CODE ALLTAG — A German-Language E-Mail Corpus

Ulrike Krieg-Holz,[1] Christian Schuschnig,[1] Franz Matthies,[2] Benjamin Redling,[2] Udo Hahn[2]

[1] Alpen-Adria-Universität Klagenfurt, Universitätsstraße 65-67, A-9020 Klagenfurt, Austria
[2] Friedrich-Schiller-Universität Jena, Fürstengraben 30, D-07743 Jena, Germany

corresponding author: ulrike.krieg-holz@aau.at

Abstract

We introduce CODE ALLTAG, a text corpus composed of German-language e-mails. It is divided into two partitions: the first of these portions, CODE ALLTAG$_{XL}$, consists of a bulk-size collection drawn from an openly accessible e-mail archive (roughly 1.5M e-mails), whereas the second portion, CODE ALLTAG$_{S+d}$, is much smaller in size (less than thousand e-mails), yet excels with demographic data from each author of an e-mail. CODE ALLTAG, thus, currently constitutes the largest e-mail corpus ever built. In this paper, we describe, for both parts, the solicitation process for gathering e-mails, present descriptive statistical properties of the corpus, and, for CODE ALLTAG$_{S+d}$, reveal a compilation of demographic features of the donors of e-mails.

Keywords: text corpus, German language, e-mails

1. Introduction

Much of the success of modern human language technology can be attributed to the provision of large amounts of linguistic data in terms of corpora of written or spoken language. With respect to written language, we observe two major streams of work. The first of these aims at building large-sized *national reference corpora* as a 'representative' mirror of authoritative, canonical language use for one national idiom. For the English language, e.g., an American (ANC) (Ide, 2009) and a British (BNC) (Aston and Burnard, 1997) variant has been assembled, for the German language, currently two synchronous resources exist, DEREKO (Kupietz et al., 2010) and DWDS CORE (Geyken, 2007); as a diachronic extension going back until 950 AD, the DTA (*Deutsches Text-Archiv*) (Geyken and Gloning, 2015) is currently going to be set up as an historic complement. These reference corpora originally started as raw data collections aiming at high coverage, while in the meantime they have been continuously enriched by linguistic metadata, such as part-of-speech tags, lemma information and even parsing data (cf., e.g., Andersen et al. (2008)).

The second stream of work is dominated by computational concerns in terms of *(meta)language banking*. NLP researchers took mainly newspaper/newswire material as a basis for in-depth annotation efforts, on a much smaller scale than the national reference corpora though, and built up value-added resources containing, e.g., extensive syntactic or semantic metadata associated with linguistic raw data. This approach is most prominently featured in the seminal work on the English PENN TREEBANK (Marcus et al., 1993) and PROPBANK (Palmer et al., 2005), respectively. For German, there are currently two major competing syntactic annotation efforts, namely the TIGER (Brants et al., 2004) and the TÜBA-D/Z Treebank (Telljohann et al., 2004), whereas propositional information is primarily featured in the SALSA corpus which builds on the syntactic annotations of TIGER (Burchardt et al., 2006).

In basically all these efforts related to national reference corpora or metadata banks, for German and other languages, a preference for documents from skilled professional, educated writers can be observed since newspaper/journal articles, books (including both literary works as well as non-fiction material), technical papers and manuals constitute the vast majority of input texts to both types of resources. As a consequence, proper language use from high-end language performers usually exhibiting a maximum of conformance with formal language rules and a profound repertoire of communication habits is overrepresented in these collections.

With the rapidly emerging interest of (computational) linguists in the communication patterns of the World Wide Web, this high-performance bias is going to be replaced by the opposite extreme. We currently witness many activities directed at the investigation of computer-mediated natural language communication in social media, such as blogs, chats or tweets (Atefeh and Inkpen, 2015). These sources cover a heterogeneous variety of language performers and thus incorporate very diverse socio-economic backgrounds and different degrees of writing proficiency. Rather than the monologic format favored in national reference and language banking corpora, communication on these platforms is usually dialogue-oriented (often, occurring as n-party dialogues, $n \gg 2$), with evident preference for highly informal, colloquial language usage, often a written surrogate for sloppy spoken language behavior.

As a compromise bridging these two extremes of language use, we concentrate our activities on e-mail communication. E-mails are a particularly interesting text genre since they cover a wide range of everyday language usage, ranging, if properly selected, from private and personal communication among family members, friends, and colleagues (thus close to the informal communication habits of blogs, chats or tweets), to (semi-)formal interactions among employees and employers, clients and companies, citizens and public administration bodies, etc. (thus close to the requirements of formal communication). Yet, e-mails (just as contributions to blogs or chats) originate more likely from nonprofessional writers than the material contained in national reference corpora. On the other hand, writers of e-mails may face formal communication requirements much more often than authors of blogs, chats or tweets. Hence, the mixture of discourse constraints and contexts, formality requirements and levels of professionalism as writers make e-mails a unique text genre worth being studied in more depth than has been done, up until now.

Accordingly, we started to set up CodE Alltag, a mixed corpus of everyday German-language use embodied in e-mail discourse (**Co**rpus **d**eutschsprachiger **E**-Mails der **Alltag**skommunikation). This corpus currently consists of two partitions—one contains a bulk collection of e-mails publicly available from an e-mail archive, the other not only supplies e-mails as raw data but, in addition, each collected e-mail is accompanied by demographic data of its writer. Only very few written language corpora currently dispose of such author-specific information which may be crucial for explaining language use patterns. One of these rare examples is the study by Schwartz et al. (2013) who investigate the language in social media for the purpose of identifying personality traits of subjects by exploring their wording in a sample of 14.3M Facebook messages using standard personality questionnaires, including reports for age and gender, for approximately 75,000 volunteers.

