At the Lower End of Language—
Exploring the Vulgar and Obscene Side of German

Elisabeth Eder **Ulrike Krieg-Holz**
Institut für Germanistik
Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria
{elisabeth.eder | ulrike.krieg-holz}@aau.at

Udo Hahn
Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Jena, Germany
udo.hahn@uni-jena.de

Abstract

In this paper, we describe a workflow for the data-driven acquisition and semantic scaling of a lexicon that covers lexical items from the lower end of the German language register—terms typically considered as rough, vulgar or obscene. Since the fine semantic representation of grades of obscenity can only inadequately be captured at the categorical level (e.g., obscene *vs.* non-obscene, or rough *vs.* vulgar), our main contribution lies in applying best-worst scaling, a rating methodology that has already been shown to be useful for emotional language, to capture the relative strength of obscenity of lexical items. We describe the empirical foundations for bootstrapping such a low-end lexicon for German by starting from manually supplied lexicographic categorizations of a small seed set of rough and vulgar lexical items and automatically enlarging this set by means of distributional semantics. We then determine the degrees of obscenity for the full set of all acquired lexical items by letting crowdworkers comparatively assess their pejorative grade using best-worst scaling. This semi-automatically enriched lexicon already comprises 3,300 lexical items and incorporates 33,000 vulgarity ratings. Using it as a seed lexicon for fully automatic lexical acquisition, we were able to raise its coverage up to slightly more than 11,000 entries.

1 Introduction

With the rapid diffusion of social media in our daily lives, we currently experience (and many of us foster) a fundamental change of social communication habits. A main feature of this new era is an unprecedented degree of public exposure and visibility of individuals via (very) large and intentionally open networks of "friends" or "followers." Blogs, chat rooms and online fora constitute even looser connected social networks with lots of personally weakly acquainted or even unknown interlocutors engaged in digital discourses. Unfortunately, the chance for malicious interactions is promoted by the sheer mass of players in these networks and easy ways of hiding real individual identities via nick names or technically slightly more advanced means of camouflage, such as fake Web identities, including non-benevolent software agents and chatbots (McIntire et al., 2010).

These promiscuous communication groups face a high risk of anti-social behavior by aggressive, ruthless or entirely hostile actors (Dadvar et al., 2014; Wester et al., 2016; Li et al., 2017b; Talukder and Carbunar, 2018). The phenomena encountered range from (political, religious, ethnic, sexual) harassment, flaming, cybertrolling, and cyberbullying to extremely evaluative (derogatory, hurtful, rough, rude, offensive, abusive, vulgar, taboo, obscene) language use (for a typological clarification attempt, cf. Waseem et al. (2017)).

NLP research has recently directed its attention towards these unwarranted effects of social media activities and targets the automatic recognition of toxic language for the purpose of alerting and warning (Huang et al., 2018), filtering and blocking (Yoon et al., 2010; Ghauth and Sukhur, 2015; Chernyak, 2017; Wu et al., 2018), or reformulating suspicious contents of this type by non-obtrusive paraphrases (Su et al., 2017; Nogueira dos Santos et al., 2018).

Yet, how can we distinguish sloppy colloquial language we all use here and there from explicitly abusive and inacceptable wording, the topic we focus on in this paper, i.e., the kind of linguistic behavior typically socially banned from civilized discourse?

Proceedings of the Third Workshop on Abusive Language Online, pages 119–128
Florence, Italy, August 1, 2019.

The standard way to deal with this challenge is to define category systems (binary ones, such as obscene *vs.* non-obscene, or staged ones, as illustrated by pejorative *vs.* rough *vs.* vulgar) and letting people decide on the assignment of lexical items to these discrete categories. Once such categorical features are available, these lexical resources can be exploited for analytic purposes. Traditionally, these decisions were made by few lexicographers but this approach suffers from subjectivity and lack of flexibility, since this lexicon of improper words is rapidly growing due to the productiveness of language and thus changing almost every day.

Alternatively, a larger number of crowdworkers can be hired to provide such category assignments which increases the level of objectivity (on the basis of inter-worker consensus) and currency (campaigns can be run without delay, on demand, with low budgets). Yet, crowdsourced assessments, as with lexicographers' judgments, inherently suffer from the problems of permeable and soft category boundaries—what is rough for one person may be vulgar for another and *vice versa*.

We challenge the established view that the representation of obscenity of language is a discrete categorical classification problem—no matter which category system is chosen—but rather assume that it is a matter of differential degree. Accordingly, we describe the empirical foundations for bootstrapping and scaling such a lexicon from the low end of stylistic conventions on *degrees of obscenity*. We start from expert-level lexicographic categorizations of a small set of pejorative/rough/vulgar lexical items, enlarge this set by distributional semantics methods and, then, determine the degree of obscenity of the items assembled this way by letting crowdworkers make individual assessments relative to the semantic poles "neutral" and "vulgar" using a best-worst scaling approach (Kiritchenko and Mohammad, 2016, 2017).

