Journal of the Linguistic Society of Papua New Guinea
ISSN 0023-1959

Special Issue 2017

Péter Maitz & Craig A. Volker (eds.):

Language Contact in the German Colonies: Papua New Guinea and beyond

DOCUMENTING UNSERDEUTSCH (RABAUL CREOLE GERMAN): A WORKSHOP REPORT

Angelika Götze & Siegwalt Lindenfelser & Salome Lipfert & Katharina Neumeier & Werner König & Péter Maitz
University of Augsburg
peter.maitz@philhist.uni-augsburg.de

ABSTRACT

This paper provides insights into the ongoing international research project *Unserdeutsch (Rabaul Creole German): Documentation of a highly endangered creole language in Papua New Guinea,* based at the University of Augsburg, Germany. It elaborates on the different stages of the project, ranging from fieldwork to corpus development, thereby outlining the methods and software background used for the intended purposes. In doing so, we also give some approaches to solving specific problems, which have arisen in the course of practical work until now.[1]

KEYWORDS

Unserdeutsch, corpus development, creole, fieldwork, German-based, language endangerment, language documentation, spoken language data, transcription

[1] Research underlying this article has been funded by the *Deutsche Forschungsgemeinschaft*/German Research Foundation (MA 6769/1-1). We are very grateful to Eva Schenzinger, Susanne Klohn and Craig Volker for their assistance in the translation of the article and we wish to thank Thomas Schmidt for his helpful advice. Of course, any remaining shortcomings are entirely ours.

1 INTRODUCTION

At least half of the world's languages can be considered endangered, i.e. facing the brink of language dormancy or extinction (cf. Thomason 2015: 2). For linguists, the impending loss of language variety means a race against time in saving valuable data. The realization of this risky situation gave rise to the linguistic subfield of language documentation in the 1990s (cf. Austin 2014: 58), which is "concerned with the methods, tools, and theoretical underpinnings for compiling a representative and lasting multipurpose record of a natural language or one of its varieties" (Gippert & Himmelmann & Mosel 2006: v).

Efforts in language documentation, above all, "strengthen the empirical foundations of those branches of linguistics and related disciplines which heavily draw on data of little-known speech communities (e.g., linguistic typology, cognitive anthropology, etc.) in that they significantly improve accountability (verifiability) and economizing research resources" (Himmelmann 2006: 1). Whalen (2004) even sees an upcoming revolution looming on the linguistic horizon, initiated by the study of endangered languages based upon large amounts of language data recently being shared on a global scale.

Language documentation may also support work in related disciplines, such as oral history and anthropology. Moreover, it constitutes "a necessary first step toward language maintenance and revitalization or as a safeguard against complete language loss" (cf. Austin & Grenoble 2007: 17). Doing research not only *on,* but likewise *for* and *with* a speech community is the basic idea (cf. Dwyer 2006: 32).

Unserdeutsch (Rabaul Creole German), the creole language this paper deals with, can be classified as "severely" to "critically" endangered (cf. Maitz & Volker 2017) according to the UNESCO *Language Vitality Index* (cf. UNESCO 2003). Unserdeutsch is therefore an urgent case for language documentation. This applies even more as Unserdeutsch has several characteristics that make it stand out in comparison to other languages.

2 THE UNSERDEUTSCH LANGUAGE

Unserdeutsch is the only known German-based creole language in the world. It emerged among mixed-race children at a missionary station of the Catholic Sacred Heart Missionaries (MSC) in Vunapope at the beginning of the 20th century. Vunapope, today part of Kokopo, is located in the northeast of the island New Britain (Gazelle Peninsula) in the Bismarck Archipelago. For generations, Unserdeutsch has been a historical connection between Papua New Guinea and Germany, even with the vast majority of the speakers living in Australia today. Tok Pisin, as the most important substrate language, shaped especially the grammar and phonology of Unserdeutsch (for aspects of the linguistic structure of Unserdeutsch refer to Lindenfelser & Maitz, this issue).