2. Related Work

There is currently an enormous interest in the automatic analysis of computer-mediated natural language communication, such as blogs, chats or tweets, most notably for applications such as sentiment analysis (Balahur, 2013), opinion mining (Sokolova and Lapalme, 2011), forensic linguistics, with focus on authorship identification (MacLeod and Grant, 2012), and cyber intelligence (Iqbal et al., 2012). Besides the rapid development of dedicated text analytics software, numerous efforts have been undertaken in building up corpus resources backing up the evaluation and, if needed, training and development of this type of software. Whereas an abundance of (even annotated) linguistic data have been assembled from blogs, chats or tweets (cf., e.g., Saurí et al. (2014), Song et al. (2014), Scheffler (2014), Uthus and Aha (2013), Beißwenger et al. (2013), Pak and Paroubek (2010)), e-mail corpora are still quite rare.

Perhaps the most famous and, so far, largest among those few e-mail text corpora is the Enron corpus (Klimt and Yang, 2004), which contains almost 620,000 English-language messages with more than 30,000 threads; it became publically available as a consequence of US court decisions. Second in size to the Enron corpus, the W3C corpus[1] was generated by Web crawlers from mailing lists and public Web pages from the World Wide Web Consortium (W3C)[2] and is composed of 200,000 documents with more than 50,000 threads. For Trec 2007, another large-sized corpus comprising approximately 237,000 English documents was assembled for a spam filtering task using Web crawlers (Cormack, 2007). The fourth major English-language e-mail corpus is the *Australian National Corpus* (Lampert, 2009) which emerged as the result of a nationally spread request for submitting e-mails (*E-Mail Australia*) based on an 8-category content scheme; it comprises 10,000 e-mails.[3]

Within Trec 2006, a bilingual e-mail corpus was assembled by Web crawlers for a spam/non-spam classification task for the English (roughly 99,000 documents) and the Chinese language (almost 65,000 documents) (Cormack, 2006). Another bilingual corpus collected for the Tat author profiling system contains about 9,800 and 8,000 documents of English and Arabic e-mails, respectively (Estival et al., 2007). Two Chinese e-mail corpora (amounting to 10,000 and 74,300 sentences) are accessible from the *King-Line Data Center*[4] (last accessed on March 10, 2016).

For the German language, two e-mail corpora have been developed, up until now. The largest one, the Flag corpus, contains approximately 120,000 sentences from the *Internet Usenet Newsgroup* (Becker et al., 2003), whereas Declerck and Klein (1997) earlier assembled a small-sized e-mail corpus (Cosma) made of 160 e-mails. CodE Alltag not only differs in size by orders of magnitude from both corpora but also in qualitative terms—its broader stylistic and linguistic variability, avoiding the focus on colloquial, typically erroneous language use only (characteristic of the Flag corpus), and its thematically much more general scope (Cosma e-mails focus on appointment scheduling, only, and thus incorporate a severe thematic bias). Unlike previous efforts, we additionally supply demographic information associated with each e-mail author in order to lay the foundation for a deeper explanatory study of discourse characteristics of e-mails in the future. A summary of currently available e-mail corpora, including basic quantitative parameters, is provided in Table 1.

Corpus	E-Mails	Sentences	Languages(s)
Enron	620,000	—	English
Trec 2007	237,000	—	English
W3C	200,000	—	English
Trec 2006	99,000	—	English
Trec 2006	65,000	—	Mandarin
Australia	10,000	—	English
Tat	9,800	—	English
Tat	8,000	—	Arabic
Cosma	160	—	German
Flag	—	120,000	German
KingLine	—	74,300	Mandarin
KingLine	—	10,000	Mandarin

Table 1: Data Sheet for Currently Available E-Mail Corpora by Number of E-Mails, Sentences, and Languages Being Covered

[1] http://research.microsoft.com/enus/um/people/nickcr/w3c-summary.html

[2] http://www.w3c.org

[3] In the Anglo-American language world, only the *Australian National Corpus* (other than the *BNC* and the *ANC*) already includes e-mail-documents as a contemporary alternative to common written letters and memoranda.

[4] The 10,000 sentence collection (KING-NLP-L-001) can be purchased at http://kingline.speechocean.com/exchange.php?id=7209&act=view, the 74,000 set (KING-NLP-037) at http://kingline.speechocean.com/exchange_new.php?id=5662&act=view&cm=1

3. Corpus Design, Corpus Construction and First Results

In this section, we present the design principles underlying CODE ALLTAG, the hybrid data elicitation strategy we pursued and sketch first statistical results. The data acquisition procedure for CODE ALLTAG is best described as a hybrid pull-push process. As far as the design principles are concerned, we were interested, on the one hand, in getting a large corpus to study structural and linguistic patterns of e-mails. In order to maximize coverage, we created a *pull* process by drawing e-mails from a super-large, publically accessible e-mail archive. The CODE ALLTAG$_{XL}$ segment resulting from this approach comes with roughly 1.5M e-mails. On the other hand, exploring in depth stylistic and variational patterns in e-mails, we believe that more background knowledge about the individual authors of e-mails could be valuable. Hence, we created a *push* process where we asked our addressees not only to donate one of their e-mails, but also to provide additional author-specific demographic information. This is realized by the CODE ALLTAG$_{S+d}$ segment which comes with a much smaller size, almost one thousand e-mails. We conclude this section with first statistical corpus data, as well as an overview of aggregated demographic data of the donors of CODE ALLTAG$_{S+d}$.

