Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6127–6135
Marseille, 11–16 May 2020

An Annotated Social Media Corpus for German

Eckhard Bick
University of Southern Denmark
Odense, Denmark
eckhard.bick@mail.dk

Abstract

This paper presents the German Twitter section of a large (2 billion word) bilingual Social Media corpus for Hate Speech research, discussing the compilation, pseudonymization and grammatical annotation of the corpus, as well as special linguistic features and peculiarities encountered in the data. Among other things, compounding, accidental and intentional orthographic variation, gendering and the use of emoticons/emojis are addressed in a genre-specific fashion. We present the different layers of linguistic annotation (morphosyntactic, dependencies and semantic types) and explain how a general parser (GerGram) can be made to work on Social Media data, pointing out necessary adaptations and extensions. In an evaluation run on a random cross-section of tweets, the modified parser achieved F-scores of 97% for morphology (fine-grained POS) and 92% for syntax (labeled attachment score). Predictably, performance was twice as good in tweets with standard orthography than in tweets with spelling/casing irregularities or lack of sentence separation, the effect being more marked for morphology than for syntax.

Keywords: Social Media corpus, Hate Speech, German Corpus Linguistics, Constraint Grammar, Syntactic parsing, Non-standard orthography, Emoji annotation

1. Introduction

Hate Speech (HS) against ethnic, religious and national minorities is a growing concern in online discourse (e.g. Foxman & Wolf 2013), creating a conflict of interest in societies that want to advocate freedom of speech on the one hand, and to protect minorities against defamation on the other (Herz & Molnar 2012). Social networks, under political pressure to take action, have begun to filter for what they perceive as outright hate speech. Reliable data, actionable definitions and linguistic research is needed for such filtering to be effective without being disproportionate, but also in order to allow policy makers, educational institutions, journalists and other public influencers to understand and counteract the phenomenon of hate speech.

The three-year research project behind the work described here is called XPEROHS (Baumgarten et al. 2019) and investigates the expression and perception of online hate speech with a particular focus on immigrant and refugee minorities in Denmark and Germany. In addition to experimental and questionnaire studies, data is collected from two major social networks, Twitter and Facebook. The material is used to examine and quantify linguistic patterns found in hateful discourse and to identify derogatory terms and outright slurs, as well as metaphors used in a demeaning and target-specific way. Finally the corpus is used to provide graded HS examples for the project's empirical work on HS perception.

2. The Corpus

2.1 Corpus Size and Sources

The XPEROHS corpus is a monitor corpus, where posts and comments were collected continuously from Twitter (late 2017 - mid 2019) and Facebook (late 2017 - mid 2018) for both German and Danish, using the networks' query APIs. Other social networks were considered, but not used because of their lack of picture-independent text or because Danish users were writing only in English. For Twitter (TW) a very high coverage was achieved using high-frequency function words as search terms (e.g. *und/og* [and], *oder/eller* [or], *der-die-das/den-det* [the], *er-sie-es/han-hun* [he-she-it], *ist/er* [is]). Harvesting Facebook (FB), on the other hand, is only possible using specific seed pages (e.g. political parties or politicians, TV sites and news media). Therefore, a quantitative comparison is only possible cross-language, not directly between data from the two networks, not least because the pre-selection of seed sites lead to a higher incidence of minority discourse in FB. On the other hand, the two media complement each other in terms of qualitative analysis. Thus, tweets are text-only and often short, public and "one-way", while FB is more multi-modal and, with its original friends-based, more symmetrical communication, more accommodating for actual turn-taking discourse.

All in all, the corpus contains over 2 billion words:

	Twitter	Facebook	
German	1,700 M	200 M	1,900 M
Danish	270 M	60 M	330 M
	1,970 M	260 M	2,230 M

Table 1: Corpus sizes

2.2 Data Selection and Filtering

Only 5-10% of the sentences in the corpus contain insults, slurs or otherwise demeaning language, similar to numbers for e.g. English Yahoo data (3.4-16.4%) reported by Nobata et al. (2016). Therefore, in the initial phase of the project, in order to facilitate a first manual inspection of the data, and excerption of test utterances for questionnaires, a number of smaller sub-corpora with a higher density of minority and "negativity" keywords were extracted in a boot-strapping approach, where the keyword lists were iteratively expanded based on inspection results. Similar methods were also used for HS corpora in other languages (e.g. Waseem & Hovy 2016), but create a filter-dependent bias (Klubička 2018) that can be problematic for finding new, unexpected and - above all - less explicit forms of hate speech, such as metaphors.

To avoid such a bias, and to allow quantitative evaluation, the unfiltered overall corpus is regarded as the main data set. This is an important difference to corpora that were built primarily to support machine-learned (ML) hate speech recognition for the purpose of removing it (automatically) from online communication. Such training corpora do exist not only for English (e.g. Waseem & Hovy 2016), but also for Italian (Sanguinetti et al. 2018) and German (Wiegand et al.), but are much smaller (16,000, 6,000 and 8,500 tweets, respectively, for these examples), because of pre-selection and manual HS annotation. Such small samples are less ideal for linguistic purposes, since their lexical coverage is low and less common constructions may not occur, or not in significant numbers. Even the advantage of having HS marked directly is not unproblematic in linguistic terms because of low inter-annotator agreement (Ross et al. 2016). Topic-driven or author-based data sets (e.g. Kratzke 2017 on German parliamentary elections) are larger, but do not necessarily provide a good cross section of hate speech and minority discourse, let alone do so for both our target languages in a comparable fashion.