Despite its linguistically exceptional position, the language has hardly been recognized by linguists since its first documentation at the end of the 1970s. The only more or less broad description of at least the most important basics of the language is an unpublished master's thesis from the early 1980s (cf. Volker 1982). More than 35 years later, Unserdeutsch is now at the edge of its extinction: Hardly any fluent speaker is younger than 65 years, and the language has not been passed on to the following generations for decades.

It is urgent to document the language and to do research now. The ongoing decline of the number of speakers, and the increasing attrition show that the data collection cannot have been delayed. Fortunately, The German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) – the central funding agency for scientific research projects in Germany – granted a considerable third-party project in order to document Unserdeutsch (cf. Maitz et al. 2016). Maitz & Volker (2017) show the status quo of the Unserdeutsch research, and first new findings concerning language structure and language endangerment. This workshop report reflects the project status as of spring 2017.

3 PROJECT OVERVIEW

The goal of the research project under the leadership of Péter Maitz and Werner König (both at the University of Augsburg) is the comprehensive

documentation of Unserdeutsch. The funding period for the project covers three years (October 2015 to September 2018). The close cooperation with Craig A. Volker (The Cairns Institute of the James Cook University), the Institute for the German Language (*Institut für Deutsche Sprache*, IDS), the main non-university institution in Germany for studying and documenting the German language, and the advisory support of Peter Mühlhäusler, a leading expert in the field of Pacific and creole linguistics, contribute to the project with additional expertise.

The documentation of Unserdeutsch can be divided into three main stages. The first project stage, already far advanced, covers data collection by carrying out fieldwork in the Pacific. This was followed by the current second stage, which is the development of an annotated corpus of Unserdeutsch. In the third and last stage, a systematic description of the language based on the corpus will be carried out. For this description, an application for the funding of a subsequent project will be made, as the current project covers mainly the development of the corpus. Of course, the three project stages cannot be hermetically separated from each other but overlap one another. A flexible approach has seemed reasonable: While developing the corpus (stage 2), one can keep collecting new data if new speakers come to be known or if deficiencies, e.g., in terms of the balancing of the corpus (such as a balanced consideration of different varieties in the post-creole continuum) become visible. Furthermore, qualitatively poorer recordings can be replaced by better ones. The language description (stage 3) can also be started gradually during stage 2 (cf. Lindenfelser & Maitz, this issue). The basic linguistic structures have already become apparent while dealing with the data in the course of the corpus development. Even within the current project stage, e.g., in the stage of the corpus development, it is possible to start working on one of the following stages early. The advantage here is that problems that arise can be recognized at an early stage.

Even now, in the first third of the funding period, some first positive results and effects of the Unserdeutsch project can be stated:

(a) Through the project, Unserdeutsch has attracted public interest and generated great interest beyond our expectations, not only within the scientific community but also in the public. Numerous print and audio-visual media in the German-speaking countries, Australia, Papua New

Guinea, and beyond have been reporting about the language, the speech community, and the project since it was launched. This is a factor that is of great importance in giving prestige to the project, and at the same time a good basis for efforts to officially accept Unserdeutsch as an endangered language and to take revitalising measures.

(b) The speakers' attitudes towards their language has changed positively. Until now, they have thought that Unserdeutsch was, in contrast to *hohe Deutsch* (Standard German), a deficient L-variety. Therefore, at the beginning of the project some persons did not want to speak Unserdeutsch with a competent Standard German speaker being present (linguistic shame). Through the interest in research and documentation of their language, the speakers have become aware of their language, its uniqueness, and the cultural heritage it transports. The desire has arisen to preserve Unserdeutsch for their descendants and to maintain the language in their community.

