostan and Klinger, 2018).

In German, a few dictionaries have been created for dimensional approaches. Among them is BAWL-R, a list of words rated with arousal, valence and imageability features (Vo et al., 2009; Briesemeister et al., 2011), where the nouns of the lexicon have been assigned to emotion intensities, amongst other values. Still, German resources are rare in comparison to English ones. To our knowledge, corpora with sentence-wise emotion annotations are not available for this language.

In particular, there is no German corpus with speakers' descriptions of emotionally intense events similar to the English ISEAR. ISEAR, the "International Survey on Emotion Antecedents and Reactions" (Scherer and Wallbott, 1997), was conducted by a group of psychologists who collected emotion data in the form of self-reports. The aim of the survey was to probe that emotions are invariant over cultures, and are characterized by patterns of bodily and behavioral changes (*e.g.*, change in breathing, felt temperature, speech behaviors). In order to investigate such view, they administered an anonymous questionnaire to 3000 students all over the world, in which participants were asked to reconstruct an emotion episode associated to one of seven basic emotions (anger, disgust, fear, guilt, joy, sadness, shame), and to recall both their evaluation of the stimulus and their reaction to it. For the final dataset, all the reports were translated to English, and accordingly, the responses of, *e.g.*, German speakers who took part in the survey are not available in their original language.

In this paper, we follow Scherer and Wallbott (1997) by re-using their set of seven basic emotions and recreating part of their questionnaire both in English and German. In contrast to ISEAR, we account for the fact that a description can be re-

lated to different emotions by its writer and its readers. Affective analyses have rendered evidence that emotional standpoints affect the quality of annotation tasks (Buechel and Hahn, 2017). For instance, annotation results vary depending on whether workers are asked if a text is *associated* with an emotion and if it *evokes* an emotion, with the first phrasing downplaying the reader's perspective and inducing higher inter-annotator agreement (Mohammad and Turney, 2013). We take notice of these findings to design our annotation guidelines.

3 Crowdsourcing-based Corpus Creation

We developed a two-phase crowdsourcing experiment: one for generating descriptions, the other for rating the emotions of the descriptions. Phase 1 can be understood as sampling from $P(\text{description}|\text{emotion})$, obtaining likely descriptions for given emotions. Phase 2 estimates $P(\text{emotion}|\text{description})$, evaluating the association between a given description and all emotions. The participants' intuitions gathered this way are interpretable as a measure for the interpersonal validity of the descriptions, and as a point of comparison for our classification results.

The two crowdsourcing phases targeted both German and English. This enabled us to tease apart the effects of the change of setup and change of language compared to the original ISEAR collection.

Phase 1: Generation. We used the Figure-Eight (<https://www.figure-eight.com>) crowdsourcing platform. Following the ISEAR questionnaire, we presented annotators with one of the seven emotions in Scherer and Wallbott's setup, and asked them to produce a textual description of an event in which they felt that emotion. The task of description generation was formulated as one of sentence completion (*e.g.*, "Ich fühlte Freude, als/weil/...", "I felt joy when/because ..."), after observing that this strategy made the job easier for laypersons, without inducing any restriction on sentence structure (for details, see Suppl. Mat., Section A). Further, we asked annotators to specify their gender (male, female, other), the temporal distance of the event (*i.e.*, whether the event took place days, weeks, months, or years before the time of text production), and the intensity and duration of the ensuing emotion (*i.e.*, whether the experience was not very intense, moderately intense, intense and very intense, and whether it lasted a few minutes, one hour, multiple hours, or more than one day). To obtain an

	Emotion	Statistics	Temporal Distance				Intensity				Duration				Gender		
		#tok	D	W	M	Y	NV	M	I	VI	min	h	>h	≥d	M	F	O
German	Anger	15.1	46	25	31	41	3	25	67	48	23	29	39	52	112	31	–
	Disgust	13.1	38	38	42	25	12	52	48	31	95	37	8	3	110	33	–
	Fear	14.0	25	32	37	49	4	24	58	57	50	32	31	30	109	34	–
	Guilt	13.8	36	27	30	50	8	57	54	24	41	29	43	30	116	27	–
	Joy	11.6	40	30	29	44	2	18	60	63	14	18	42	69	107	35	1
	Sadness	11.5	29	26	42	46	3	31	43	66	16	9	27	91	113	30	–
	Shame	13.2	25	28	36	54	24	56	41	22	72	28	24	19	116	27	–
	Sum	13.2	239	206	247	309	56	263	371	311	311	182	214	294	783	217	1
English	Anger	28.3	45	29	25	44	9	34	48	52	30	23	36	54	62	81	–
	Disgust	22.4	57	25	21	40	12	51	37	43	66	27	24	26	57	86	–
	Fear	27.0	19	29	36	59	2	30	57	54	52	29	35	27	66	77	–
	Guilt	25.5	33	24	27	59	25	52	43	23	26	39	28	50	59	84	–
	Joy	23.6	32	24	31	56	2	27	48	66	14	13	43	73	60	83	–
	Sadness	21.6	40	24	31	48	10	45	38	50	17	21	23	82	62	81	–
	Shame	24.8	21	22	19	81	16	51	42	34	29	25	39	50	57	86	–
	Sum	24.7	247	177	190	387	76	290	313	322	234	177	228	362	423	578	–

Table 1: Statistics for prompting emotions across the average number of tokens (#tok) and the extra-linguistic labels of the descriptions. Temporal Distance, Intensity and Duration report the number of descriptions for events which took place days (D), weeks (W), months (M) or years (Y) ago, which caused an emotion of a specific intensity (NV: not very intense, M: moderate, I: intense, VI: very intense) and duration (min: a few minutes, one hour: h, multiple hours: >h, one or multiple days ≥d); Gender counts of the annotators are reported in the last column (male: M, female: F, other: O).