The resulting lexicon targeting that lower end of German language comprises already 3,300 lexical items, incorporates 33,000 human ratings, and serves as a seed lexicon for fully automatically acquiring and scoring new lexical items from the same register. After several iterations, we finally come up with VULGER, a lexicon of VULgar GERman, totalling slightly more than 11,000 entries.

2 Related Work

Lexicons covering offensive language are almost only available for the English language. Perhaps the earliest collection of such lexical items (including phrases and multi-word expressions) is due to Razavi et al. (2010) who manually assembled approximately 2,700 dictionary entries. More recent work on an alternative verb-centered lexicon (size is not specified) with a focus on hate speech is reported by Gitari et al. (2015). The currently largest and most up to date English lexicon of abusive words is provided by Wiegand et al. (2018a) who manually and automatically collected around 8,500 lexical items.[1]

Languages other than English are incorporated in HURTLEX[2] (Bassignana et al., 2018) which forms a multilingual lexical resource of words that hurt for 53 languages, among them Italian, Spanish, English and German. This lexicon grew out of a manual selection of roughly 1,000 Italian hate words originally organized around 17 categories, with particular focus on derogatory words. It was further semi-automatically extended with complementary borrowings from the Italian MULTIWORDNET[3] and BABELNET.[4] HURTLEX also excels with additional linguistic information (parts of speech, lexicographic definitions) for its lemmas. The lexicon integration step yields roughly 1,160 multilingual lexical items (with the help of the BABELNET API).

Manual curation (for the Italian portion) included a categorization step for each lemma sense into one of three categories: 'Not Offensive'–'Neutral'–'Offensive'. In a subsequent step, the 'Neutral' category was split into 'Not Literally Pejorative' (insult by means of a semantic shift, e.g. metaphorically) and 'Negative Connotation' (not necessarily a direct derogatory use but used in a derogatory way). 2-expert agreements plunged from 87.6% for the 3-category decisions to 61% for the extended 5-category decisions. Clearly, an indicator that such categorical decisions are hard to make even for competent native speakers.

As far as canonical German lexical resources are concerned, their coverage at the low end of lan-

[1] https://github.com/uds-lsv/lexicon-of-abusive-words

[2] http://hatespeech.di.unito.it/resources.html

[3] http://multiwordnet.fbk.eu/english/home.php

[4] https://babelnet.org/

guage is, not surprisingly, more than incomplete. In effect, GERMANET V13.0,[5] for instance, covers only 1,774 lexical items from our seed lexicon (3,300 lexical items, in total). Yet, this ratio is even higher than for other lexical resources such as HATEBASE,[6] a repository which covers 95 languages (with 2,691 hate terms), yet only enumerates 95 manually provided German hate speech entries at all.

In conclusion, the compilation of lexicons for offensive, abusive or hate language typically consists of two steps. First, already available lexical resources covering such pejorative lexical items are identified and bundled in a seed lexicon. Next, this seed is incrementally enlarged—using additional lexical resources (such as WORDNETS, WIKTIONARY, or BABELNET), or employing some sort of machine learning process (Wiegand et al., 2018a). Yet, the semantic core of such lexicons are (manual or automatic) categorical assignments of either bi-polar (e.g., 'Offensive' *vs.* 'Non-Offensive') or multi-polar categories (e.g., 'Colloquial' *vs.* 'Rough' *vs.* 'Obscene').

As an alternative to this scheme, our work focuses on substituting discrete categorical decisions by continuous grading of the above distinctions based on Best-Worst Scaling (Louviere et al., 2015). We thus target a research desideratum already described by Schmidt and Wiegand (2017, p.3-4) in the following way: *"Despite their general effectiveness, relatively little is known about the creation process and the theoretical concepts that underlie the lexical resources that have been specially compiled for hate speech detection."*

3 (Tentatively) Characterizing Vulgar Language

In our study, we not only consider hate speech and abusive terms, but take a much broader perspective on the topic of offensive language and its lexicalizations. Still, this goal is very hard to characterize by distinctive criteria since many lexical-semantic dimensions seem to be involved and strongly interact.

Vulgar language, as we conceive it, is predominantly signalled by an overly lowered language register, the taboo layer, with disgusting and obscene lexicalizations generally banned from any type of civilized discourse. Primarily (yet not only), it addresses the lexical fields of sexuality (sexual organs and activities, in particular), as well as body orifices or other specific body parts (e.g., *"Fresse" ("puss")* as a negative denotation for *"Gesicht/Mund" ("face/mouth")*) and scatologic expressions. One often also observes meaning transfers from animals with culture-dependent negative connotations to humans (e.g., *"Ratte" ("rat")*). Pejorative words with marked negative connotation also play a significant role here (e.g., *"abkratzen" ("croak")*). Especially religious, ethnic and political orientations, the primary targets of hate speech, gain a strong vulgar status when they are combined with (animal-related) swearwords such as *"Schwein" ("pig")*.