2.3 Preprocessing, Anonymization and Pseudonymization

In order to comply with recent European legislation regarding the protection of personal data, meta data such as user identity, addressee and timestamp are stored in separate files, while the corpus itself only contains a number key for each tweet, post or comment. When performing corpus searches, the latter will allow project members to follow a link button to the original internet URL and study discourse interaction and rhetorical structure, or contextualize an utterance multi-modally (pictures, video, sound files), without being able to see metadata in the corpus extracts themselves.

User names occurring in the text itself cannot be safely removed without endangering the syntactic cohesion of the sentence, because they may function as e.g. subjects, objects or vocatives. Therefore, pseudonymization was used instead of anonymization, replacing user names with a dummy "twittername" throughout the corpus. For larger excerpts, in particular n-gram statistics, we also pseudonomize URLs, person names, publication titles, numerical expressions and dates. In addition to data protection, lumping e.g. person names together as one "word" has the advantage of making linguistic patterns more salient and statistically more significant[1].

A specific problem for the Danish Twitter data was "noise" from other languages, because the Twitter API, while featuring a language parameter, suffers from a certain amount of language confusion due to the character string similarity between Danish and the other Scandinavian languages, especially Norwegian variants and Swedish, as well as sometimes Dutch. We therefore used additional language filtering (Google, lexicon-based weighting, letter cluster weighting).

3. Linguistic Annotation

Many types of linguistic statistics, pattern identification and comparison are difficult or impossible to perform on text-only corpora, making it necessary to enrich the corpus with grammatical and lexico-semantic information. Thus, lemmatization hugely simplifies corpus searches, while part-of-speech (POS) and syntactic annotation and disambiguation facilitates pattern generalization and reduces the number of false positive hits. Specifically for German and Danish, morphological analysis is useful, since many words are out-of-lexicon compounds. Not least in HS research in-word modification with (derogatory) prefixes, suffixes and attributes is an important linguistic parameter. A more ambitious, semantic annotation will even allow the corpus user to look for patterns involving categories based on named entity recognition (NER), ontologies, verb frames and semantic roles. One example from the realm of hate speech is using the semantic type of animal to generalize searches for dehumanization patterns involving animal metaphors[2].

3.1 Choice of parser

Computer-mediated communication (CMC) is known to be a difficult genre to parse, as pointed out by e.g. Proisl (2018) in his work on a Social Media and Web Tagger, due to problems like out-of-vocabulary words (OOV), emoticons/emojis, interaction words (*lach* [laugh], *heul* [cry]), hash tags, URL's, onomatopoeia, orthographic variation and contractions (e.g. *'stimmts?* - is that correct?). Beißwenger et al. (2016) adds further features to this list, such as emphasis by upper-casing or letter repetition, discourse links (hashtags and user address), as well as the prevalence of colloquial syntax and colloquial particles (interjections, intensifiers, focus and gradation particles, modal particles and down-toners). Often a number of such non-standard traits is found in one single utterance (underlined):

> *In 5..10 Jahren sind die Deutschen eine Minderheit*
> *u.können die heutigen #Asylanten kostenlos*
> *verklagen*❣☺*Ich find' d.#Toleranzgesetz cool*
> [Germans will be a minority in 5-10 years and can sue
> today's #refugees for free. I just love the
> Discrimination Act]

In a sobering comparative study of 5 state-of-the-art parsers, Giesbrecht & Evert (2009) showed that performance dropped dramatically from the originally reported ~ 97% accuracies for part-of-speech to around 93%[3] when trained and evaluated on web data (DeWaC corpus). Proisl cites similar results for his own tagger, with 93.75% accuracy for CMC and 91.06% for web data. It is part of this cautionary tale that errors were not spread evenly across word classes. Thus, for the CMC domain, verbs had sub-par performance, with accuracies of 89.1% (finite verbs), 87.4% (infinitives) and 80% (imperatives).

[1] Because the lumped-together category is much more frequent than the individual name.

[2] To clarify: The corpus is not (manually) annotated for such metaphorical usage, but a semantically informed search will yield a concordance with a reasonable hit rate of interesting cases.

[3] The best accuracy for web texts was 93.8%, with a simplified (coarse-grained) tag set, and under 93% for the full tag set.

Proper nouns had an accuracy of 17.4%, and emoticons were not recognized in over half the cases.

Even within the online genre, cross-domain performance appears to be a problem. Therefore, it would be problematic to use training data, even where they do exist. For instance, Neunerdt (2013) found that their WebTrain training corpus improved tagging accuracy by 5% to 93.7% for web data (compared to TIGER treebank training), but still only achieved 89% accuracy for chat and 84% for YouTube comments.

Apart from non-standard domain traits, lexicon coverage appears to be decisive. Thus, Neunerdt reports an accuracy of 95.8% for known words, but only 68% for OOV words. Also, their tagger could "simulate" the change to web training data by adding a web lexicon to the original tagger, achieving only 1% less in accuracy.

In the light of these results, and in the absence of a dedicated training corpus for Twitter or Facebook, using an off-the-shelf tagger for our corpus would not have been ideal. Also, even the general web training data would only have been available for German, not for Danish. In addition, even a POS accuracy of 93% would mean a couple of POS errors in every sentence, each potentially propagating into multiple (e.g. syntactic) errors at higher-level parsing stages. Thus, while parsers trained on general treebanks have been adapted for Twitter, results at the depencency level are not yet ideal, even for English, e.g. 80% unlabeled attachment score for TWEEBOPARSER (Kong et al. 2014). Improved results (79.4% *labeled* attachment score) are reported by Liu et al. (2018) for a 20-parser ensemble, but were achieved with an in-domain (twitter) training treebank not available for our language pair.