English equivalent to deISEAR, we crowd-sourced the same set of questions in English, creating a comparable English corpus (enISEAR). The generation task was published in two slices (Nov/Dec 2018 and Jan 2019). It was crucial for data quality to restrict the countries of origin (for German, DE/A; for English, UK/IR) – this prevented a substantial number of non-native participants who are proficient users of machine translation services from submitting answers. For each generated description, we paid 15 cents (see Suppl. Material, Section A for details).

Phase 2: Emotion Labeling. To verify to what extent the collected descriptions convey the emotions for which they were produced, we presented a new set of annotators with ten randomly sampled descriptions, omitting the emotion word (e.g., “*I felt ... when/because ...*”), together with the list of seven emotions. The task was to choose the emotion the original author most likely felt during the described event. Each description was judged by 5 annotators. We paid 15 cents per task.

4 Corpus Analysis

Descriptive analysis. We include all descriptions from Phase 1 in the final resource and the upcoming discussion, regardless of the inter-annotator agreement from Phase 2. Both deISEAR and enISEAR comprise 1001 event-centered descrip-

tions: deISEAR includes 1084 sentences and 2613 distinct tokens, with a 0.19 type-token ratio; enISEAR contains 1366 sentences and a vocabulary of 3066 terms, with a type-token ratio of 0.12. Table 1 summarizes the Phase 1 annotation. For each prompting label¹, we report average description length, annotators’ gender, duration, intensity and temporal distance of the emotional events.

The main difference between the two languages is description length: English instances are almost twice as long (24.7 tokens) as German ones (13.2 tokens). These differences may be related to the differences in gender distribution between languages.

Most patterns are similar across German and English. In both corpora, Anger and Sadness receive the longest and shortest descriptions, respectively. Enraging facts are usually depicted through the specific aspects that irritated their experiencers, like “*when a superior at work decided to make a huge issue out of something very petty just to [...] prove they have power over me*”. In contrast, sad events are reported with fewer details, possibly because they are often conventionally associated with pain and require little elaboration, such as “*my grandmother had passed away*”. Also the perceptual assessments of emotion episodes, as given by the extra-linguistic labels, are comparable between lan-

¹Transl. de→en: Angst-Fear, Ekel-Disgust, Freude-Joy, Scham-Shame, Schuld-Guilt, Traurigkeit-Sadness, Wut-Anger

Emotion	German					English				
	≥ 1	≥ 2	≥ 3	≥ 4	$= 5$	≥ 1	≥ 2	≥ 3	≥ 4	$= 5$
Anger	135	125	107	81	52	137	129	112	89	59
Disgust	139	134	130	124	91	118	101	84	76	53
Fear	134	124	108	99	78	136	131	124	116	86
Guilt	137	126	102	67	31	137	130	124	89	44
Joy	142	142	142	140	136	143	143	143	143	137
Sadness	132	123	113	97	76	140	133	131	116	97
Shame	128	109	86	66	41	116	92	64	41	23
<i>Sum</i>	947	883	788	674	505	927	859	782	670	499

Table 2: Number of descriptions whose prompting label (column Emotion) agrees with the emotion labeled by all Phase-2 annotators ($=5$), by at least four (≥ 4), at least three (≥ 3), at least two (≥ 2), at least one (≥ 1).

guages. The majority of descriptions are located at the high end of the scale both for intensity and temporal distance, *i.e.*, they point to “milestone” events that are both remote and emotionally striking.

Agreement on emotions. We next analyze to what extent the emotions labelled in Phase 2 agree with the prompting emotion presented in Phase 1. Table 2 reports for how many descriptions (out of 143) the prompting emotion was selected one, two, three, four, or five (out of five) times in Phase 2. Agreement is similar between deISEAR and enISEAR. This indicates that the German items, although short, are sufficiently informative. In both languages, the agreement drops across the columns, yet half of the descriptions show perfect intersubjective validity ($=5$): 505 for German, 499 for English. We interpret this as a sign of quality.

Again, we find differences among emotions. Agreement is nearly perfect for Joy and rather low for Shame. These patterns can arise due to different processes. Certain emotions are easier to recognize from language (*e.g.*, “*when I saw someone else got stabbed near me*”: Fear) than others (*e.g.*, “*when my daughter was rude to my wife*”: elicited for Shame, arguably also associated with Anger or Sadness). Patterns may also indicate closer conceptual similarity among specific emotions (Russell and Mehrabian, 1977, *cf.*).

To follow up on this observation, Figure 1 shows two confusion matrices for German and English which plot the frequency with which annotators selected emotion labels (Phase 2, rows) for prompting emotions (Phase 1, columns). The results in the diagonals correspond to the $=5$ columns in Table 2, mirroring the overall high level of validity of the descriptions, and spanning the range between Joy (very high agreement) and Shame (low agreement).

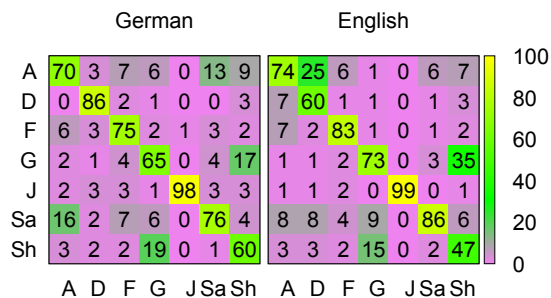


Figure 1: Confusion matrices for emotions. Columns: prompting emotions; rows: labeled emotions.