We are aware of the preliminary status of this characterization of vulgar language, but consider our work as a starting point for clarifying its nature and systematicity in more depth.

4 Lexicon Acquisition Method

Since a broad-coverage lexicon of obscene German (ranging on an interval from neutral to vulgar) is missing, we decided on a weakly supervised approach to lexicon acquisition based on bootstrapping. It consists of the following steps (the over-all workflow is fundamentally inspired by the work of Wiegand et al. (2018a), yet complements it by a hitherto unexplored methodology to scale the degree of obscenity of lexical items based on best-worst scaling):

1. **Language Resources**: Select a *seed lexicon* (possibly combining numerous relevant resources) which contains a collection of lexical items already tagged as rough and vulgar. Typically, this step reuses manually pre-categorized lexical items (work typically due to experienced lexicographers). Further, this lexical collection can be enhanced by exploiting large-scale *corpora*—these can either be already annotated for (some degree of) vulgarity or lack any annotation of this kind at all—or representational derivatives therefrom, such as (word) embeddings.

2. **Human Assessment**: Refine the seed lexicon by complementary human assessments of obscenity/vulgarity on the basis of *crowdsourcing* using differential *best-worst scaling*.

[5] http://www.sfs.uni-tuebingen.de/GermaNet/

[6] https://hatebase.org

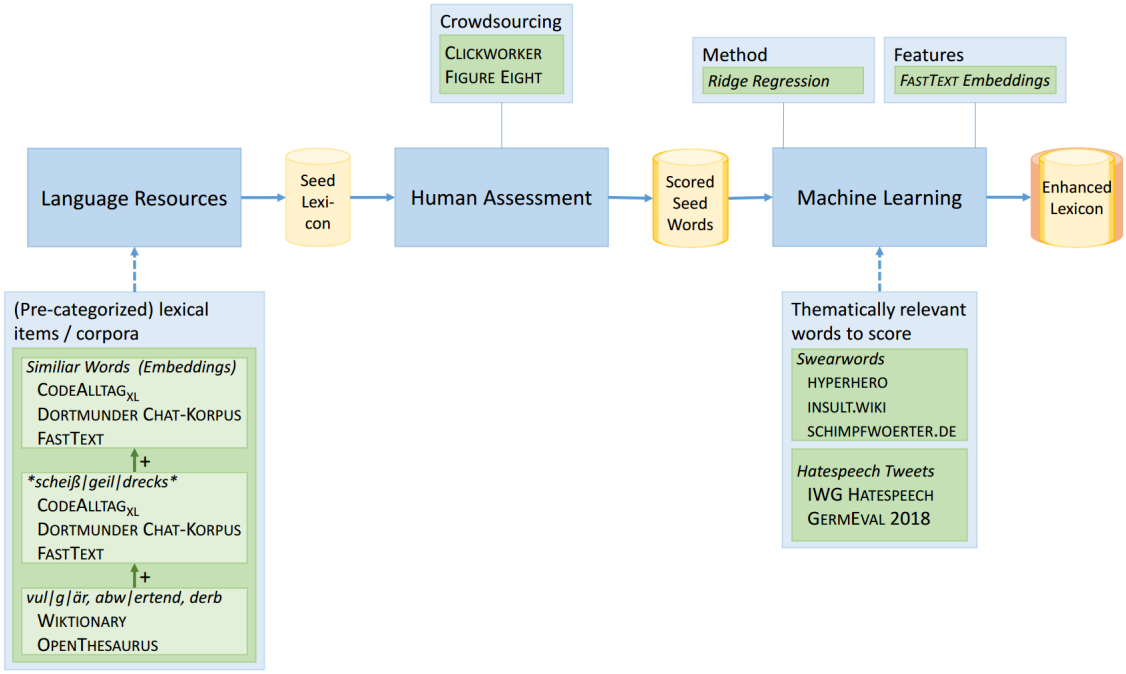

Figure 1: Generic language-independent workflow for lexicon acquisition (in blue) and its instantiation for German (in green); solid blue arrows indicate control flow, data flow is represented by dashed blue arrows, green arrows and '+' stand for lexical data harvesting (with RegEx-style expressions for matching search terms), thin blue lines link particular choices to realizations (implementations) of the blue main components of VULGER's acquisition system

3. **Machine Learning**: Use the resulting lexicon scored on a continuous neutrality-vulgarity scale as training data for automatically identifying and scoring new, thematically relevant lexical items, ideally from corpora containing a high amount of words regarding the property of interest (rough and vulgar wording).