3.1. Elicitation Process for Gathering E-Mails for CODE ALLTAG$_{XL}$

The most common way to assemble corpora from electronically mediated communication—chats, blogs, tweets, SMS messages, or e-mail—is to harvest digital archives or ad-hoc collections, such as tweets with special hashtags. Following the archive-based elicitation strategy of the FLAG corpus (Becker et al., 2003), we also made use of the German part of archived e-mails from various *Internet Usenet Newsgroup*s, accessible via `https://archive.org/details/usenet-de`. This is a huge collection comprising more than 35M e-mails from 841 thematic categories (roughly 78GB of storage, raw data only). Fortunately, e-mails stored in this archive are not only publically accessible but there is also no legal entity which is in possession of the IPRs for these e-mails. Hence, the data are open for scientific exploitation. We, finally, selected seven categories from the whole collection of 841 categories:

- `de.etc.finanz.boerse` (FINANCE) is a collection about financial issues, including stock exchange news,
- `de.etc.sprache.deutsch` (GERMAN) gathers topics related to the German language,
- `de.rec.film.misc` (MOVIES) contains information and discussions about movies,
- `de.rec.reisen.misc` (TRAVELS) deals with travel reports and tourism,
- `de.sci.philosophie` (PHILOSOPHY) covers philosophical issues,
- `de.talk.jugend` (TEENS) contains topics of interest for young people, such as teenagers,
- `de.talk.tagesgeschehen` (EVENTS) relates to events of the day.

In general, the archival material constitutes a 'dirty' document collection containing spam and fake mails on the one hand, and posing lots of character encoding problems (due to non-UTF-8-compliance), on the other hand. A notorious problem, e.g., relates to improperly representing the Germans umlauts (e.g., 'ä' vs. 'ae' vs. 'a' vs. some garbage sequence of symbols resulting from arbitrary mappings from the original encoding scheme to some unintended target code), not to mention an excessive level of erroneous language use (Becker et al., 2003).

By splitting the selected categories into single e-mails using MBOXSPLIT from the APACHE SPAMASSASSIN^{TM} project,[5] we gathered roughly 1.5M e-mails. The resulting documents' new line delimiters were unified via TOFRODOS and the first header was removed (we removed all line-up, too, including the first empty line) with the GNU stream editor SED[6]).[7]

3.2. Elicitation Process for Gathering E-Mails for CODE ALLTAG$_{S+d}$

One of the major hurdles in assembling and, later on, publically distributing e-mail corpora are privacy regulations related to the sender and recipient(s) of an e-mail. From a legal perspective, each sender of an e-mail is the owner of the text embodied in an e-mail. The ENRON corpus can clearly be considered an exception, since public accessibility was enforced by a U.S. court decision, whereas the privacy statuses of the e-mails collected for the W3C and the TREC corpora remain undecided. Privacy concerns related to material extracted from publicly accessible e-mail archives (as with FLAG or CODE ALLTAG$_{XL}$) are hard to substantiate since legal entities which are in possession of the IPRs of the e-mails can hardly be traced.

In order to avoid such legal issues, several e-mail corpus construction campaigns (*E-Mail Australia*, TAT, etc.), including ours for the second segment of CODE ALLTAG, as an alternative, rely on a proactive data elicitation strategy. They follow the push-based *donor model* where personal e-mails are requested from individual e-mail authors on the basis of a volunteer act and each donation can thus be considered as an act of free will to contribute to the corpus

[5] http://svn.apache.org/repos/asf/spamassassin/trunk/tools/mboxsplit

[6] https://www.gnu.org/software/sed/manual/sed.html

[7] We are fully aware that this crude procedure can be signifiantly improved by more sophisticated parsing of the mail body. The *Usenet Historical Collection* site (https://archive.org/details/usenethistorical) states that "This historical collection of Usenet spans more than 30 years and was given to us by a generous donor." A fully adequate parser would thus have to adhere to the plethora of RFCs starting with RFC822 (Standard for the Format for ARPA Internet Text Messages, 1982) and following all its successors to RFC6854 (2013) for basic e-mail, orthogonally tracking RFC1341 (Multipurpose Internet Mail Extensions MIME, 1992) and all its successors and updates, among others RFC6532 (Internationalized Email Headers, 2012). Regarding CODE ALLTAG$_{XL}$, we neither analyzed the time frame of the selected subset, nor did we map all the possibly relevant RFCs.

(still, the explicit consent from the donor has to be gained and was implemented in our questionnaire).