Therefore, instead of using mainstream ML systems, we opted for rule-driven parsers, GerGram for German[4] and DanGram for Danish[5], that allow transparent genre adaptation at all levels. Both parsers use the Constraint Grammar formalism (Karlsson 1990, Bick & Didriksen 2015) to implement linguistic rules based on lexical information and contextual features, treating consecutive annotation layers of increasing complexity in a chain of modules, progressing from morphological analysis and POS disambiguation to syntactic function annotation, dependency trees and semantic annotation.

Using a rule-driven system has the advantage that all errors are completely transparent, and can be identified and addressed - given time and man-power - one by one. For instance, verb tagging can be improved by relaxing the bias against imperatives (they are very rare in news texts) and by adapting a few disambiguation rules to the fact that CMC sentences often have subject ellipsis, starting with a finite verb in the first person singular.

Also, both parsers feature a full morphological analysis with a reliable compound and affixation analysis, rather than just a lexicon with full-forms. For both German and Danish, this considerably reduces the problem with unknown words.

The following annotation fields can be distinguished, and are expressed as token-based tag lines:

(a) Wordform - can be a multi-word expression (MWE)

(b) Lemma - e.g. ‘wähle/wählst/wählt/wählte’ -> ‘wählen’ [choose]

(c) POS - e.g. **N**on, **V**erb, **ADJ**ektive, **ADV**erb, DETerminer)

(d) Inflection - e.g. **PR**esent, **NOM**inative, **S**ingular

(e) Syntactic function - e.g. **@SUBJ**ekt, **@ADVerbiaL**

(f) Secondary categories - e.g. **<mv>** main verb, **<dem>** demonstrative

(g) Semantic Categories - e.g. **<party>**, **<H...>** person, **<Q->** negative polarization)

(h) Dependency tag - e.g #2->5 (token 2 attaches to head token 5)

It should be noted that Constraint Grammar (CG) assigns these tag fields individually, even at the structural level (syntax, dependency and frame relations), and does not need complete "generative" parses to do so. In other words, there are no "unparsed" utterances, only local errors, which is a great robustness feature in the face of non-standard language input such as CMC.

3.2 Morphological Annotation

The morphological annotation level comprises wordform, lemma, part-of-speech and inflection categories. In order to improve readability and to facilitate feature searches, inflection categories are not fused with the POS tag, but kept in separate fields for gender, number, case, tense, person, mood etc.

Word recognition

A large portion of words in the CMC genre is not immediately recognizable even with inflectional analysis, and a number of strategies is used to handle such OOV words. Even non-analytical, single-word normalization can, according to Sidarenka et al. (2013), reduce the OOV rate in German Twitter data by 5-9% and improve overall POS tagging accuracy by 6.4%. Apart from general Twitter features (e.g. hashtags), the most common problems in our corpus were lower-casing of words that the regular parser expected to be uppercased (German nouns), upper-casing for emphasis and spelling errors/variation. Because of the risk of ambiguity, casing can not always be normalized through simple lexicon lookup, so a frequency-based heuristics is used, and in some cases, both lc and uc forms are passed to the contextual disambiguation rules.

We added a specially developed, automated spellchecking module to the morphological analyzer, limiting changes to the Levenshtein 1 level (1 letter substituted, inserted or deleted)[6]. For compounding languages like German and Danish, automatic spellchecking cannot be performed as a simple preprocessing step. Rather, compound analysis has

[4] https://visl.sdu.dk/visl/de/parsing/automatic/

[5] https://visl.sdu.dk/visl/da/parsing/automatic/

[6] Some common spelling errors above Levenshtein 1 were treated through lexicon additions.

to be performed first. Hereafter, we employ spellchecking (1) first for entire words, then (2) for potential compounds (i.e. words with a recognizable first or second part, but no match in the other half), and finally (3) for potential roots, after stripping off recognizable inflection endings.

All of the above normalization techniques are directed at individual words. However, non-standard tokenization is also an important problem. Thus, our data contain both non-standard "English style" compound splitting (a) and "colloquial" contractions (b) or elisions (c) typical of spoken language:

a1) *Kanaken Gang* [Middle-East-immigrant-slur gang]
a2) *Terroristen Pack* [terrorist scum]
b1) *packen <u>wirs</u> (wir-es) an* [let us- do -it]
b2) *<u>haste</u> (hast-du) das gesehen?* [have-you seen it]
c) *ich <u>find</u>' (finde) ihn geil* [I think he's cute]

We split the contractions (b) and assign full analyses to both parts, maintaining the fullform on the first part and marking the split on both. Fusion is marked by assigning a <pre-n> marker, a PREF (prefix) word class and a @PREF syntactic function tag, maintaining the original sub-tokens in order to preserve their individual tagging (e.g. semantic types).

Abbreviations

Abbreviations are a common, time- and space-saving feature of CMC text, creating both lexical and orthographic challenges. Orthographically, recognition is hampered by a frequent lack of either abbreviation dots (a) or the inter-token space (b), often in combination with casing irregularities.

a1) *zB, zB., z.B., z.b.* (zum Beispiel - e.g.)
a2) vll, vllt, vlt (vielleicht - maybe)
b) *Ja die kleinen PATEIEN haben erst Angefangen die,die <u>36Jahren</u> Geschlafen haben <u>u.kein</u> Geld für Einheimische ... <u>u.nicht</u> nur Miliarten <u>f.Flüchtlinge</u> <u>u.Ausländische</u> Bürger !* (Yes the small PATIES_SIC have only just Begun who, who have Slept for 36Years_CASE_ERROR a.no money for locals ... and not just bilioms_SIC f.refugees a.Foreign citizens!)