The first step of this workflow (illustrated in Fig. 1), consisting of the assembly of relevant lexical material from scratch, will be described in Section 5. The second one, adding human assessments for that lexical material, is dealt with in Section 6, while the third step, automatic lexicon enhancement, is described in Section 7.

5 Building the Seed Lexicon

From the German slice of WIKTIONARY,[7] we extracted all words marked as vulgar, rough and pejorative.[8] Additionally, we gathered entries tagged with corresponding categories[9] from the German OPENTHESAURUS.[10] As the focus of our corpus lies on single words[11] from a vulgar vocabulary, we excluded phrases from this processing step.

The list resulting from this first step contained some entries used as affixes in morphologically productive word formation, such as *"*geil"*, *"scheiß*"*, *"Scheiß*"* and *"Drecks*"*, the latter ones denoting variants of *"shit"* and *"dirt"*. We cancelled these rudimentary entries from the list because there is no way to get meaningful judgments for them in isolation due to the many possible combinations yielding highly diverse degrees of vulgarity.

In order to account for these terms in a reasonable way we extended our list by harvesting rough and vulgar word forms concatenated with the affixes mentioned above from the CODE ALLTAG$_{S+d}$ email corpus[12] (Krieg-Holz et al., 2016), the DORTMUNDER CHAT KORPUS (Beißwenger, 2013)[12] and from entries in

[7] https://de.wiktionary.org

[8] The exact terms and corresponding abbreviations are: 'vulgär', 'vulg.', 'vul.', 'derb' and 'abwertend', 'abw.'.

[9] 'vulg.', 'derb', 'abwertend'

[10] https://www.openthesaurus.de

[11] The vast potential of the German language for productive noun composition within single compounds makes this decision less restrictive than it seems; cf., e.g., the English phrase form *"son of a bitch"* and its German single compound equivalent *"Hurensohn"*.

[12] We only took words with a minimum frequency of 3.

FASTTEXT (Grave et al., 2018) word embeddings, the latter being based on COMMON CRAWL and WIKIPEDIA.

Yet, we not only incorporated plain text corpora or computationally derived lexical items (exploiting the FASTTEXT embeddings) into our study, but also included word embeddings as a representation format based on the distributional semantics hypothesis and computationally derived from corpora (see also Tulkens et al. (2016); Wiegand et al. (2018a). Utilizing the word embeddings from the corpora mentioned above and the GENSIM module (Řehůřek and Sojka, 2010) we further generated, using the lexical seeds from the previous round, *similar words*, i.e., close semantic neighbors of these seed words by iteratively minimizing the threshold for similarity, until we found too much noise was returned (a common procedure, cf. also Tulkens et al. (2016)). We manually edited the resulting list in regard to inflected forms, misspellings and case sensitivity, but we intentionally kept the 'lexical noise', i.e., presumably neutral words. Since we planned to annotate the lexical items identified this way by crowdsourcing in a later phase, these neutral words also help counterbalance the impact of rough and vulgar expressions during assessments. In total, based on this procedure we gathered a seed lexicon with 3,300 entries.

6 Enriching the Seed Lexicon: Scaling Degrees of Vulgarity

We chose to annotate our seed words with Best-Worst-Scaling (BWS), because it delivers high-quality annotations with only a relatively small number of annotation steps. BWS is an extension of the method of paired comparison to multiple choices, originally developed by Louviere et al. (2015) and introduced into NLP for emotion scaling by Kiritchenko and Mohammad (2016, 2017). For BWS, annotators are presented with n items at a time (an n-tuple, where n ¿ 1, and typically n = 4). They then have to decide which item from the n-tuple under scrutiny is the *best* (highest in terms of the property of interest) and which is the *worst* (lowest in terms of the property of interest).

In our case, judges had to select the *most neutral* and the *most vulgar* terms per given n-tuple. We used the BWS tool[13] from Kiritchenko and Mohammad (2016, 2017) to generate $2N$ decision alternatives (N denotes the size of our seed lexicon) and thus came up with 6,600 4-tuples to be assessed. Tuples were produced randomly under the premise that each term has to occur only once in eight different tuples and each tuple is unique.

For the annotation process proper, we used the crowdsourcing platforms FIGURE EIGHT[14] and CLICKWORKER,[15] where each n-tuple was assessed by five annotators. In order to get real-valued scores from the BWS annotations we applied COUNTS ANALYSIS (Orme, 2009)[16] and thus got scores between $+1$ (most neutral) and -1 (most vulgar). Scores were calculated by subtracting the percentage of times the term was chosen as worst from the percentage of times the term was chosen as best. We computed the split-half reliability[16] like Kiritchenko and Mohammad (2017) by randomly splitting the annotations of a tuple into two halves, calculating scores independently for these halves and measuring the correlation between the resulting two sets of scores. We got an average Pearson correlation of 0.9102 $(+/-$ $0.0022)$ over 100 trials.