As far as CODE ALLTAG$_{S+d}$ is concerned, two elicitation rounds can be distinguished. We started our project at the University of Leipzig (Germany) and asked students in the introductory undergraduate courses for German linguistics to provide us with one of their personal e-mails, without any changes of the original e-mail body. We then spread our request over the university's e-mail system and so extended the range of addressees to all other institutes and administrative units of the University of Leipzig. In order to go beyond a purely academic audience, we asked everybody from our addressees to invite their families, friends, colleagues from sport clubs, etc. to contribute an e-mail, as well. The selection of a particular e-mail was not biased at all (although we assume that thematic considerations and private concerns played a decisive role for the final choice) and occurred at the free will of each donor. In particular, we did not articulate any content or formal preferences. In the corpus, we distinguish the original single donated e-mail from the thread it is (sometimes) linked to. In general, it is not our goal to collect threads (for an analysis of the thread structures of e-mails, cf. Sharaff and Nagwani (2016)), mainly because of the ownership issues discussed above—note that quoted (parts of) e-mails are owned by the original author, not by the quoter.

This 'Leipzig segment' is currently going to be complemented by the 'Klagenfurt segment'. At the Alpen-Adria-Universität Klagenfurt (Austria) a similar campaign as in Leipzig has been started and is still running at the time of this writing. Unlike Leipzig, much more public activities have been planned and executed by student staff (e.g., using flyers in public places, personally addressing people in the urban environment at cultural and sports events, involving the public media for advertising the project, etc.).

3.3. Pseudonymization of E-Mails

As already mentioned in the previous section, privacy preservation is a crucial issue for handling donated e-mails in a trustworthy manner. Especially, the donor model we have implemented for CODE ALLTAG$_{S+d}$ requires that privacy concerns have to be obeyed in a strict way to protect the right to informational self-determination of each individual author. This requirement, in principle, also holds for archival material from which CODE ALLTAG$_{XL}$ was derived.

A standard procedure to get around with this challenge is to either anonymize or pseudonymize raw linguistic data, preferably automatically (Medlock, 2006).[8] This problem resembles requirements imposed on the accessibility and subsequent processing of clinical reports in the medical domain (Meystre et al., 2010). The 2007 I2B2 de-identification challenge yielded promising results for automatic anonymization in the range of 98% F-score (Uzuner et al., 2007). As far as CODE ALLTAG is concerned, despite its undisputable relevance, we leave this topic for future research.

3.4. Descriptive Properties of CODE ALLTAG

In this section, we briefly introduce some descriptive properties of CODE ALLTAG. We start with an overview of basic statistics of CODE ALLTAG$_{XL}$ (cf. Table 2. A+B). The number of e-mails (E-MAIL) we identified in the seven categories varies around 200,000 documents per category and amounts to a total of roughly 1.5M. We further counted the number of sentences (SENTENCE) and of text tokens in all e-mails (TOKEN counts all running tokens, whereas TOKEN– is a filtered set, with stop words[9] and punctuation marks removed). We, furthermore, determined the number of different text tokens (TYPE–) with reference to TOKEN–, and normalized morphological variants in TYPE– which yielded all lemmata (LEMMA–). The reduction to TOKEN– typically halved the number of TOKENs, whereas the reduction rate from TOKEN– to TYPES– varies roughly between 95% to 97%.

For these basic computations, we used the sentence splitter and tokenizer from OPENNLP (with German models)[10] and a JAVA wrapper for the lemmatization function of the TREETAGGER software.[11]

From these basic parameters, we determined the average length of sentences (M_SENT) measured by the number of text tokens in a sentence (together with their standard deviation), as well as the average length of text tokens (M_TOKEN) measured by the number of characters (together with their standard deviation)—both including stop words and excluding punctuation marks. The number of tokens in sentences varies between 10 to 16 (with large standard deviations), whereas the average token length is quite stable (around 5 characters).

The latter two parameters focus on general corpus properties, whereas the following two parameters emphasize specific properties of e-mails. We distinguish here the average number of sentences per e-mail (M_SENT_E) from the average number of text tokens per e-mail (M_TOKEN_E), again including stop words and excluding punctuation marks. E-mails contain typically between 9 to 14 sentences (again, with large standard deviations), the PHILOSOPHY category being somewhat an outlier, with more than 20 sentences. The number of text tokens per e-mail typically ranges from 180 to 100. There are two deviant categories—on the low end, we find the TEENS category with 72 text tokens on average, whereas at the high end, the PHILOSOPHY category peaks (again) with more than 280 tokens. The standard deviations are quite large, so there is a lot of variance in the data.

Finally, two common lexico-statistical metrics are applied. Lexical diversity in CODE ALLTAG can be quantified by the type-token ratio

$$TTR := \frac{TYPE-}{TOKEN-} \tag{1}$$

[8] Efforts relying on self-anonymization by the donors in the *Australian National Corpus* were doomed to failure (Lampert, 2009).

[9] We used the German stop word list available from http://snowball.tartarus.org/algorithms/german/stop.txt which contains 232 entries.

[10] https://opennlp.apache.org/

[11] TREETAGGER is available from http://www.cis.uni-muenchen.de/∼schmid/tools/TreeTagger/.