(c) Last but not least, the social relationships within the group have been re-intensified through the project. The Unserdeutsch community was close-knit until the post-war period, as all speakers were living in and around the Vunapope Mission. With the emigration of most of the speakers as a consequence of the independence of Papua New Guinea in 1975, social contacts and with them community ties were more and more weakened by geographical dispersion. Now, the group identity is being revived, strengthened by its own closed Facebook group, where memories and news can be shared and discussed.

4 DATA COLLECTION, FIELDWORK

Unserdeutsch is spoken by only about 100 elderly people today. There are only a few fluent or semi-fluent speakers below the age of 65. Most of the speakers' families who lived in and around Vunapope have emigrated to Australia as a result of the independence of Papua New Guinea, and live today in the metropolitan areas along the east coast of Australia in and around Brisbane, the Gold Coast, Cairns and Sydney. Only a small number remained in Papua New Guinea, now scattered on different islands of the country. The main distribution area of the language has thus shifted from Papua New

Guinea to eastern Australia since 1975. For this reason, fieldwork is taking place mainly in Australia. Only a small handful of speakers have been interviewed in Papua New Guinea, among those living in Kokopo, Kavieng and the Duke of York Islands.

A large part of the data could be collected from 2014 to 2017 during five fieldwork trips. We started with the transcription of the first recordings made during the first fieldwork trips in spring 2016. At this stage, there is only one further fieldwork trip scheduled for autumn 2017. The data collection is done by means of a method triangulation, with a partly controlled narrative interview at its core. In order to reduce the relative unnaturalness of the survey situation and to improve the authenticity of the data, two peer group members (friends or relatives) are interviewed at the same time, so that a conversation can develop during the interview. The interviewer's questions mainly relate to the linguistic and social circumstances at the mission in Vunapope, to the past and the present life of the speakers, and to their families and the speech community as a whole. By this thematic focus, valuable metalinguistic information relevant for the linguistic interpretation of the primary data can be gained. The interview is supplemented by a questionnaire with about 320 stimuli (sentences and phrases) in English and Tok Pisin, the principal languages of Papua New Guinea spoken by all Unserdeutsch speakers. The language of the stimuli is determined or selected by the interviewees depending on their preference and competence. These stimuli are reproduced orally by the interviewees in Unserdeutsch and recorded. The stimuli are designed in a way that the basic vocabulary and the most important morphological and syntactic variables are elicited. At the end, the interviewees fill out a questionnaire to provide metalinguistic data concerning their language biography, their self-identity, and their language attitudes. This information is essential for both the reconstruction of the language history of the community, which is to be reconstructed systematically as part of a doctoral dissertation project, as well as for the appropriate linguistic interpretation of the primary linguistic data.

Approximately 50 hours of spontaneous speech and the according transcripts are to be integrated in the corpus. Basilectal, mesolectal and acrolectal varieties of Unserdeutsch will be represented in these 50 hours so that, if possible, the whole variation along the post-creole continuum is displayed. When collecting data, one is inevitably confronted with numerous

methodological problems. The first and probably biggest one is the observer's paradox (Labov 1972): We try to record natural, spontaneous speech in an investigative situation, which is unnatural when produced only by the presence of an outsider investigator. Unserdeutsch is, and always was, the medium of informal oral in-group communication within the Vunapope mixed-race community. Therefore, the language does not show any considerable stylistic variation. However, the observer's paradox is a severe methodological challenge, especially regarding mesolectal and acrolectal speakers, who control a wider range of the creole continuum, and, at the same time, have a greater language awareness.

A second methodological difficulty arises from standard language ideology being present within the speech community and especially among acrolectal speakers:

(1) *immä* *wi* *geht* *spiel-en* *mit* *alle* *kind-ä* *fi* *die,*
every_time 1PL go play-V with PL child-PL of 3PL

die *sa:* *in* *mein* *haus* *du* *spreh-en* *deutsch* *odä*
3PL say in 1SG.POSS haus 2SG speak-V German or

englisch, *kein* *kaputt-e* *deutsch*
English no broken-ATTR German

'Every time we went playing with their children, they [their parents] said: In my house, you speak German or English, no broken German.'