The off-diagonal cells indicate disagreements. In both languages, annotators perceive Shame descriptions as expressing Guilt, and vice versa (35% and 15% for English, 17% and 19% for German). In fact, Shame and Guilt “occur when events are attributed to internal causes” (Tracy and Robins, 2006), and thus they may appear overlapping.

We also see an interesting cross-lingual divergence. In deISEAR, Sadness is comparably often confused with Anger (13% of items), while in enISEAR it is Disgust that is regularly interpreted as Anger (25% of items). This might result from differences in the connotations of the prompting emotion words in the two languages. For Disgust (“*Ekel*”), German descriptions concentrate on physical repulsion, while the English descriptions also include metaphorical disgust which is more easily confounded with other emotions such as Anger.

Post-hoc Event type analysis. After the preceding analyses, we returned to the Phase 1 descriptions and performed a post-hoc annotation ourselves on a sample of 385 English and 385 German descriptions (balanced across emotions). We tagged them with dimensions motivated by Smith and Ellsworth (1985): whether the event was re-occurring (*general*), whether the event was in the *future* or in the *past*; whether it was a *prospective* emotion or actually felt; whether it had a *social* characteristic (involving other people or animals); whether the event had *self consequences* or *consequences for others*; and whether the author presumably had *situational control* or *responsibility*².

Table 3 shows the results. In both English and German, only a few units depict general and future events, in line with the annotation guidelines. Fear more often targets the future than other emotions. Most event descriptions involve other participants, especially in English. In general, events seem to

²One may be responsible, but not in control of the situation (*e.g.*, “*when I forgot to set an alarm*”).

<i>Dimension</i>		Anger	Disgust	Fear	Guilt	Joy	Sadness	Shame
German	General event	4	2	1	0	0	1	0
	Future event	0	0	1	0	0	0	0
	Past event	51	53	53	55	55	54	55
	Prospective	1	0	4	0	1	1	0
	Social	30	28	24	29	24	40	25
	Self conseq.	37	34	37	26	44	21	37
	Conseq. oth.	21	9	19	34	16	34	14
	Situat. control	2	5	4	24	9	3	19
	Responsible	20	31	17	51	26	23	40
English	General event	2	2	2	2	0	3	0
	Future event	0	0	0	0	0	0	0
	Past event	53	53	53	53	55	52	55
	Prospective	0	0	14	0	1	0	0
	Social	50	37	30	41	39	49	41
	Self conseq.	29	26	42	20	35	16	32
	Conseq. oth.	29	23	19	34	24	43	29
	Situat. control	3	7	8	31	15	2	24
	Responsible	13	29	34	53	34	16	43

Table 3: Event type analysis: Cells are counts of post-annotation out of 55 descriptions for each emotion.

affect authors themselves more than other people, particularly in the case of Joy and Fear. Exceptions are Guilt and Sadness, for which there is a predominance of events whose effects bear down on others. Regarding the aspect of situational control, Shame and Guilt dominate. Guilt is particularly more frequent in descriptions in which the author is presumably responsible. These observations echo the findings by Tracy and Robins (2006).

Modeling. As a final analysis, we tested the compatibility of our created data with the original ISEAR corpus for emotion classification. We trained a maximum entropy classifier with L2 regularization with boolean unigram features on the original ISEAR corpus (7665 instances) and evaluated it on all instances collected in Phase 1 (with liblinear, Fan et al., 2008). We chose MaxEnt as a method as it constitutes a comparably strong baseline which is, in contrast to most neural classifiers, more easy to reproduce due to the convex optimization function and fewer hyper-parameters. We applied it to enISEAR and to a version of deISEAR translated with Google Translate³, an effective baseline strategy for cross-lingual modeling (Barnes et al., 2016). In accord with the Phase 2 experiment, the emotion words present in the sentences were obscured. Table 4 shows a decent performance of the ISEAR model on our novel corpora, with similar scores and performance

³<http://translate.google.com>, applied on February 25, 2019

Dataset	μF_1	An	Di	Fe	Gu	Jo	Sa	Sh
deISEAR	47	29	49	48	42	68	53	39
enISEAR	47	27	45	57	41	67	58	32

Table 4: Performance of ISEAR-trained classifier on our crowdsourced corpora, per emotion and micro-average F_1 (μF_1).

differences between emotion classes to previous studies (Bostan and Klinger, 2018).

Modeling performance and inter-annotator disagreement are correlated: emotions that are difficult to annotate are also difficult to predict (Spearman’s ρ between F_1 and the diagonal in Figure 1 is 0.85 for German, $p = .01$, and 0.75 for English, $p = .05$). It is notable that results for German are on a level with English despite the translation step and the shorter length of the German descriptions. That goes against our expectations, as previous studies showed that translation is only sentiment-preserving to some degree (Salameh et al., 2015; Lohar et al., 2018). We take this outcome as evidence for the cross-lingual comparability of deISEAR and enISEAR, and our general method.

5 Conclusion

We presented (a) deISEAR, a corpus of 1001 event descriptions in German, annotated with seven emotion classes; and (b) enISEAR, a companion English resource build analogously, to disentangle effects of annotation setup and English when comparing to the original ISEAR resource. Our two-phase annotation setup shows that perceived emotions can be different from expressed emotions in such event-focused corpus, which also affects classification performance.