Parameter	FINANCE$_{XL}$	GERMAN$_{XL}$	MOVIES$_{XL}$	TRAVELS$_{XL}$
E-MAIL	174,375	240,780	205,960	154,184
SENTENCE	2,462,152	2,359,301	2,098,171	1,401,026
TOKEN	36,349,694	30,235,557	34,533,372	21,150,739
TOKEN–	20,453,407	14,507,782	17,914,162	11,144,957
TYPE–	915,970	597,385	691,578	505,832
LEMMA–	730,627	520,619	610,428	446,392
M_SENT	12.56 (+/- 15.92)	10.64 (+/- 4.45)	16.35 (+/- 143.36)	15.88 (+/- 137.68)
M_TOKEN	5.68 (+/- 2.86)	5.47 (+/- 0.98)	5.35 (+/- 1.12)	5.55 (+/- 1.43)
M_SENT_E	14.12 (+/- 43.52)	9.8 (+/- 13.52)	10.19 (+/- 18.39)	9.09 (+/- 9.62)
M_TOKEN_E	178.03 (+/- 527.79)	103.3 (+/- 145.76)	141.81 (+/- 441.76)	116.85 (+/- 350.64)
TTR	0.045	0.041	0.039	0.045
LTR	0.036	0.036	0.034	0.040

Table 2: A: Breakdown of Basic Descriptive Statistical Data of Seven Segments for CODE ALLTAG$_{XL}$, standard deviation in rounded brackets

Parameter	PHILOSOPHY$_{XL}$	TEENS$_{XL}$	EVENTS$_{XL}$	**Σ: CODE ALLTAG$_{XL}$**
E-MAIL	209,558	239,008	246,556	1,470,421
SENTENCE	4,244,784	2,141,299	3,097,027	17,803,760
TOKEN	70,347,007	21,566,493	43,694,803	257,877,665
TOKEN–	31,407,208	10,370,899	20,963,113	126,761,528
TYPE–	918,106	316,511	699,116	3,097,518
LEMMA–	830,521	266,987	605,470	2,779,125
M_SENT	13.68 (+/- 45.10)	8.45 (+/- 4.16)	11.36 (+/- 9.00)	12.41 (+/- 72.20)
M_TOKEN	5.61 (+/- 1.06)	5.26 (+/- 0.66)	5.77 (+/- 1.97)	5.52 (+/- 1.56)
M_SENT_E	20.26 (+/- 31.67)	8.96 (+/- 8.30)	12.56 (+/- 16.17)	12.11 (+/- 22.88)
M_TOKEN_E	281.64 (+/- 479.51)	72.3 (+/- 108.95)	150.05 (+/- 222.51)	147.19 (+/- 351.79)
TTR	0.029	0.030	0.033	0.024
LTR	0.026	0.026	0.029	0.022

Table 2: B: Breakdown of Basic Descriptive Statistical Data of Seven Segments for CODE ALLTAG$_{XL}$, standard deviation in rounded brackets

and the lemma-token ratio

$$LTR := \frac{LEMMA-}{TOKEN-} \tag{2}$$

There is a division into two clusters for TTR. Four from seven categories (FINANCE, GERMAN, MOVIES and TRAVELS) vary around 0.04, whereas three from seven categories (PHILOSOPHY, TEENS and EVENTS) are below 0.03. Similar observations are made for the LTR values.

Table 3 renders statistical counts for CODE ALLTAG$_{S+d}$. This corpus is considerably smaller and less varied than CODE ALLTAG$_{XL}$. Altogether, we have 847 e-mails, excluding threaded e-mails (E-MAIL–), we end up with 297 items. Token, type and lemma counts are based on the sentence splitting, tokenization and stemming option from the TREETAGGER. Yet, it is almost impossible to reasonably compare both portions because of the low numbers in CODE ALLTAG$_{S+d}$. Still, M_SENT and M_TOKEN lie within the ranges we determined for CODE ALLTAG$_{XL}$. Not so surprisingly given the majority of our subjects, the M_SENT_E and M_TOKEN_E values are closest to those from the TEENS category in CODE ALLTAG$_{XL}$. TTR and LTR values in CODE ALLTAG$_{S+d}$ deviate markedly from those in CODE ALLTAG$_{XL}$ due to the limited size of the corpus. This is also reflected in the much smaller reduction rates from TOKEN– via TYPE– to LEMMA– (compared with CODE ALLTAG$_{XL}$).

Parameter	CODE ALLTAG$_{S+d}$
E-MAIL	847
E-MAIL–	297
SENTENCE	6,340
TOKEN	72,723
TOKEN–	39,280
TYPE–	12,078
LEMMA–	6,543
M_SENT	11,47 (+/- 3,86)
M_TOKEN	4.74 (+/- 3.75)
M_SENT_E	7.49 (+/- 6.19)
M_TOKEN_E	85.86 (+/- 81.39)
TTR	0.307
LTR	0.167

Table 3: Breakdown of Basic Descriptive Statistical Data for CODE ALLTAG$_{S+d}$, standard deviation in rounded brackets

3.5. Demographic Features of E-Mail Donors in CODE ALLTAG$_{S+d}$

When subjects submitted their e-mail donation to CODE ALLTAG$_{S+d}$, as a follow-up, a questionnaire was immediately sent to each donor asking for eight demographic variables and self-assessments of the donor, including gender, age, regional provenance, educational and professional background, language provenance and preferences, proficiency of writing and frequency of e-mailing. Finally, each donor was requested to explicitly grant permission to use the donated e-mail for scientific purposes (if the donor refrained from giving that permission, the donated e-mail was excluded from CODE ALLTAG$_{S+d}$).

In our study, the number of females clearly dominates the number of males by a factor of 4:1 (see Table 4). That table also reveals that CODE ALLTAG$_{S+d}$ donors are almost only native speakers of German, yet they speak predominantly German dialect(s) rather than high-standard German.