More interestingly, many abbreviations are abbreviations of multi-word expressions (MWE) and/or English expressions and have become lexemes in their own right, carrying a jargon specific meaning. These were collected and added to the parser lexicon:

WTF (what the fuck)
omg (oh my God)
ka (keine Ahnung - no idea)
kb (kein Bock - no desire to)

Compound analysis

In an evaluation of part of the immigrant-filtered subsection of the German Twitter corpus (11.8 million words), about 10% of all words were compounds or derivations (not counting separable verbs), with the lion share (2/3) being noun compounds. Of these, 84% had a lexicon-entry[7], albeit with a 5% error rate for the listed compound analysis. One in six (16%) were OOV, i.e. found through live compound analysis, and about 2/3 of the latter were flagged as high-confidence compounds by the parser, 1/3 as low-confidence. 3% of the high-confidence OOV compounds were false positives (e.g. '*Profiteure*' [profiteers] = *Profi+teuer* [professional + expensive]), but rarely in the sense of a wrong compound-split. The most common error was the analysis of proper noun as a compound common noun (e.g. *Nickel~s+dorf*)[8], followed by foreign words and name derivations.

Low-confidence compounds had 6 times as many (17%) false positives. However, only 1/4 of these were ordinary errors regarding correctly spelled words (again often names), while most were last-ditch efforts to assign an analysis to words with missing spaces ('dieMehrheit' [the majority]), spelling errors and casing anomalies ('moslem-wichser' [muslim jerk]) or orthographically puns ('umFAIRteilen' [distribute fairly]).

By comparison, the more standard genre of online news text[9], using the same metrics, had more compounds and derivations (12.8% in all, 8.8% for nouns), albeit with a slightly lower OOV ratio (1/7)[10]. Escartín et al. (2014) reported an even higher density of 11-14% for nominal compounds in German technical texts, as well as a much higher rate of OOV compounds (over 60%) based on a German monolingual dictionary. The relative distribution of good and "maybe" compounds was the same, as were the error percentage and types of the high-confidence (good) compounds, while the low-confidence compounds in the news corpus had half as many false positives, mostly because orthographical errors and variation was rare.

Since there is no universally accepted definition of what a compound is, and because the lack of a gold standard for Twitter data, it is difficult to make a direct comparison with other results. Using alignment of NPs and PPs in the German-English section of the Europarl Parallel corpus, Koehn & Knight (2003) achieved a precision of 93.8% and a recall of 90.1% for that genre. For our data, the easiest to approximate is precision: Thus, with a correctness rate of 95% for the 84% in-lexicon words and a correct share of 92% (97%*2/3+83%*1/3) of the 16% OOV cases, a rough estimate is a precision of 0.95*84% + 0.92*16% = 94.5%.

Gendering

Contemporary German has become fairly gender-aware in orthographical terms, not least in opinionated discourse on social media.

[7] The parser lexicon also marks Greek-Latin word parts, e.g. 'Atmo+Sphäre' (atmosphere) or even 'Tour+ismus' (tourism), which don't fit everybody's idea of a compound, so depending on the definition, the compound share could be lower. The OOV compounds, on the other hand, draw their building blocks only from regular German vocabulary, and percentages are thus less open to discussion.

[8] This error is typical of German, because common nouns are uppercased and thus not easily distinguishable from proper nouns.

[9] A 2015 news sample from the Leipzig Corpora collection: http://wortschatz.uni-leipzig.de/en/download/

[10] The OOV difference may have been influenced by orthographic variation specific to the Twitter corpus.

While not introducing new pronouns like Swedish, German has long provided an option for marking gender-neutrality on person nouns by adding *-In* (Sg.) or *-Innen* (Pl.), optionally preceded by a separator (*, /, # or _). In the presence of a separator, the suffix can also be found in lower case. This variation presented a challenge to the unmodified parser that would break these words on the separator or - if not - fail to match them in its lexicon. Our solution is preprocessing and passing a standardized female form on to the analyzer while maintaining the marked surface form in parallel.

In the corpus as a whole, the most frequent were *UnterstützerInnen, KollegInnen, SchülerInnen, BürgerInnen, MitarbeiterInnen, JournalistInnen, WählerInnen, LehrerInnen, PolitikerInnen, TeilnehmerInnen* etc. Most are profession or agent terms ending in *-er* or *-ist*, that in traditional grammar both allow the female *-in* affix. However, already at rank 14, we find a term like *GrünInnen* that did not exist before gendering (**Grünin)*, and there are others like e.g. *RabaukInnen (*Rabaukin).* Interestingly, and problematically for morphological analysis, the suffix also appears after a plural ending (rather than between stem and plural): *GrüneInnen, FreundeInnen* (correct: *FreundInnen*) or even with a singular gendering suffix attached to a plural form: PädagogenIn. Singular forms were rare (0.4%), and the gendering suffix does not occur with female forms: *MasseurIn*, but not **MasseusIN* or **MasseuseIN.*

Almost 2/3 of instances had a separator (table 2), the most common one (*) being slightly more frequent (37.9%) than not using a separator (35.6%). Singular forms were rare (11.7% and 4.7%, respectively, for words with and without a separator).