7 Automatic Lexicon Extension

7.1 Regression Models

In order to further extend the lexicon in a purely automatic way and also inspired by studies on automatic word emotion induction (especially by Li et al. (2017a) and Buechel and Hahn (2018)) we employed regression models to predict scores for input words. The seed words served as training and testing data for a linear regression and a ridge regression model (linear regression with L_2 regularization during training).[17] As features for the words we used their respective word embeddings (this, obviously, excludes lexical items from further consideration for which no embeddings exist).

We experimented with different word embeddings. We built 100-dimensional word embeddings from CODE ALLTAG$_{XL}$ (Krieg-Holz et al., 2016) using WORD2VEC (Mikolov et al., 2013) for all words occurring at least 3 times in CODE ALLTAG$_{XL}$. Furthermore, we employed WORD2VEC word embeddings from Reimers

[13] http://www.saifmohammad.com/WebPages/BestWorst.html

[14] https://www.figure-eight.com

[15] https://www.clickworker.de

[16] Again we used the scripts from Kiritchenko and Mohammad (2016, 2017).

[17] For both we used the scikit-learn.org implementation using the default parameters.

et al. (2014) with a minimum word frequency of 5 and 100 dimensions (UKP), 300-dimensional FASTTEXT word embeddings from SPINNING-BYTES (Cieliebak et al., 2017) trained on German tweets (TWITTER) and, finally, FASTTEXT word embeddings (Grave et al., 2018) based on COMMON CRAWL and WIKIPEDIA (FASTTEXT). We also tried to utilize embeddings generated from the German TWITTER HATESPEECH corpora from Ross et al. (2016) and Wiegand et al. (2018b) under the assumption that they might contain a large number of rough and vulgar words. But due to their small size and their nevertheless high proportion of out-of-vocabulary words we had to exclude both of these resources from further consideration.

Table 1 shows that the ridge regression model performs equally or slightly better compared to the linear regression model. Regarding the input features the FASTTEXT token embeddings performed best (see Table 2).

Embeddings	LinReg	RidgeReg	p
CODE ALLTAG$_{XL}$	0.574	0.575	0.004
UKP	0.682	0.682	0.121
TWITTER	0.735	0.735	0.073
FASTTEXT	0.766	0.779	0.001

Table 1: Averaged Pearson correlation (10-fold cross validation) and p-value (two-sided t-test) for Linear Regression (LinReg) and Ridge Regression (RidgeReg)

Embeddings	Pearson r	p
CODE ALLTAG$_{XL}$	0.575	< 0.001
UKP	0.682	< 0.001
TWITTER	0.735	< 0.001
FASTTEXT	0.779	—

Table 2: Averaged Pearson correlation (10-fold cross validation) for different embeddings with Ridge Regression, with significance difference to best performing embeddings (p-value from two-sided t-test)

7.2 Applying Regression Models to Enhance the Lexicon

We used the best method (ridge regression and FASTTEXT embeddings) to extend our lexicon with three German swearword lists.[18] There is an overlap between swearwords and vulgar lexicalizations, but not every swearword has strong vulgar status,[19] e.g., *"Schwein" ("pig")*, a subtle distinction which our scaling approach accounts for (cf. also the remarks made in Section 3).

We trained a ridge regression model on the seed words (cf. Sections 5 and 6), i.e., the respective word embeddings and the scores. This model was then applied to the input swearwords (from the three sources mentioned above), which do not occur in the seed lexicon already, and predicted the neutrality/vulgarity scores of the remaining entries on the basis of their word embeddings provided that an embedding for the respective word was found in the FASTTEXT embeddings.[20] We excluded out-of-vocabulary words in order to avoid getting too much noise in terms of wrongly scored lexical items in our lexicon. Further we thus dropped really rare words. With the words already contained in our seed lexicon and words not present as embeddings removed, we assembled 2,046 additional entries following this approach.