Gender	Male		Female		no answer	
	%	n	%	n	%	n
	21.2	63	77.1	229	1.7	5
German	**Native**		**Foreign**		**no answer**	
	%	n	%	n	%	n
	94,3	281	5,2	15	0,4	1
Standard	**German**		**Dialect**		**no answer**	
	%	n	%	n	%	n
	39.4	117	60.3	179	0,3	1

Table 4: Fundamental Demographic Data for the CODE ALLTAG$_{S+d}$ Corpus; the number of subjects equals E-MAIL– (= 297) from Table 3

The age distribution (see Figure 1) reveals that more than two thirds of our subjects are between 18 and 34 years old.

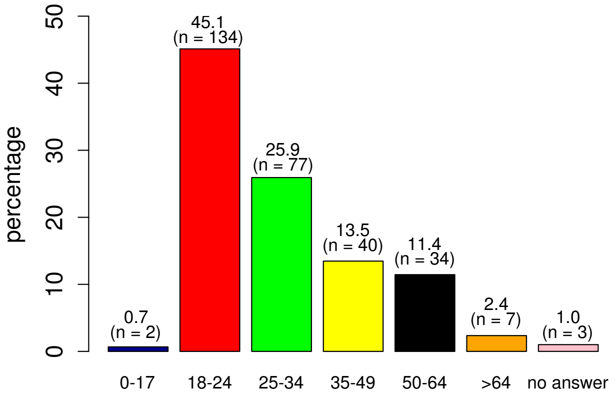

Figure 1: Age Distribution of E-Mail Authors in CODE ALLTAG$_{S+d}$

The vast majority of our subjects are currently either working (employed) or studying (students, pupils), at a level of almost 85% (see Figure 2). The level of work being performed is biased towards the top level segment (usually associated with academic degrees); almost 76% of our subjects are assigned to this group (see Figure 3).

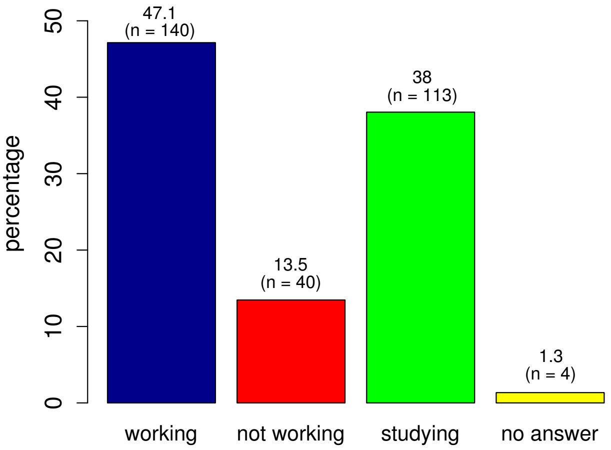

Figure 2: Distribution of Working Statuses of E-Mail Authors in CODE ALLTAG$_{S+d}$

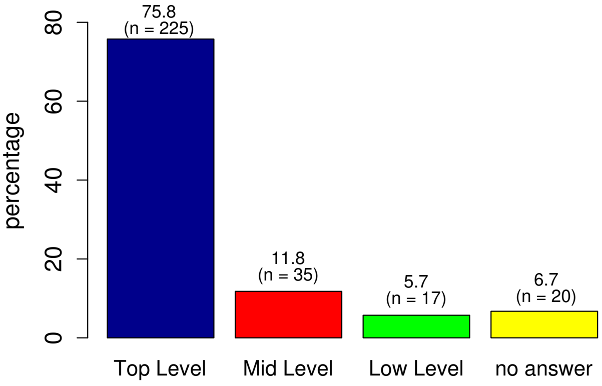

Figure 3: Distribution of Employment Levels of E-Mail Authors in CODE ALLTAG$_{S+d}$

As far as exposure to e-mail activities is concerned, roughly 93% of the subjects in our sample (see Figure 4) have a long-standing experience in using e-mail (more than 5 years, at least). Also the frequency of e-mail use is high: 50% are active daily e-mailers, 35% active on a weekly basis, at least (see Figure 5).

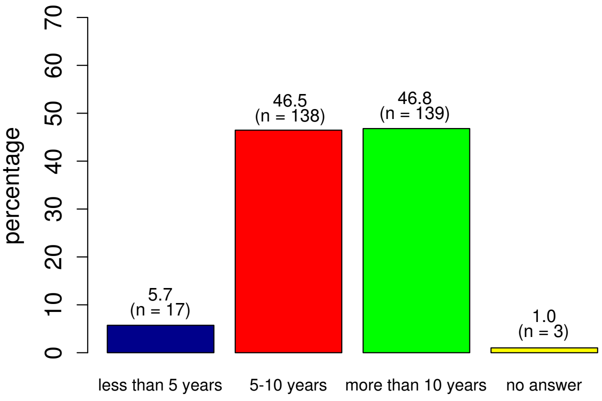

Figure 4: Distribution of E-Mailing Experience of E-Mail Authors in CODE ALLTAG$_{S+d}$