This leads the speakers to classify Standard German as *Proper German* and Unserdeutsch as its corrupted form. This becomes apparent inter alia in emic language designations for Unserdeutsch, such as *Falsche Deutsch* ("wrong German") or *Kaputtene Deutsch* ("broken German"). This standard language ideology leads to the problem that especially acrolectal and mesolectal speakers, who already show a bigger language awareness, might consciously avoid basilectal features in formal out-group communication, including the interview situations.

These difficulties have been dealt with in four different ways during fieldwork. Firstly, we try to establish a trustful relationship with the speakers through informal personal meetings (e.g., gatherings, lunch, dinner etc.) before the interviews. Secondly, most of the interviews take place at the

speakers' homes, i.e. in a familiar environment. Thirdly, the interviews are, as already mentioned, conducted with two familiar peer-group members. All these measures are supposed to reduce the unnaturalness of the recording situation. Last but not least, the interviewers speak (as much as they can) Unserdeutsch during the interviews. This is a necessary strategy as Standard German is only partly intelligible for the speakers of Unserdeutsch. Moreover, the use of the in-group language as a contextualization cue should contribute to dissolve the group boundary between interviewer and interviewee (Gumperz 1982). In addition, the avoidance of Standard German as an interview language is supposed to prevent echo effects, and, hence, the evocation of unauthentic acrolectal features.

At the beginning of the fieldwork conducted for the Augsburg Unserdeutsch Project in autumn 2014, Unserdeutsch was hardly documented. A large part of the recordings made by Volker at the end of the 1970s and the beginning of the 1980s had been lost; one copy is missing due to the carelessness of the university library in Australia where it had been deposited, another copy was destroyed by the great volcanic eruption in Rabaul in 1994. Only a small rest containing about six hours of bad or very bad quality has survived.[2] Almost half of this material, however, is actually Standard German, because the older generation interviewed at that time was also able to speak Standard German fluently thanks to school lessons partly held in German in Vunapope during the interwar period.

Since the beginning of the project, 62 hours of recordings with 52 speakers in total have been made from 2014 to 2017; 47 hours of these are narrative interviews and about 15 hours are questionnaires. The latter will not be integrated into the corpus but will be essential when describing the language structure after the language documentation.

The data show considerable variation in numerous respects. The first concern the languages used. Unserdeutsch, Tok Pisin and English play an important role in the language biography of all speakers. This trilingual competence shows varying symmetry or asymmetry depending on the speakers' individual language biographies.

[2] The remaining data is now stored at the Archive of Spoken German (*Archiv für Gesprochenes Deutsch*, AGD) in Mannheim; cf. the corpus German in Oceania (*Deutsch in Ozeanien*, OZ): http://agd.ids-mannheim.de/korpus_index.shtml (13.11.2017).

This is why speakers shift or switch from Unserdeutsch to English or Tok Pisin to a different extent and often vary between two or three of these languages during the interviews:

(2)

orait	*i*	*wid*	*ni*	*resign*	*i*	*wid*	*bleib*
TP	UD	UD	UD	EN	UD	UD	UD
all_right	1SG	AUX.FUT	NEG	resign	1SG	AUX.FUT	stay

arbeit	*weiter […]*	*aba*	*i*	*own-im*	*de*	*trade*	*store*
UD	UD	UD	UD	EN-TP	UD	EN	EN
work	further	but	1SG	own-TR	ART.DEF	trade	store

'All right, I will not resign; I will stay and work further on […], but I (will) own the trade store.'

Secondly, the speakers use lexically and grammatically differently elaborated basilectal, mesolectal or acrolectal varieties of Unserdeutsch. Thirdly, the recordings differ in terms of the extent and kind of attrition phenomena: lexical gaps, problems with word finding, code-switching, and phonological, grammatical and/or lexical interference of English and/or Tok Pisin – languages that have been used as functional first languages for decades.