Separator	-In(nen)	-in(nen)	
*	**36.6%**	1.3%	37.9%
_	11.2%	0.5%	11.7%
/	9.2%	**0.5%**	9.7%
#	5.1%	0.04%	5.1%
none	**35.6%**	-	35.6%
	97.7%	2.3%	100%

Table 2: Orthographic variation of gendering suffixes

This paper is about the corpus and its annotation rather than the actual HS research based on these data, but assuming that political correctness is bundled across topics, than one would assume that explicitly neutral gendering inversely correlates with minority discrimination and HS expressions. A tentative analysis of tweets containing '*MuslimInnen/Musliminnen'* does indeed support this (table), even though the lower-case variant is still slightly more marked than '*Muslim*' and contains examples of the alternative gendering strategy '*Musliminnen und Muslime*'. Thus, '*-Innen*'-tweets rarely express a negative attitude towards the target group (table 3), and have a higher incidence of counter-speech (i.e. criticizing discrimination).

attitude	Musliminnen	MuslimInnen
negative	17%	5%
counter-speech	40%	48%
hedged criticism	3%	2%
unclear, neutral	40%	45%

Table 3: Attitude linked to -Innen/-innen

3.3 POS and Syntactic Annotation

The main task of the parsing grammars is morphosyntactic (or semantic) disambiguation and the mapping of function tags and dependency links. However, a number of non-standard syntactic constructions found in our data had adverse effects on parser performance and called for additional rules and the tuning of existing rules.

For instance, subject-less finite sentences (a1), otherwise impossible in written German, do occur in 1. person CMC discourse, but without the pronoun, the 1. person verb reading risks being removed in favour of a homonymous and grammatically correct imperative (e.g. *schaue* - look) or noun (e.g. *Glaube* - 'faith' vs. 'think'). The distinction is subject to the additional twist that a hashtag, instead of being just a comment, can fill the missing subject slot (a2).

Similarly, default German parsing rules assume that every main clause contains a finite verb, and will therefore misread homonymous infinitives (b) or participles (c) as finite forms, if confronted with non-finite main clauses. Infinitive main clauses do occur in recipes and with imperative function (*nicht aufmachen* - don't open), but are absent in ordinary text corpora. In spoken language and CMC, however, the construction is normal, possibly as a generalization tool (b1).

(*a1*) *Schaue aus dem Fenster, Sonne scheint.* (Looking out of the window, sun is shining)

(a2) *#Italien wollte Einwanderunspolitik für #Europa machen* (#Italy wanted to make immigration policies for #Europe)

(b1) *Am Bahnhof Flüchtlinge vor Kameras beklatschen und im nächsten Jahr die eigenen Kinder auf ne Privatschule schicken.* (Applaud_INF refugees at the station and the next year send_INF your own children to a private school)

(b2) *am freitag gehn wir in die rofa, wär bock har mirzugehn, einfach mal melden* (On Friday we'll go to Rofa, who wants to join, simply let_INF me know)

Participle main clauses in the passive are typical of headlines (e.g. *Tourist von Bär gebissen* - 'Tourist bit by bear'), and were already handled by the parser. Active participle clauses (c), however, are not found in ordinary written German, so the parser would misread them for passives or - sometimes - finite forms (e.g. *erhalten* - 'got').

(c1) *Schon zurück, weil Vorlesung geschwänzt* (Already back, because [have] skipped_PCP lecture)

(c2) *Gestern statt Einkaufszettel Handyfotos leerer Packungen einzukaufender Dinge gemacht* . (Yesterday instead of shopping list [have] made_PCP phone pictures of empty boxes of buyables)

For German, the assignment of syntactic functions is heavily case-dependent (e.g. nominative -> subject or subject complement, accusative -> direct object of measuring adject). At the clause level, the parser exploits the uniqueness principle to disambiguate case and function. However, this method breaks down if additional nouns are introduced, without coordination, in a non-grammatical fashion. In CMC, this can happen by omitting list- or clause-delimiting punctuation or by faulty, space-split compounding (d). We address the former by punctuation-mapping rules and by relaxing punctuation-dependent rules, and the latter by introducing special prefixing tags (<pre-n>, PREF and @PREF for secondary tagging, POS and syntactic function, respectively).

(d) *wenn der alman der einzige ist mit dem Führerschein und Papas Mercedes in der Kanaken Gang*. (When the German_SLANG is the only one with a driving license and dad's Mercedes in the foreigner_SLUR gang)

However, recognizing a noun as a (wrongly) split first part of a compound, rather than an individual constituent, is by no means a trivial task - in part because many compounds are productive and not listed in the lexicon, but also because each noun has its own rules as to case and number when attaching to another noun. We therefore computed likelihoods for these features from existing compounds in the lexicon. Rules can then use both contextual clues (e.g. prenominal gender-number-case agreement) and morphological probability thresholds to identify compounding errors.

3.4 Semantic type annotation

Semantic information constitutes a valuable - and less common - additional layer of corpus annotation. For the parser itself, semantic information is used contextually in the disambiguation of other, lower-level categories. In the context of the hate-speech project, it facilitates the batch-wise extraction and comparison of e.g. nationality or ideology terms or the search for animal, disease and other dehumanizing metaphors.