Emotions vary substantially in their properties, both linguistic and extra-linguistic, which affects both annotation and modeling, while there is high consistency across the language pair English-German. Our modeling experiment shows that the straightforward application of machine translation for model transfer to another language does not lead to a drop in prediction performance.

Acknowledgments

This work was supported by Leibniz WissenschaftsCampus Tübingen “Cognitive Interfaces” and Deutsche Forschungsgemeinschaft (project SEAT, KL 2869/1-1). We thank Kai Sassenberg for inspiration and fruitful discussions.

References

- Alexandra Balahur, Jesús M. Hermida, Andrés Montoyo, and Rafael Muñoz. 2011. [EmotiNet: A knowledge base for emotion detection in text built on the appraisal theories](#). In *Natural Language Processing and Information Systems*, pages 27–39, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Jeremy Barnes, Patrik Lambert, and Toni Badia. 2016. [Exploring distributional representations and machine translation for aspect-based cross-lingual sentiment classification](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1613–1623, Osaka, Japan. The COLING 2016 Organizing Committee.
- Katarina Boland, Andias Wira-Alam, and Reinhard Messerschmidt. 2013. [Creating an annotated corpus for sentiment analysis of German product reviews](#). Technical Report 2013/05, GESIS.
- Laura Ana Maria Bostan and Roman Klinger. 2018. [An analysis of annotated corpora for emotion classification in text](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119. Association for Computational Linguistics.
- Benny B. Briesemeister, Lars Kuchinke, and Arthur M. Jacobs. 2011. [Discrete emotion norms for nouns: Berlin affective word list \(DENN–BAWL\)](#). *Behavior Research Methods*, 43(2):441.
- Sven Buechel and Udo Hahn. 2016. [Emotion analysis as a regression problem – dimensional models and their implications on emotion representation and metrical evaluation](#). In *Proceedings of the 22nd European Conference on Artificial Intelligence*, pages 1114–1122, The Hague, The Netherlands.
- Sven Buechel and Udo Hahn. 2017. [Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation](#). pages 1–12.
- Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. [Liblinear: A library for large linear classification](#). *Journal of machine learning research*, 9(Aug):1871–1874.
- Roman Klinger and Philipp Cimiano. 2014. [The US-AGE review corpus for fine grained multi lingual opinion analysis](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 2211–2218, Reykjavik, Iceland. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1656.
- Roman Klinger, Orphee De Clercq, Saif Mohammad, and Alexandra Balahur. 2018. [IEST: WASSA-2018 Implicit Emotions Shared Task](#). In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 31–42. Association for Computational Linguistics.
- Ya Li, Jianhua Tao, Linlin Chao, Wei Bao, and Yazhu Liu. 2017. [Cheavd: a chinese natural emotional audio–visual database](#). *Journal of Ambient Intelligence and Humanized Computing*, 8(6):913–924.
- Pintu Lohar, Haithem Afli, and Andy Way. 2018. [Balancing translation quality and sentiment preservation \(non-archival extended abstract\)](#). In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 81–88, Boston, MA. Association for Machine Translation in the Americas.
- Saif Mohammad. 2011. [From once upon a time to happily ever after: Tracking emotions in novels and fairy tales](#). In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114. Association for Computational Linguistics.
- Saif M Mohammad and Peter D Turney. 2013. [Crowdsourcing a word–emotion association lexicon](#). *Computational Intelligence*, 29(3):436–465.
- Odbal and Zengfu Wang. 2014. [Segment-based fine-grained emotion detection for chinese text](#). In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 52–60, Wuhan, China. Association for Computational Linguistics.
- Chris Pool and Malvina Nissim. 2016. [Distant supervision for emotion detection using facebook reactions](#). In *Proceedings of the Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 30–39. The COLING 2016 Organizing Committee.
- Josef Ruppenhofer, Roman Klinger, Julia Maria Struß Jonathan Sonntag, and Michael Wiegand. 2014. [IG-GSA Shared Tasks on German Sentiment Analysis](#). In *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, Hildesheim, Germany. University of Hildesheim.
- Josef Ruppenhofer, Petra Steiner, and Michael Wiegand. 2017. [Evaluating the morphological compositionality of polarity](#). In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 625–633. INCOMA Ltd.
- Bertrand Russell. 1945. *A History of Western Philosophy*. Routledge Classics.
- James A Russell and Albert Mehrabian. 1977. [Evidence for a three-factor theory of emotions](#). *Journal of research in Personality*, 11(3):273–294.
- Mohammad Salameh, Saif Mohammad, and Svetlana Kiritchenko. 2015. [Sentiment after translation: A](#)