Assuming that corpora for hate speech detection include a higher amount of vulgar and rough words, we also made use of such datasets. There exist two publicly available German-language text corpora annotated for hate speech from which we extracted lexical material. The first of them, IWG HATESPEECH, originating from Ross et al. (2016), contains about 500 tweets which were annotated by two judges using a binary categorization scheme ("hate speech": Yes or No) and a 6-point Likert scale ranging from "not offensive" to "very offensive".[21] The second corpus collected by Wiegand et al. (2018b) contains more than 8,500 tweets and was compiled for GERMEVAL 2018, a challenge task addressing the recognition and classification of offensive German language.[22] The latter corpus was coarsely annotated with binary 'Offense' and 'Other' categories, but it also comes with a 4-way classification schema where besides the non-offensive 'Other' class 'Offense' was subdivided in three ways: 'Profanity' (no intent to insult someone, yet the lexical choice is negatively marked, with swearwords such as sca-

[18] These lists were retrieved from http://www.hyperhero.com/de/insults.htm, http://www.insult.wiki/wiki/Schimpfwort-Liste and https://www.schimpfwoerter.de

[19] Also not every vulgar word is a swearword.

[20] We also checked for different spellings regarding case sensitivity.

[21] The corpus is available at https://github.com/UCSM-DUE/IWG_hatespeech_public

[22] The corpus is available at https://projects.cai.fbi.h-da.de/iggsa/

tologic *"Scheiße" (shit)*), 'Insult' (clear intent to offend someone) and 'Abuse' (an even stronger form of 'Insult', i.e., an abusive utterance that degrades a target person/group by ascribing a social identity to a person/group that is judged negatively by a (perceived) majority of society).

From these two corpora we extracted words from all tweets marked as 'Offense' = 'YES' by one of the annotators and further removed stop words, hashtags and words with non-alphabetic characters excluding hyphens or a word length smaller than 4. We also tried to lemmatize the words[23] and normalize spellings in regard to case sensitivity, but admittedly inserted some noise into our input words, i.e., some inflected forms and other forms of semantic duplication could not be normalized. After excluding words already present in the seed lexicon or in the German swearwords lists we applied the same procedure as used for the swearwords and obtained another 5,700 new scored lexical entries.

Due to the lack of better resources we tried to measure the reliability of the resulting scores in a preliminary way by calculating the correlation between the probability of a word being in an offensive post and its score. We got a Pearson correlation coefficient of only -0.35, probably also caused by many words occurring just once, but the correlation may also be inherently weak. In future work, we plan to evaluate the automatically determined extension of our seed word lexicon by feeding the lexical items back into another crowdsourcing round and determining the correlation between the human assessment and the automatically derived scoring values.

The final version of VULGER, a lexicon with VULgarity ratings of GERman words, enhanced with swearwords and words from the two hate speech corpora in the end comprises 11,046 entries (see Table 3).

Resource	# Lexical Items
Seed Words	3,300
German Swearwords	2,046
Twitter Hate Speech Corpora	5,700
Total	11,046

Table 3: Decomposition of contributions from various language resources for VULGER, the current version of the lexicon of VULgar GERman

[23]We used SPACY: https://spacy.io/

8 Conclusion

In this paper, we are concerned with the lexical segment at the lower stylistic end of each natural language often referred to as rough, vulgar and obscene. This register typically covers very explicit and rude linguistic expressions (taboo words). Standard lexical repositories have mostly neglected lots of these expressions on purpose although a pressing need can now be derived for such an extension, e.g., for the purpose of identifying and neutralizing or blocking offensive and humiliating utterances in social media.

Our workflow for building such a lower-end lexicon is based on three steps: assembling already existing lexicons (or fragments therefrom) for this stylistic subvariety of language, assigning degrees of vulgarity for each lexical item included, and using this seed for continuous automatic enhancement by weakly supervised machine learning procedures.

As far as the representation of the semantics of these lexical items are concerned, we propose a continuous grading system to substitute overly simplistic discrete categorical schemata which have been prevailing so far. Still, the claim that such a fine-grained representation is helpful at all must also be demonstrated by experiments in the future. In any case, we plan to use and iteratively extend our newly developed lexicon on text corpora with similar biases into pejorative languages (including scores for obscenity). However, merely (automatically) extending a specialized lexicon might not necessarily prove beneficial as evidenced by the results of Tulkens et al. (2016) that showed no performance boost for a system using such an extended dictionary, at least for detecting Dutch racist language.

In order to by-pass the sparse data problem, methods like transfer learning might also be appropriate here (Sahlgren et al., 2018). Still, the validity of these new items and their scores have to be experimentally validated, e.g., by feeding newly found lexical material back to annotators and compare their judgments with automatically predicted ones.

We are also aware of the fact that purely lexically driven approaches to account for obscene, offensive or vulgar language may not be sufficient to solve the recognition problem completely and that a broader discourse context has to be taken into account, as well as the linguistic conventions in

different communities (Owsley Sood et al., 2012). Still, a lexicon of significant size and quality might form the backbone for machines sensitive to rude and vulgar language.

References

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. HURTLEX: a multilingual lexicon of words to hurt. In *CLiC-it 2018 — Proceedings of the 5th Italian Conference on Computational Linguistics. Torino, Italy, December 10-12, 2018*, number 2253 in CEUR Workshop Proceedings, page #49.