For both Danish and German, semantic lexical types are annotated for nouns, adjectives and some adverbs, while the Danish parser (not described here) also addresses semantic function, assigning verb frames and semantic roles. The backbone of the semantic type system is a shallow noun hierarchy[11] with about 200 categories. Upper-level categories such as <H> (human), <food>, <tool> or <L> (location) are further subdivided into lower-level categories such as <Hprof> (profession), <Hideo> (follower of an ideology), <Hnat> (national), <Hfam> (family term), <Lh> (human-functional place), <Ltop> (natural-topological place), <Lciv> (civitas/town/country) etc. The scheme provides an easy way to lump categories and to work with either fine-grained or coarse-grained features. For proper nouns, 7 main categories are recognized: <hum> human, <org> organization, <inst> institution, <occ> organized event, <brand>, <tit> (title/work-of-art) and <L> location. The latter contains a number of sub-classes such as <Ltown>, <Lcountry> and <Lwater>. In addition, two special <org> categories (<media> and <party>) are distinguished as well as a handful of special categories.

The semantic scheme for adjectives contains about 110 categories grouped into 14 primary and 25 secondary umbrella categories. People adjectives, for instance can be <jpsych> (feelings), <janat> (body features), <jage>, <jsick> etc. The largest/default umbrella category, "property", is itself subdivided into secondary hypernym categories such as "measurable property" (<jsize>, <jweight>, <jtemp> [temperature], <jspeed>) and "physical aspects" (<jshape>, <jcol> [color], <jsub> [composition], <jmat> [material]. The semantic type tags are supplemented by additional tagging for domain and polarity. The purpose of polarity tagging is two-fold: First, it allows binary distinctions, such as <jtemp> <Q+> = *warm*, compared to <jtemp> <Q-> = *cold*. Second, the Q+/Q- tags double as sentiment markers, with Q+ chosen for the polarity that either literally or metaphorically is the one more often associated with a positive sentiment. Where this is impossible, or contradictory, Q0 (no polarity) or Q+/Q- (double polarity) is used.

3.5 Emoticons and Emojis

Emoticons are pictorial representations of facial expressions, using punctuation symbols and, to a lesser degree, a few numbers and characters. Emojis are a newer, pictorial version, resembling ideograms or actual pictures, covering not only emotions, but also a wide range of objects, actions, places etc. Emoticons/emojis are an interesting annotation topic for two very different reasons: First, they present a formidable obstacle for a parser that has not been designed to handle them. All other things being equal, text emoticons will end up split into punctuation "atoms", and emoji strings as OOV foreign nouns. The former can lead to sentence discontinuities and structural parsing errors, while the extra constituents spawned by the latter may affect NP cohesion, interfere with uniqueness rules, or even mask adjacent words into OOVs in the absence of a separating space character. The second reason for taking emoticons seriously in our project is the obvious one - their emotional content. Thus, if correctly recognized and annotated, emoticons can help to decide the degree of HS of a given utterance, or even help to search for new hateful content.

Several parser adaptations were necessary to handle this new category, the first step being a pattern-based recognition of text emoticons as character/punctuation strings and a separation and marking of emojis based on Unicode blocks. These will then be annotated semantically and assigned a part-of-speech. Because of its relatively free distribution in the sentence, ADV (adverb) is used for for all emoticons and most emojis. An exception are flag emojis, that get tagged as proper nouns, because they are sometimes used instead of country names in our data (a).

The semantic tagging lumps emoticons into 10 emotional categories (e.g. "emo-happy", "emo-love", "emo-sad", "emo-angry" etc.) that are used as "lemma" for a group of emoticons. Emojis get individual lemmas (e.g. "emo-gesture-Left-Facing-Fist"), but - where relevant - with a prefix indicating one of the 10 emotional umbrella categories (e.g. "emo-laugh-Face-With-Tears-of-Joy").

[11] For a full list of types, cf. http://visl.sdu.dk/semantic_prototypes_overview.pdf

(a) *er ist gut für ein starkes 🇫🇷 während 🇩🇪 vor die Hunde geht* (he is good for a strong France, while Germany goes down the drain)

(b) *Rentner sammeln Flaschen und Flüchtlinge leben auf großem Fuß* 😠 🤛 (Pensioners collect bottles and refugees live in style)

(c) *Unsere fiese , unmenschliche Regierung ist einfach zu gemein zu den armen Flüchtlingen . :(* (Our nasty, inhumane government is simply sooo mean to the poor refugees)

Though it nicely fits the "living-at-our-expense" narrative, interpreting the example in (b) as a hateful remark and finding it in a billion-word corpus is difficult with ordinary annotation alone. Even if "*auf großem Fuß*" gets recognized as a fixed expression, it risks getting flagged as positive in isolation. Together with an angry-face emoji and a left-facing fist the meaning is clear, and the example likely to find its way into a concordance. The textual sadness emoticon in (c), finally, is a means to underline (and identify) the intended ironic interpretation of the utterance.

4. Parser Evaluation

Though Twitter data make up part of existing CMC evaluation data sets, this is - to the best of our knowledge - the first time that a German parser has been evaluated specifically on separate, mono-modal Twitter or Facebook data, possibly due to the fact that most state-of-the-art parsers are ML systems and require a sufficient amount of separate in-domain training data to work optimally on a specific subdomain.

For the standard treebank domain, the CoNLL X shared task on dependency parsing (Buchwald & Marsi 2006) provides an early general baseline for dependency parsing that can be regarded as a kind of minimum performance that current parsers should be able to surpass. Here, the best system for German achieved a labeled attachment score (LAS) of 87.3 .

More specifically for the German CMC domain, in another shared task, EmpiriciST 2015 (Beißwenger et al. 2016), the best system achieved an accuracy of 87.33% for the original tag set and 90.20% with the simplified STTS[12] tagset, where domain specific tags, such as modal and focus particles, emoticons, colloquial verb contractions and "e-words" (URL, emails, @user, #hashtag) were not counted as errors if confused with their respective parent POS categories (adverb, finite verb, XY, respectively). However, syntactic labels and dependencies were not part of the EmpiriciST task.