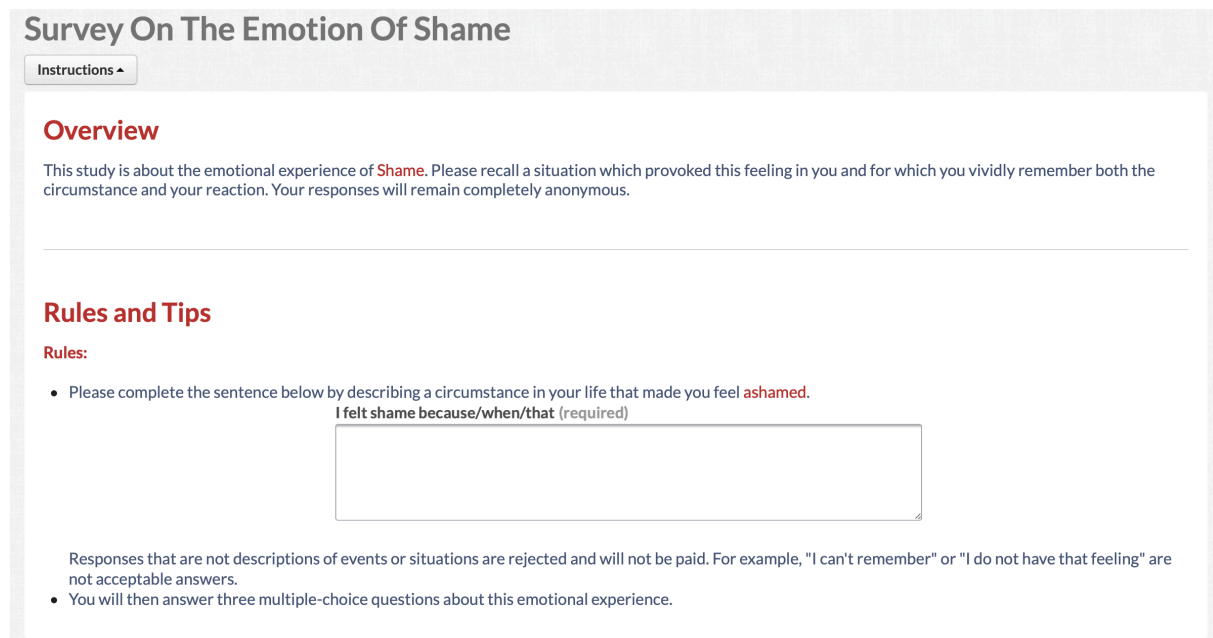
- case-study on arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, Denver, Colorado. Association for Computational Linguistics.
- Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. [One million posts: A data set of german on-line discussions](#). In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17*, pages 1241–1244, New York, NY, USA. ACM.
- Klaus R. Scherer and Harald G. Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.
- Klaus R. Scherer and Harald G. Wallbott. 1997. [The ISEAR questionnaire and codebook](#). *Geneva Emotion Research Group*.
- Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. [Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus](#). In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark. Association for Computational Linguistics.
- Craig A Smith and Phoebe C Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, 48(4):813.
- Jessica L. Tracy and Richard W. Robins. 2006. Appraisal antecedents of shame and guilt: Support for a theoretical model. *Personality and social psychology bulletin*, 32(10):1339–1351.
- Melissa Vo, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus J. Hofmann, and Arthur M. Jacobs. 2009. The Berlin affective word list reloaded (BAWL-R). *Behavior research methods*, 41(2):534–538.
- Jiahong Yuan, Liqin Shen, and Fangxin Chen. 2002. [The acoustic realization of anger, fear, joy and sadness in chinese](#). In *Seventh International Conference on Spoken Language Processing*, pages 2025–2028, Denver, Colorado, USA.

A Corpus Generation and Labelling

For experimental reproducibility, we detail here our crowdsourcing approach. Figure 2 illustrates the instructions presented to the annotators for sentence generation (Phase 1), Figure 3 shows a preview of the task itself. The labelling task of Phase 2 is presented in Figure 4.

To build deISEAR, we targeted Figure-Eight contributors from Germany and Austria, while the English experiment was restricted to United Kingdom and Ireland. As a quality check, we required all workers to be level-3 contributors, i.e., the most experienced ones, who reached the highest accuracy in previous Figure-Eight jobs. It should be noted that these laypeople received only minimal and distant training, while participants of ISEAR were directly instructed by the experimenters. We aimed at adapting their questionnaire to a crowdsourcing framework, by formulating the task of sentence generation as one of sentence completion (e.g. “*Ich fühlte Freude, als/weil/...*”, “*I felt Joy when/because ...*”). Preliminary experiments showed that people provided more coherent and grammatically correct sentences than when they were presented with a faithful translation of the original survey.

Phase 1 involved 121 English jobs and 116 German jobs after filtering unacceptable answers (e.g. nonsensical items), totalling 2002 tasks (hits). The two languages required a diverse amount of jobs because ungrammatical and nonsensical descriptions were (manually) discarded. In the second Phase, 34 jobs were launched for English and 23 for German. This way we collected 5005 annotations for each language (i.e. 5 annotations per description). Overall, data collection and annotation was finalized in three months. The total cost was 300\$ for Phase 1, and 150\$ for Phase 2.



Survey On The Emotion Of Shame

Instructions ▾

Overview

This study is about the emotional experience of **Shame**. Please recall a situation which provoked this feeling in you and for which you vividly remember both the circumstance and your reaction. Your responses will remain completely anonymous.

Rules and Tips

Rules:

- Please complete the sentence below by describing a circumstance in your life that made you feel **ashamed**.
I felt shame because/when/that (required)

Responses that are not descriptions of events or situations are rejected and will not be paid. For example, "I can't remember" or "I do not have that feeling" are not acceptable answers.

- You will then answer three multiple-choice questions about this emotional experience.

Figure 2: Instructions for the Generation Task

Complete the sentence by describing a situation or event—in as much detail as possible—in which you felt **Shame**.

I felt shame because/when/that (required)

When did this happen? (required)

- days ago
- weeks ago
- months ago
- years ago

How long did you feel the emotion? (required)

- a few minutes
- an hour
- several hours
- a day or more

How intense was this feeling? (required)

- not very
- moderately intense
- intense
- very intense

You are (required)

Figure 3: Preview of the Generation Task

Emotion Rating

Instructions ▾

Overview

In an experiment we asked participants to describe emotional situations. Your task now is to guess which emotion was felt.