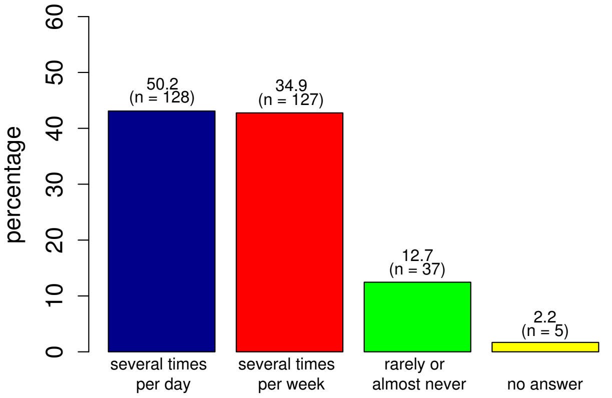

Figure 5: Distribution of E-Mailing Frequency of E-Mail Authors in CODE ALLTAG$_{S+d}$

4. Conclusions

We here presented the first overview of CODE ALLTAG, a text corpus of everyday German language use as embodied in private and semi-formal e-mails. CODE ALLTAG is a hybrid corpus composed of two different parts:

- CODE ALLTAG$_{XL}$ extracted from a publicly accessible e-mail archive is by far the largest portion, with approximately 1.5M e-mails. With these size numbers, CODE ALLTAG ranks on the top position among the largest e-mail corpora available worldwide.
- CODE ALLTAG$_{S+d}$, with less than 1,000 e-mails considerably smaller than CODE ALLTAG$_{XL}$, was newly generated from scratch on the basis of volunteer e-mail donations, together with demographic data from each e-mail donor.

Both elicitations campaigns are on-going. CODE ALLTAG$_{XL}$ should grow up to 35M e-mails, the total number of e-mails accessible in the German partition of the *Usenet* archive. However, these data are inherently 'dirty' and need a lot of curation prior to further processing and distribution. For the full version of CODE ALLTAG$_{S+d}$, we plan to elicit around 3,000 to 5,000 e-mails, plus demographic data of their donors.

Prior to public distribution, both partitions will undergo safe pseudonymization procedures. Once this is done, we hope to supply the community with a balanced corpus that should cover the language in e-mails in many different communication contexts, from (more or less) formal to informal language use. This balance seems currently not given in other corpus building and maintenance initiatives, for German, as well as for other natural languages.

Furthermore, in the future, we will extend CODE ALLTAG by syntactic metadata (such as POS tags, chunks, parsing structures, predicate argument structures), as well as semantic metadata (e.g., for named entity recognition, emotional language use, etc.).(Lawson et al., 2010) The former will lay the foundations for an even deeper comparison with structural and statistical properties of German-language newspaper (TIGER (Brants et al., 2004), TÜBA-D/Z (Telljohann et al., 2004)) and national reference corpora (DEREKO (Kupietz et al., 2010) and DWDS CORE (Geyken, 2007)).

In summary, we not only envisage a multitude of studies investigating the text genre specificity of e-mails, their comparison with other text corpora and their inherent sampling biases in order to create more balanced corpora. We also believe that the availability of complementary demographic data, such as those contained in CODE ALLTAG$_{S+d}$, will spur future investigations focusing on language style and different language varieties by exploiting heterogeneous and large corpora, such as CODE ALLTAG$_{XL}$.

5. Bibliographical References

Andersen, O., Nioche, J., Briscoe, E. J., and Carroll, J. A. (2008). The BNC parsed with RASP4UIMA. In *LREC 2008 — Proceedings of the 6th International Conference on Language Resources and Evaluation. Marrakech, Morocco, 26 May - June 1, 2008*, pages 865–869.

Aston, G. and Burnard, L. (1997). *The BNC Handbook. Exploring the British National Corpus with* SARA. Edinburgh University Press.

Atefeh, F. and Inkpen, D. Z. (2015). *Natural Language Processing for Social Media.* Morgan & Claypool.

Balahur, A. (2013). Sentiment analysis in social media texts. In *WASSA 2013 — Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis @ NAACL-HLT 2013. Atlanta, Georgia, USA, 14 June 2013*, pages 120–128.

Becker, M., Bredenkamp, A., Crysmann, B., and Klein, J. (2003). Annotation of error types for German newsgroup corpus. In Anne Abeillé, editor, *Treebanks. Building and Using Parsed Corpora*, pages 89–100. Springer.

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L., and Storrer, A. (2013). DERIK: A German reference corpus of computer-mediated communication. *Literary and Linguistic Computing*, 28(4):531–537.

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., and Pinkal, M. (2006). The SALSA corpus: A German corpus resource for lexical semantics. In *LREC 2006 – Proceedings of the 5th International Conference on Language Resources and Evaluation. Genoa, Italy, 22-28 May, 2006*, pages 969–974.

Cormack, G. V. (2006). TREC 2006 Spam Track overview. In *TREC 2006 – Proceedings of the 15th Text REtrieval Conference. Gaithersburg, Maryland, November 14-17, 2006*, pages 117–127.

Cormack, G. V. (2007). TREC 2007 Spam Track overview. In *TREC 2007 – Proceedings of the 16th Text REtrieval Conference. Gaithersburg, Maryland, November 5-9, 2007.*

Declerck, T. and Klein, J. (1997). Ein Email-Korpus zur Entwicklung und Evaluierung der Analysekomponente eines Terminvereinbarungssystems. In *Proceedings 6. Fachtagung der Sektion Computerlinguistik der Deutschen Gesellschaft für Sprachwissenschaft: Integrative Ansätze in der Computerlinguistik. DGfS-CL '97. Heidelberg, Germany, October 8-10, 1997.*

Estival, D., Gaustad, T., Pham, S., Radford, W., and Hutchinson, B. (2007). TAT: An author profiling tool with application to Arabic emails. In *Proceedings of the Australasian Language Technology Workshop 2007. Melbourne, Australia, 10-11 Dec. 2007*, pages 21–30.