Our own evaluation was carried out on a 5000 token cross section of the Twitter corpus, selecting tweets with id's *ending* in '00000' in order to avoid adjacent tweets and a period- or topic-dependent bias. A few broken tweets, caused by initial harvesting problems, were removed, the rest was annotated with the improved/appended version of GerGram, and evaluated through a 2-pass manual inspection. Obviously, this is a softer evaluation method than the ones used in CoNLL X and EmpiriciST, that had access to separate, multi-annotator gold corpora. On the other hand, when using training data for ML, a gold corpus is simply a reserved sub-section of the corpus, and compatibility of the tag set is automatically ensured when using an ML method. However, when a parser is trained on one treebank and measured on another, evaluation can be biased due to different encoding schemes (Rehbein & van Genabith 2007). A rule-based system, with its own tagset, is a worst-case scenario in this regard, because it cannot be re-trained. Therefore, in our case, evaluation with an external gold corpus was impractical, and creating one from scratch manually was deemed excessive, since the resource would be needed only for evaluation, not for training, and because parser evaluation is an - unfunded - side issue for the HS project as a hole[13].

Still, though not directly comparable, evaluation results (table 4) were encouraging, with F1 scores of 97% for POS+morphology and 92.3% for labeled attachment (LAS), respectively. These numbers were calculated for syntactic words rather than tokens, ignoring punctuation[14] and sentence-initial/final Twitter names that were not part of the syntactic tree. Computed as a token ratio, numbers would be about 0.9 percentage points higher on average.

	Recall %	Precision %	F1 %
POS (coarse)	98.54	98.54	98.54
POS + inflection + lemma	96.94	96.94	96.94
Syntactic function	93.55	93.60	93.58
Dependency links	94.18	94.18	94.18
LAS (synt. + dep.)	92.02	92.08	92.05
All errors	91.32	91.38	91.35

Table 4: Parser performance

These results should be seen against the backdrop of the general difficulty of the genre, with its many orthographical aberrations, missing punctuation etc. Thus, 40-50% of all POS and syntactic function errors (table 5) occurred in tweets with orthographic issues (e.g. spelling errors, creative abbreviations, casing etc.):

	% of error type occurring in problem tweets	error % in problem tweets (~1/4)	error % in other tweets (~3/4)
POS (coarse)	53.45	2.92	0.92
POS + inflection + lemma	44.26	5.09	2.33
Syntactic function	39.31	9.53	5.30
Dependency links	33.62	7.36	5.27
LAS (synt. + dep.)	39.30	11.79	6.60
All errors	38.44	12.55	7.28

Table 5: Parser performance

[12] Stuttgart-Tübingen tagset for spoken German

[13] That said, our inspection-evaluated sample could in theory be used by others as a small future gold corpus, given a compatible tokenization and tagging scheme.

[14] With the exception of punctuation emoticons

As can be seen from table 5, error density is about twice as high in orthographically problematic tweets, the difference being more marked for POS/morphology than for syntax.

A break-down of error types (table 6) identified false positive noun readings and intra-class confusions between the four pronoun/article classes[15] as the most common POS error sources. Inspection suggests that nouns (N) end up as a default (mis)reading of most other classes when the latter are OOV or orthographically unrecognizable. At first glance, verbs appear to be a fairly stable category, but the intra-POS confusion between finite verbs and active/passive participles or infinitives (in parentheses) is more important than for most other POS, because for verbs it propagates into structural errors and disambiguation errors in clause constituent categories.

correct POS tagged POS	N	ADJ	ADV	PRON ART	V	PR OP	rest	
N	-	3	2	-	3	6	7	21
ADJ	-	-	-	-	2	-	1	3
ADV	1	3	-	2	2	-	1	9
PRON/ ART	-	-	-	7	1	-	2	10
V	1	1	1	-	(12)	-	-	3
PROP	2	-	-	-	-	-	2	4
rest	-	-	4	1	-	-	-	5
	4	7	7	10	8	6	13	55

Table 6: POS error confusion table

At the syntactic level, due to the larger set of function categories, there were 128 different confusion pairs, even without taking into account attachment errors. However, only 13 pairs occurred more than 3 times, the most frequent being the classical ambiguity between postnominal and adverbial PP-attachment and the semantically important confusion between subject and accusative or dative objects. Judging from false positives and negatives, the GerGram parser has a bias towards adverbial over postnominal attachment, but is "error-balanced" with regard to subjects and objects. Because German has case constraints both for clause-level constituents and for arguments of prepositions, there is also some confusion between e.g. direct objects and accusative preposition arguments.

5. Conclusion and Outlook

We have presented a new German Social Media corpus for Hate Speech research and shown how a general parser can be adapted and extended to annotate such data at various linguistic levels, achieving satisfactory results for both POS/morphology (F=97%) and LAS/syntax (F=92%). Apart from lexicographical work (jargon, special abbreviations etc.) and better morphological analysis (compounding, gendering, emoticons), syntactic rule changes and additions were also necessary due to (a) non-standard clause structures (e.g. infinitive and participle clauses) and (b) the ambiguous role of hashtagged words as either add-on comments or integral parts of the syntactic tree.