I felt ... because I received more holidays than I thought I would get, so I could spend more time on my hobbies.

Which emotion, do you think, did the writer of the sentence most likely feel? (required)

- Anger
- Disgust
- Fear
- Guilt
- Joy
- Sadness
- Shame

Figure 4: Preview of the Emotion Validation Task

B Descriptive Analysis

Table 5 and Table 6 present a compact description of the corpora, normalizing the counts by column and by row blocks, as reported in Section 4 in the main paper.

Table 5 highlights differences in the distribution of emotions across different temporal distances, intensities, durations, and annotators’ gender. We see for instance that Shame is outstanding in English for long-distant events, while Anger and Disgust (depending on language) are more dominant in events that happened a few days prior to description production. For intensities, the distribution across emotions is most unbalanced for the label “Not Very”; for duration, Disgust is the prevailing emotion among those which lasted only a few minutes, while it is the less frequent among those which persisted for one or multiple days. The exact opposite holds for Joy and Sadness, which appear to be more durable states.

Table 6 highlights differences in the distribution of extra-linguistic labels across different emotions. A few commonalities emerge between the two languages. The majority of descriptions are referred to remote emotion episodes. Moreover, Anger-, Fear-, Joy- and Sadness-related descriptions are mostly about events which caused very intense affective states. For duration, most occurrences of Anger and Sadness lasted longer than one day both in German and English, while Fear episodes are more short-termed, similar to Disgust.

		Temporal Distance				Intensity				Duration				Gender		
		Emotion	D	W	M	Y	NV	M	I	VI	min	h	>h	≥d	M	F
German	Anger	.19	.12	.13	.13	.05	.10	.18	.15	.07	.16	.18	.18	.14	.14	0
	Disgust	.16	.18	.17	.08	.21	.20	.13	.10	.31	.20	.04	.01	.14	.15	0
	Fear	.10	.16	.15	.16	.07	.09	.16	.18	.16	.18	.14	.10	.14	.16	0
	Guilt	.15	.13	.12	.16	.14	.22	.15	.08	.13	.16	.20	.10	.15	.12	0
	Joy	.17	.15	.12	.14	.04	.07	.16	.20	.05	.10	.20	.23	.14	.16	1
	Sadness	.12	.13	.17	.15	.05	.12	.12	.21	.05	.05	.13	.31	.14	.14	0
	Shame	.10	.14	.15	.17	.43	.21	.11	.07	.23	.15	.11	.06	.15	.12	0
English	Anger	.18	.16	.13	.11	.12	.12	.15	.16	.13	.13	.16	.15	.15	.14	0
	Disgust	.23	.14	.11	.10	.16	.18	.12	.13	.28	.15	.11	.07	.13	.15	0
	Fear	.08	.16	.19	.15	.03	.10	.18	.17	.22	.16	.15	.07	.16	.13	0
	Guilt	.13	.14	.14	.15	.33	.18	.14	.07	.11	.22	.12	.14	.14	.15	0
	Joy	.13	.14	.16	.14	.03	.09	.15	.20	.06	.07	.19	.20	.14	.14	0
	Sadness	.16	.14	.16	.12	.13	.16	.12	.16	.07	.12	.10	.23	.15	.14	0
	Shame	.09	.12	.10	.21	.21	.18	.13	.11	.12	.14	.17	.14	.13	.15	0

Table 5: Statistics normalized by column. The unnormalized counts are shown in the paper in Table 1.

		Temporal Distance				Intensity				Duration				Gender		
		Emotion	D	W	M	Y	NV	M	I	VI	min	h	>h	≥d	M	F
German	Anger	.32	.17	.22	.29	.02	.17	.47	.34	.16	.20	.27	.36	.78	.22	0
	Disgust	.27	.27	.29	.17	.08	.36	.34	.22	.66	.26	.06	.02	.77	.23	0
	Fear	.17	.22	.26	.34	.03	.17	.41	.40	.35	.22	.22	.21	.76	.24	0
	Guilt	.25	.19	.21	.35	.06	.40	.38	.17	.29	.20	.30	.21	.81	.19	0
	Joy	.28	.21	.20	.31	.01	.13	.42	.44	.10	.13	.29	.48	.75	.24	.01
	Sadness	.20	.18	.29	.32	.02	.22	.30	.46	.11	.06	.19	.64	.79	.21	0
	Shame	.17	.20	.25	.38	.17	.39	.29	.15	.50	.20	.17	.13	.81	.19	0
English	Anger	.31	.20	.17	.31	.06	.24	.34	.36	.21	.16	.25	.38	.43	.57	0
	Disgust	.40	.17	.15	.28	.08	.36	.26	.30	.46	.19	.17	.18	.40	.60	0
	Fear	.13	.20	.25	.41	.01	.21	.40	.38	.36	.20	.24	.19	.46	.54	0
	Guilt	.23	.17	.19	.41	.17	.36	.30	.16	.18	.27	.20	.35	.41	.59	0
	Joy	.22	.17	.22	.39	.01	.19	.34	.46	.10	.09	.30	.51	.42	.58	0
	Sadness	.28	.17	.22	.34	.07	.31	.27	.35	.12	.15	.16	.57	.43	.57	0
	Shame	.15	.15	.13	.57	.11	.36	.29	.24	.20	.17	.27	.35	.40	.60	0

Table 6: Statistics normalized by partial row. The unnormalized counts are shown in the paper in Table 1.