5 DEVELOPMENT OF THE UNSERDEUTSCH CORPUS

The development of the Unserdeutsch corpus can be divided into five phases: 1. Transcription, 2. Normalization, 3. Lemmatization, 4. Annotation, and 5. Database implementation. The most time-consuming part is the present phase of transcription.

Phase 1 – Transcription

The transcription of the audio recordings is conducted with the Partitur Editor of the software EXMARaLDA (Extensible Markup Language for Discourse Analysis). EXMARaLDA (cf. www.exmaralda.org) is a free tool for computer-assisted speech transcription and for the administration of spoken language corpora (cf. Schmidt & Wörner 2014). It is compatible with all common operating systems and offers interfaces to other common transcription tools and formats (e.g., ELAN, Praat, TEI, Transcriber) as well

as to standard applications such as Microsoft Word, internet browsers and text editors via import and export functions. The transcription appears time-aligned to the audio file in musical score (*Partitur*) notation:

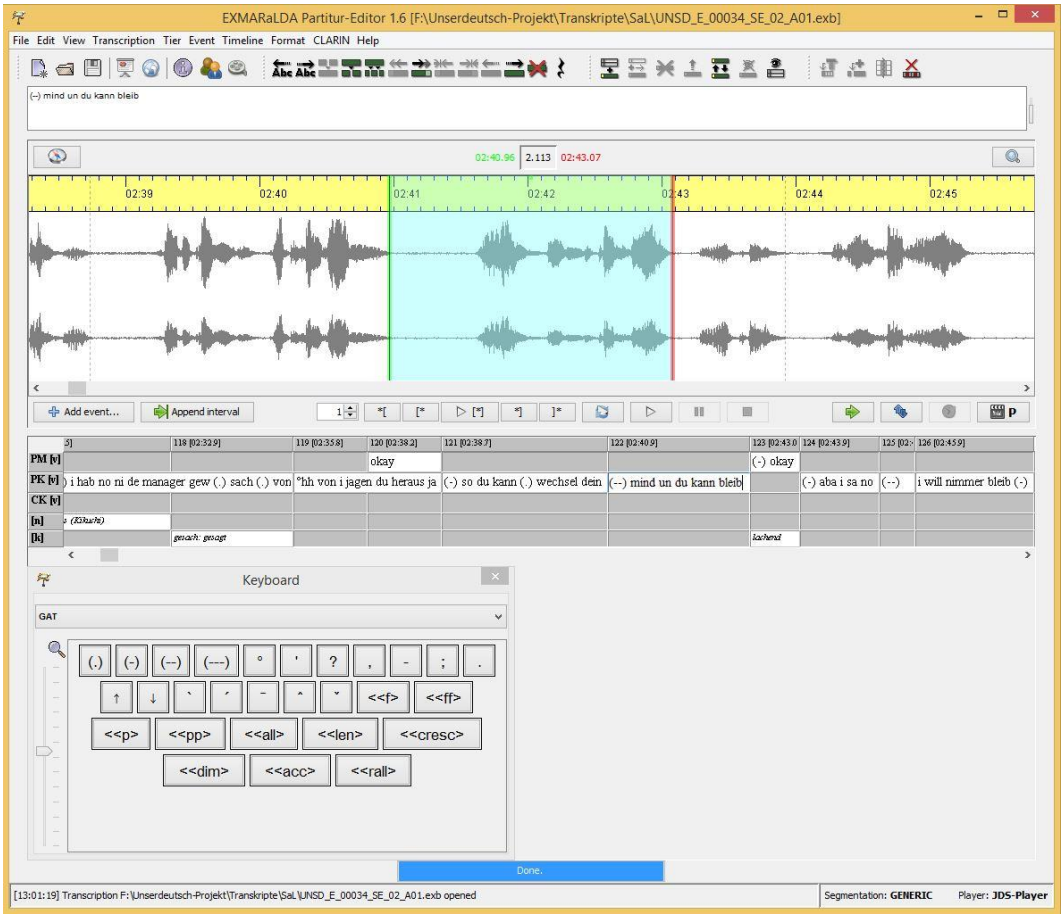

Figure 1. The work surface of the EXMARaLDA Partitur Editor

One transcription tier is assigned to every speaker, further description tiers and annotation tiers can be added optionally. The possibility to play intervals in a modified speed has been helpful. Work is done with two different files that are linked to one another: the recording itself and the transcription file. EXMARaLDA has dedicated support for a number of widely used transcription systems such as GAT 2 (*Gesprächsanalytisches Transkriptions-system* 2, cf. Selting et al. 2009) or HIAT (*Halbinterpretative Arbeits-transkriptionen*, cf. Ehlich & Rehbein 1976) that are commonly used in German-speaking countries and beyond. For the purpose of the Unserdeutsch project, the system cGAT (cf. Schmidt et al. 2015) was taken as a basis, which was developed in the style of GAT 2 minimum transcript. It presumes the use

of standard orthography when a deviation is not significant from the standard phonology. Expressions that show significant deviation are noted in a so-called "literary transcription", i.e. phonetically using the Latin alphabet (a modified standard orthography). It is based on grapheme-phoneme correspondences; the literary transcription is therefore easily applicable and readable. It makes sure that structurally relevant features of Unserdeutsch – but not every small phonetic alternation that are irrelevant for grammatical description purposes – are visible in the transcript and searchable. This concerns phonetic phenomena such as the g-spirantization (Standard German *Berg* 'mountain' transcribed as <berch> according to the actual pronunciation influenced by northern German