A high incidence of orthographic errors and variation proved to be disruptive to ordinary parsing, and had to be addressed by integrated spellchecking and various normalization techniques. Even so, 1/4 of tweets still contained surviving abnormalities, causing a two-fold increase in tagging errors, obviously worthy of further work. In addition, future work should focus on the semantic level, complementing the existing semantic type annotation with a sematic role and frame layer, something that has been implemented for one of the project languages, Danish, but not yet for the German section of the corpus.

6. Acknowledgements

The XPEROHS project as a whole is funded jointly by the Velux Foundations and the University of Southern Denmark. Work on the GerGram and DanGram parsers was carried out in cooperation with GrammarSoft ApS.

7. Bibliographical References

Baumgarten, N., Bick, E., Geyer, K., Iversen, D. A., Kleene, A., Lindø, A. V., Neitsch, J., Niebuhr, O., Nielsen, R., Petersen, E. N. (2019). Towards Balance and Boundaries in Public Discourse: Expressing and Perceiving Online Hate Speech (XPEROHS). In: Mey, J., Holsting, A., Johannessen, C. (ed.): RASK - International Journal of Language and Communication. Vol. 50., pp. 87-108. University of Southern Denmark.

Beißwenger, M., Bartsch, S., Evert, S., and Würzner, K.-M. (2016). EmpiriST 2015: A shared task on the automatic linguistic annotation of computer-mediated communication and web corpora. In: Paul Cook et al. (ed.): Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task. pp. 44-56. Berlin: Association for Computational Linguistics.

Bick, E., Didriksen, T. (2015). CG-3 - Beyond Classical Constraint Grammar. In: Beáta Megyesi: Proceedings of NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania. pp. 31-39. Linköping: LiU Electronic Press

Buchholz, S.; Marsi, E. (2006). CoNLL-X shared task on Multilingual Dependency Parsing. In: Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X). pp. 149-164. ACL

Escartín, C, Peitz, S., Ney, H. (2014). German Compounds and Statistical Machine Translation. Can they get along? In: Proceedings of Proceedings of the 10th Workshop on Multiword Expressions (MWE). ACL. pp. 48-56

Foxman, A, Wolf, C. 2013. Viral Hate: Containing Its Spread on the Internet. New York: St. Martin's Press.

Giesbrecht, E, Evert, S. (2009). Is Part-of-Speech Tagging a Solved Task?An Evaluation of POS Taggers for the German Web as Corpus. In: Proceedings of the 5th Web as Corpus Workshop (WAC5), San Sebastian, Spain, (2009)

Herz, M., Molnar, P. (eds.). 2012. The Content and Context of Hate Speech. Rethinking Regulation and Responses. Cambridge: Cambridge University Press.

[15] Personal pronouns (PERS), independent pronouns (INDP), determiners (DET) and articles (ART). In addition, relatives and interrogatives were treated as different POS

Karlsson, F. (1990). Constraint Grammar as a Framework for Parsing Running Text. In: Proceedings of the 13th conference on Computational Linguistics - Vol. 3, pp. 168-173. ACL

Koehn, Ph., Knight, K. (2003). Empirical Methods for Compounds Splitting. In: Proceedings of EACL 2003 (Budapest). ACL. pp. 187-193

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., Smith, N. A. (2014). A dependency parser for tweets. In: Proceedings of EMNLP (Doha, Qatar). pp. 1001–1012.

Liu, Y., Zhu, Y., Che, W., Qin, B., Schneider, N., Smith, N. A. (2018). Parsing Tweets into Universal Dependencies. In: Proceedings of NAACL: Human Language Technologies (2018 (New Orleans). pp. 965-975

Neunerdt, M., Trevisan, B.,Reyer, M., Mathar, R. Part-of-Speech Tagging for Social Media Texts. (2013). Language Processing andKnowledge in the Web. pp. 139-150. Springer

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive Language Detection in Online User Content. In: Proceedings of the 25th International Conference on World Wide Web, WWW'16 (Montréal). pp. 145-153

Kratzke, N. (2017). #BTW17 Twitter Dataset - Recorded Tweets of the Federal Election Campaigns of 2017 for the 19th German Bundestag. [Data set]. Data. Zenodo. http://doi.org/10.5281/zenodo.835735

Proisl, T. SoMeWeTa: A Part-of-Speech Tagger for German Social Media and Web Texts. In: Proceedings of ELREC 2018. pp. 665-670

Rehbein, I.; van Genabith, J. (2007). Tree Annotation Schemes and Parser Evaluation for German. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). pp. 630-639. ACL

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., Stranisci, M. (2018). An Italian Twitter Corpus of Hate Speech against Immigrants. In: Proceedings of the 11th Conference on Language Resources and Evaluation (LREC2018), Miyazaki, Japan. pp. 2798-2895

Sidarenka, U., Scheffler, T., Stede, M. Rule-Based Normalization of German Twitter Messages. (2013). In: Proceedings of the GSCL Workshop: Verarbeitung und Annotation von Sprachdaten aus Genres internetbasierter Kommunikation.

Waseem, Z., Hovy, D. (2016). Hateful Symbols or Hateful People? - Predictive Features for Hate Speech Detection on on Twitter. In: Proceedings of the NAACL Student Research Workshop (San Diego, California). ACL. pp. 88-93

Wiegand, M., Siegel, M., Ruppenhofer, J. (2018). Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. (2018). In: Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna. pp. 1-10

8. Language Resource References

For the time being, the XPEROHS corpus is accessible only to project members at a secure server through a password-protected search interface. Due to GDPR concerns and data provider provisions it is still unclear if, when and how (parts of) the corpus can be made accessible to external research.