C Event-type Analysis

The event-type analysis presented in Section 4 targeted 385 items per language (55 descriptions per emotion). Table 2 in the paper shows the counts of instances associated to the psychological labels across the seven emotions.

For each description, we annotated the following boolean variables:

- About the event time:
 - Does the text describe a *general event*?
 - Does the text describe a *future event*?
 - Does the text describe a *past event*?
- About the realization of the emotion:
 - Is it an actual or a *prospective* emotion?
- About the embedding in a social environment:
 - Are other people or animals part of the event description; is it a *social* event description?
- About the consequences of the event:
 - Are there *self-consequences*?
 - Are there *consequences for others*?
- About the control of the writer:
 - Is the author presumably under *situational control*?
 - Does the author presumably have *self control/responsibility*?

While the paper describes the distribution of labels by emotion, here we expand the discussion to the extra-linguistic information collected in Phase 1. Table 7 distributes the raw counts across the annotation values. It should be noticed that the random descriptions used for this analysis were not balanced with respect to their values of each variable. For this reason, Table 8 reports relative counts (i.e. counts of descriptions normalized by the number of instances within the label Day, Week, Month etc.).

Some regularities can be observed across all columns of Table 8. For instance, events which involved a purposeful participation of their experiencer are a minority in both languages (Sit. control), and approximately 50% of the descriptions mention individuals other than the writer (Social). The latter proportion, however, is higher for English than for German.

Events that are linked to consequences for the self mostly come from the German sample (Self conseq.). In German, moreover, such type of events are recalled more frequently than events that had consequences on others (Conseq. oth.). The opposite is true for English: emotions of English authors often wrote about events that affected the life of other people or animals. This holds irrespective of the temporal distance, the intensity, the duration of the experience and the gender of the experiencer. Exceptions are English descriptions of facts which only lasted a few minutes, and which appear to bring consequences for the self more than for others (Self conseq. and Conseq. oth. in column min).

As for the responsibility of events, this label is consistent across all columns in the German sample. Instead, in English we observe some marked differences. Emotions with a low intensity (column NV) followed an event which was directly triggered by their experiencer, but very intense emotions are less frequently associated to responsibility (column VI). Lastly, shorter events (min) imply the responsibility dimension more than long ones ($\geq d$).

Dimension	Temporal Distance				Intensity				Duration				Gender			
	D	W	M	Y	NV	M	I	VI	min	h	>h	≥d	M	F	O	
German	General Event	2	3	1	2	0	1	4	3	4	0	1	3	6	2	0
	Future Event	0	0	1	0	0	0	1	0	0	0	1	0	1	0	0
	Past Event	98	76	101	101	22	92	141	121	121	66	83	106	287	89	0
	Prospective	3	2	2	0	0	3	3	1	2	2	1	2	5	2	0
	Social	55	41	53	51	13	43	80	64	70	32	42	56	152	48	0
	Self conseq.	54	45	70	67	15	52	94	75	74	36	59	67	176	60	0
	Conseq. oth.	42	30	34	41	10	35	54	48	52	25	28	42	110	37	0
	Sit. ctrl.	17	13	18	18	2	17	29	18	21	10	14	21	56	10	0
	Responsib.	53	37	63	55	11	57	76	64	68	40	45	55	160	48	0
	Sum	226	171	242	234	51	208	341	273	291	145	191	246	666	207	0
English	General Event	6	2	2	1	2	2	5	2	5	2	1	3	3	8	0
	Future Event	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Past Event	88	61	73	152	23	104	122	125	76	74	85	139	155	219	0
	Prospective	3	4	3	5	0	2	8	5	5	4	3	3	7	8	0
	Social	73	51	56	107	14	72	94	107	49	52	66	120	103	184	0
	Self conseq.	46	30	34	90	14	57	71	58	52	32	47	69	89	111	0
	Conseq. oth.	51	38	39	73	8	49	69	75	30	47	40	84	73	128	0
	Sit. ctrl.	15	17	16	42	12	30	25	23	21	19	17	33	40	50	0
	Responsib.	50	36	47	89	20	71	80	51	57	50	53	62	104	118	0
	Sum	244	178	197	407	70	283	352	321	219	206	227	374	419	607	0

Table 7: Event-type analysis: Raw counts of the labels which were manually assigned to a subset of enISEAR and deISEAR, across the extra-linguistic information collected in Phase 1. See the text for the explanation of variables.

Dimension	Temporal Distance				Intensity				Duration				Gender			
	D	W	M	Y	NV	M	I	VI	min	h	>h	≥d	M	F	O	
German	General Event	.02	.04	.01	.02	0	.01	.03	.02	.03	0	.01	.03	.02	.02	0
	Future Event	0	0	.01	0	0	0	.01	0	0	0	.01	0	0	0	0
	Past Event	.98	.96	.98	.98	1	.99	.97	.98	.97	1	.98	.97	.98	.98	0
	Prospective	.03	.03	.02	0	0	.03	.02	.01	.02	.03	.01	.02	.02	.02	0
	Social	.55	.52	.51	.50	.59	.46	.55	.52	.56	.48	.49	.51	.52	.53	0
	Self conseq.	.54	.57	.68	.65	.68	.56	.64	.60	.59	.55	.69	.61	.60	.66	0
	Conseq. oth.	.42	.38	.33	.40	.45	.38	.37	.39	.42	.38	.33	.39	.37	.41	0
	Sit. ctrl.	.17	.16	.17	.17	.09	.18	.20	.15	.17	.15	.16	.19	.19	.11	0
	Responsib.	.53	.47	.61	.53	.50	.61	.52	.52	.54	.61	.53	.50	.54	.53	0
	English	General Event	.06	.03	.03	.01	.08	.02	.04	.02	.06	.03	.01	.02	.02	.04
Future Event		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Past Event		.94	.97	.97	.99	.92	.98	.96	.98	.94	.97	.99	.98	.98	.96	0
Prospective		.03	.06	.04	.03	0	.02	.06	.04	.06	.05	.03	.02	.04	.04	0
Social		.78	.81	.75	.70	.56	.68	.74	.84	.60	.68	.77	.85	.65	.81	0
Self conseq.		.49	.48	.45	.59	.56	.54	.56	.46	.64	.42	.55	.49	.56	.49	0
Conseq. oth.		.54	.60	.52	.48	.32	.46	.54	.59	.37	.62	.47	.59	.46	.56	0
Sit. ctrl.		.16	.27	.21	.27	.48	.28	.20	.18	.26	.25	.20	.23	.25	.22	0
Responsib.		.53	.57	.63	.58	.80	.67	.63	.40	.70	.66	.62	.44	.66	.52	0