Michael Beißwenger. 2013. Das Dortmunder Chat-Korpus. *Zeitschrift für germanistische Linguistik*, 41(1):161–164.

Sven Buechel and Udo Hahn. 2018. Word emotion induction for multiple languages as a deep multi-task learning problem. In *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, Louisiana, USA, June 1-6, 2018*, volume 1: Long Papers, pages 1907–1918, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Ekaterina Chernyak. 2017. Comparison of string similarity measures for obscenity filtering. In *BSNLP 2017 — Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing @ EACL 2017. Valencia, Spain, April 4, 2017*, pages 97–101, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Mark Cieliebak, Jan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A Twitter corpus and benchmark resources for German sentiment analysis. In *SocialNLP 2017 — Proceedings of the 5th International Workshop on Natural Language Processing for Social Media of the AFNLP SIG SocialNLP @ EACL 2017. Valencia, Spain, April 3, 2017*, pages 45–51, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Maral Dadvar, Dolf Trieschnigg, and Franciska M. G. De Jong. 2014. Experts and machines against bullies: a hybrid approach to detect cyberbullies. In *Advances in Artificial Intelligence. Canadian AI 2014 — Proceedings of the 27th Canadian Conference on Artificial Intelligence. Montréal, Québec, Canada, May 6-9, 2014*, number 8436 in Lecture Notes in Articificial Intelligence (LNAI), pages 275–281, Cham, Switzerland. Springer International Publishing.

Khairil Imran Ghauth and Muhammad Shurazi Sukhur. 2015. Text censoring system for filtering malicious content using approximate string matching and Bayesian filtering. In *Computational Intelligence in Information Systems. INNS-CIIS 2014 — Proceedings of the 4th INNS Symposia Series on Computational Intelligence in Information Systems. Bandar Seri Begawan, Brunei, November 2014*, number 331 in Advances in Intelligent Systems and Computing Book Series (AISC), pages 149–158, Cham, Switzerland. Springer International Publishing.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomáš Mikolov. 2018. Learning word vectors for 157 languages. In *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation. Miyazaki, Japan, May 7-12, 2018*, pages 3483–3487, Paris. European Language Resources Association (ELRA).

Qianjia Huang, Jianhong Zhang, Diana Z. Inkpen, and David Van Bruwaene. 2018. Cyberbullying intervention interface based on convolutional neural networks. In *TRAC 2018 — Proceedings of the 1st Workshop on Trolling, Aggression and Cyberbullying @ COLING 2018. Santa Fe, New Mexico, USA, 25 August, 2018*, pages 42–51, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling. In *NAACL-HLT 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California, USA, June 12-17, 2016*, pages 811–817, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Svetlana Kiritchenko and Saif M. Mohammad. 2017. Best-worst scaling more reliable than rating scales: a case study on sentiment intensity annotation. In *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, British Columbia, Canada, July 30 - August 4, 2017*, volume 2: Short Papers, pages 465–470, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Ulrike Krieg-Holz, Christian Schuschnig, Franz Matthies, Benjamin Redling, and Udo Hahn. 2016. CODE ALLTAG: A German-language e-mail corpus. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016*, pages 2543–2550, Paris. European Language Resources Association (ELRA-ELDA).

Minglei Li, Qin Lu, Yunfei Long, and Lin Gui. 2017a. Inferring affective meanings of words from word embedding. *IEEE Transactions on Affective Computing*, 8(4):443–456.

Tai Ching Li, Joobin Gharibshah, Evangelos E. Papalexakis, and Michalis Faloutsos. 2017b. TROLLSPOT: detecting misbehavior in commenting platforms. In *ASONAM 2017 — Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. Sydney, Australia, July 31 - August 03, 2017*, pages 171–175, New York/NY. Association for Computing Machinery (ACM).

Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press, Cambridge, U.K.

John P. McIntire, Lindsey K. McIntire, and Paul R. Havig. 2010. Methods for chatbot detection in distributed text-based communications. In *CTS 2010 — Proceedings of the 2010 International Symposium on Collaborative Technologies and Systems. Chicago, Illinois, USA, 17-21 May 2010*, pages 463–472. IEEE.

Tomáš Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 — NIPS 2013. Proceedings of the 27th Annual Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA, December 5-10, 2013*, pages 3111–3119, Red Hook/NY. Curran Associates, Inc.

Bryan Orme. 2009. Maxdiff analysis: simple counting, individual-level logit, and HB. *Sawtooth Software, Inc.*

Sara Owsley Sood, Judd Antin, and Elizabeth F. Churchill. 2012. Profanity use in online communities. In *CHI 2012 — Proceedings of the 30th ACM SIGCHI Conference on Human Factors in Computing Systems. Austin, Texas, USA, May 5-10, 2012*, pages 1481–1490, New York/NY. Association for Computing Machinery (ACM).