Geyken, A. and Gloning, T. (2015). A living text archive of 15th-19th-century German: Corpus strategies, technology, organization. In *Historical Corpora: Challenges and Perspectives. Proceedings of the Conference. Frankfurt a. Main, Germany, December 6-8, 2012*, pages 165–180. Gunter Narr, Tübingen.

Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In C. Fellbaum, editor, *Idioms and Collocations*, pages 23–40. Continuum.

Ide, N. (2009). The American National Corpus: Then, now, and tomorrow. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages*, pages 108–113.

Iqbal, F., Fung, B., and Debbabi, M. (2012). Mining criminal networks from chat log. In *WI-IAT 2012 — Proceedings of the 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. Macau, China, Dec. 4-7, 2012*, pages 332–337.

Klimt, B. and Yang, Y. (2004). The ENRON corpus: A new dataset for email classification research. In *Machine Learning. Proceedings of the 15th European Conference on Machine Learning – ECML 2004. Pisa, Italy, September 20-24, 2004*, number 3201 in Lecture Notes in Computer Science, pages 217–226, Berlin. Springer.

Kupietz, M., Belica, C., Keibel, H., and Witt, A. (2010). The German reference corpus DEREKO: A primordial sample for linguistic research. In *LREC 2010 — Proceedings of the 7th International Conference on Language Resources and Evaluation. La Valletta, Malta, May 17-23, 2010*, pages 1848–1854.

Lampert, A. (2009). Email in the Australian National Corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages*, pages 55–60.

Lawson, N., Eustice, K., Perkowitz, M., and Yetisgen-Yildiz, M. (2010). Annotating large email datasets for named entity recognition with Mechanical Turk. In *Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk @ NAACL-HLT 2010. Los Angeles, CA, June 6, 2010*, pages 71–79.

MacLeod, N. and Grant, T. D. (2012). Whose tweet? Authorship analysis of micro-blogs and other short-form. In *Proceedings of the International Association of Forensic Linguists' 10th Biennial Conference. Aston University, Birmingham, July 2012*, pages 210–224.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The PENN TREEBANK. *Computational Linguistics*, 19(2):313–330.

Medlock, B. (2006). An introduction to NLP-based textual anonymisation. In *LREC 2006 — Proceedings of the 5th International Conference on Language Resources and Evaluation. Genoa, Italy, 22-28 May, 2006*, pages 1051–1056.

Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S., and Samore, M. H. (2010). Automatic de-identification of textual documents in the electronic health record: A review of recent research. *BMC Medical Research Methodology*, 10:#70.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREC 2010 — Proceedings of the 7th International Conference on Language Resources and Evaluation. La Valletta, Malta, May 17-23, 2010*, pages 1320–1326.

Palmer, M. S., Gildea, D., and Kingsbury, P. (2005). The PROPOSITION BANK: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Saurí R., Domingo, J., and Badia, T. (2014). The NEWSOME corpus: A unifying opinion annotation framework across genres and in multiple languages. In *LREC 2014 — Proceedings of the 9th International Conference on Language Resources and Evaluation. Reykjavik, Iceland, May 26-31, 2014*, pages 2229–2236.

Scheffler, T. (2014). A German Twitter snapshot. In *LREC 2014 — Proceedings of the 9th International Conference on Language Resources and Evaluation. Reykjavik, Iceland, May 26-31, 2014*, pages 2284–2289.

Schwartz, H. A., Eichstaedt, J. C., Dziurzynski, L., Kern, M. L., Seligman, M. E. P., Ungar, L. H., Blanco, E., Kosinski, M., and Stillwell, D. (2013). Toward personality insights from language exploration in social media. In *Analyzing Microtext: Papers from the 2013 AAAI Spring Symposium*, pages 72–79. AAAI Press.

Sharaff, A. and Nagwani, N. K. (2016). Email thread identification using latent Dirichlet allocation and non-negative matrix factorization based clustering techniques. *Journal of Information Science*, 42(2):200–212.

Sokolova, M. and Lapalme, G. (2011). Learning opinions in user-generated Web content. *Natural Language Engineering*, 17(4):541–567.

Song, Z., Strassel, S., Lee, H., Walker, K., Wright, J., Garland, J., Fore, D., Gainor, B., Cabe, P., Thomas, T., Callahan, B., and Sawyer, A. (2014). Collecting natural SMS and chat conversations in multiple languages: The BOLT phase 2 corpus. In *LREC 2014 — Proceedings of the 9th International Conference on Language Resources and Evaluation. Reykjavik, Iceland, May 26-31, 2014*, pages 1699–1704.

Telljohann, H., Hinrichs, E. W., and Kübler, S. (2004). The TÜBA-D/Z TREEBANK: Annotating German with a context-free backbone. In *LREC 2004 — Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon, Portugal, 24-30 May, 2004*, pages 2229–2235.

Uthus, D. C. and Aha, D. W. (2013). The UBUNTU chat corpus for multi-participant chat analysis. In *Analyzing Microtext: Papers from the 2013 AAAI Spring Symposium*, pages 99–102. AAAI Press.

Uzuner, O., Luo, Y., and Szolovits, P. (2007). Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.