Table 8: Event-type analysis: Counts are normalized by instances with the particular value, e.g., the count in the cell “Time General”–“D” is normalized by the number of all instances with the associated value D (temporal distance of days).

D Annotator Agreement

Section 4.1 discussed the agreement reached by different subsets of annotators *at each generation label*. We report relative counts in Table 9 and we extend the analysis in Table 10, summing over the prompting emotions. This table shows the interannotator agreement of Phase-2 annotators with respect to the meta-information given by the participants of Phase 1, i.e., all the alternatives for gender, intensity, duration and temporal distance under the column Labels.

These numbers represent the count of descriptions within a corpus – and *not within a generation label*, for which the annotation label is the same as the generation label. One can read the table as follows: 177 descriptions from deISEAR, which were labeled as VI by Phase 1 annotators, were then labelled by 5 Phase 2 annotators with their original prompting emotion; 506 instances provided by female annotators for enISEAR were labelled by at least 2 Phase 2 annotators with their original prompting emotion, and so on.

Notably, in the table of Section 4.2, the maximum value that each cell can reach is 143, i.e., the total number of descriptions prompted by a specific emotion. Here, the maximum value varies by cell, because each meta-data label is assigned to a different number of descriptions⁴. Accordingly, higher counts do not necessarily indicate stronger agreement.

Emotion	German					English				
	≥ 1	≥ 2	≥ 3	≥ 4	=5	≥ 1	≥ 2	≥ 3	≥ 4	=5
Anger	.94	.87	.75	.57	.36	.96	.90	.78	.62	.41
Disgust	.97	.94	.91	.87	.64	.83	.71	.59	.53	.37
Fear	.94	.87	.76	.69	.55	.95	.92	.87	.81	.60
Guilt	.96	.88	.71	.47	.22	.96	.91	.87	.62	.31
Joy	.99	.99	.99	.98	.95	1	1	1	1	.96
Sadness	.92	.86	.79	.68	.53	.98	.93	.92	.81	.68
Shame	.90	.76	.60	.46	.29	.81	.64	.45	.29	.16
<i>Sum</i>	6.62	6.17	5.51	4.71	3.53	6.48	6.01	5.47	4.69	3.49

Table 9: Relative agreement counts.

	Labels	German					English				
		≥ 1	≥ 2	≥ 3	≥ 4	=5	≥ 1	≥ 2	≥ 3	≥ 4	=5
When	D	226	157	184	209	226	229	211	189	161	115
	W	197	184	169	143	108	168	152	137	112	79
	M	229	215	198	174	125	177	165	154	138	109
	Y	295	275	237	200	161	353	331	302	259	196
Length	min	291	275	245	213	145	223	208	185	162	115
	h	173	162	151	127	99	162	145	130	106	74
	>h	205	188	164	139	103	210	197	178	158	118
	$\geq d$	278	258	228	195	158	332	309	289	244	192
Intense	NV	52	46	38	32	18	74	69	61	51	31
	M	241	224	194	162	113	264	240	217	185	128
	I	352	331	301	255	197	288	267	247	213	165
	VI	302	282	255	225	177	301	283	257	221	172
Gender	M	738	684	604	510	392	386	353	316	273	200
	F	208	198	183	163	112	541	506	466	397	299
	O	1	1	1	1	1	–	–	–	–	–

Table 10: Full agreement information for both German and English crowd-sourced corpora.

⁴For an overview of the distribution of meta-data labels over the descriptions, refer to Section 4.1.

E Modeling

Table 11 shows the results of the maximum entropy classifier across all emotions.

Emotion	deISEAR						enISEAR					
	TP	FP	FN	P	R	F1	TP	FP	FN	P	R	F1
Anger	29	30	114	.49	.20	.29	27	32	116	.46	.19	.27
Disgust	65	57	78	.53	.45	.49	67	85	76	.44	.47	.45
Fear	70	77	73	.48	.49	.48	85	69	58	.55	.59	.57
Guilt	75	140	68	.35	.52	.42	79	161	64	.33	.55	.41
Joy	106	61	37	.63	.74	.68	94	43	49	.69	.66	.67
Sadness	63	31	80	.67	.44	.53	70	29	73	.71	.49	.58
Shame	66	131	77	.34	.46	.39	49	111	94	.31	.34	.32
<i>Micro</i>	474	527	527	.47	.47	.47	471	530	530	.47	.47	.47

Table 11: Classification results for both corpora.