Amir Hossein Razavi, Diana Z. Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Advances in Artificial Intelligence. Canadian AI 2010 — Proceedings of the 23rd Canadian Conference on Artificial Intelligence. Ottawa, Ontario, Canada, May 31 - June 2, 2010*, number 6085 in Lecture Notes in Computer Science (LNCS), pages 16–27, Berlin, Heidelberg. Springer-Verlag.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the Workshop on New Challenges for NLP Frameworks @ LREC 2010. La Valletta, Malta, May 22, 2010*, pages 45–50, Paris. European Language Resources Association (ELRA).

Nils Reimers, Judith Eckle-Kohler, Carsten Schnober, Jungi Kim, and Iryna Gurevych. 2014. GERMEVAL2014: nested named entity recognition with neural networks. In *KONVENS 2014 — Proceedings of the Workshops of the 12th Edition of the KONVENS Conference: GermEval. Hildesheim, Germany, October 8-10, 2014*, pages 117–120, Hildesheim, Germany. Universitätsverlag Hildesheim.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the reliability of hate speech annotations: the case of the European refugee crisis. In *NLP4CMC III — Proceedings of the 3rd Workshop on Natural Language Processing for Computer-Mediated Communication. Bochum, Germany, 22 September 2016*, number 17 in Bochumer Linguistische Arbeitsberichte (BLA), pages 6–9.

Magnus Sahlgren, Tim Isbister, and Fredrik Olsson. 2018. Learning representations for detecting abusive language. In *ALW 2 — Proceedings of the 2nd Workshop on Abusive Language Online @ EMNLP 2018. Brussels, Belgium, October 31, 2018*, pages 115—123, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Cícero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *ACL 2018 — Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Victoria, Australia, July 15-20, 2018*, volume 2: Short Papers, pages 189–194, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *SocialNLP 2017 — Proceedings of the 5th International Workshop on Natural Language Processing for Social Media of the AFNLP SIG SocialNLP @ EACL 2017. Valencia, Spain, April 3, 2017*, pages 1–10. Association for Computational Linguistics (ACL).

Hui-Po Su, Zhen-Jie Huang, Hao-Tsung Chang, and Chuan-Jie Lin. 2017. Rephrasing profanity in Chinese text. In *ALW 1 — Proceedings of the 1st Workshop on Abusive Language Online @ ACL 2017. Vancouver, British Columbia, Canada, August 4, 2017*, pages 18–24, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Sajedul Talukder and Bogdan Carbunar. 2018. ABUSNIFF : automatic detection and defenses against abusive FACEBOOK friends. In *ICWSM 2018 — Proceedings of the 12th International AAAI Conference on Web and Social Media. Stanford, California, USA, June 25-28, 2018*, pages 385–394, Palo Alto/CA. AAAI Press.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A dictionary-based approach to racism detection in

Dutch social media. In *TA-COS 2016 — Proceedings of the Workshop on Text Analytics for Cybersecurity and Online Safety @ LREC 2016. Portorož, Slovenia, 23 May 2016*, pages 11–17.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: a typology of abusive language detection subtasks. In *ALW 1 — Proceedings of the 1st Workshop on Abusive Language Online @ ACL 2017. Vancouver, British Columbia, Canada, August 4, 2017*, pages 78–84, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Aksel Wester, Lilja Øvrelid, Erik Velldal, and Hugo Lewi Hammer. 2016. Threat detection in online discussions. In *WASSA 2016 — Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis @ NAACL-HLT 2016. San Diego, California, USA, June 16, 2016*, pages 66–71, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018a. Inducing a lexicon of abusive words: a feature-based approach. In *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, Louisiana, USA, June 1-6, 2018*, volume 1: Long Papers, pages 1046–1056, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018b. Overview of the GERMEVAL 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of the* GERMEVAL *2018 Workshop @ KONVENS 2018. Vienna, Austria, September 21, 2018*, pages 1–10.

Zhelun Wu, Nishant Kambhatla, and Anoop Sarkar. 2018. Decipherment for adversarial offensive language detection. In *ALW 2 — Proceedings of the 2nd Workshop on Abusive Language Online @ EMNLP 2018. Brussels, Belgium, October 31, 2018*, pages 149–159, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Taijin Yoon, Sun-Young Park, and Hwan-Gue Cho. 2010. A smart filtering system for newly coined profanities by using approximate string alignment. In *CIT 2010 — Proceedings of the IEEE 10th International Conference on Computer and Information Technology. Bradford, UK, 29 June - 1 July 2010*, pages 643–650. IEEE.