varieties), the delabialization of labial vowels (Standard German *Frühstück* 'breakfast' transcribed as <frihstick>) or the loss of final consonants (Standard German *Nachmittag* 'afternoon' transcribed as <namitta>). Sequences or words in English and Tok Pisin are reproduced according to the standard orthography of the respective language, since these languages are not in the focus of the Unserdeutsch project. For the notation of Unserdeutsch, some additional conventions were established in the interest of a more detailed possibility for analysis. In planning the time needed for transcription, a ratio of 1:60 has turned out to be realistic. For one minute of recordings, 60 minutes of transcription are therefore required (including correction steps). This is an average value, which varies depending on the comprehensibility of the recordings, the number of the people speaking, and the languages used. When choosing recordings for the initial phase of the transcriptions, we consciously started with a range of different speakers who represent various varieties along the creole continuum. By this means, we can start at an early stage to do first language analyses on a semi-representative basis and to describe the variation along the creole continuum of Unserdeutsch. Furthermore, it has become apparent that it makes sense to specialize the transcribers in terms of the recordings: The necessary period of acclimatization with a new recording decreases when individual speakers or the speech variety (basilectal vs. acrolectal speakers) are familiar to the person transcribing. The same is true for the distribution of the recordings with extensive code-switching to English or especially to Tok Pisin. Therefore, not everyone transcribing has to become acquainted with Tok Pisin equally well. By this means, every transcriber becomes an expert for a special part of the transcription and loses less time when dealing with

problems. As soon as the planned amount of data is completely transcribed and reviewed, shorter relevant passages will be additionally transcribed phonetically using the IPA. This is crucial for a detailed phonological description of Unserdeutsch. EXMARaLDA supports phonetic transcription, but we intend to use the software Praat, which was created especially for that purpose. Praat offers detailed measuring, e.g., in terms of vowel quantity which is especially relevant for the analysis of Unserdeutsch, as the data show different vowel lengths, but the phonological distinctiveness of the vowel quantity seems very doubtful at first sight, at least in the basilect (cf. Maitz & Volker forthc.).

Phase 2 – Normalization

The use of literary transcription means a high error ratio for automatic part-of-speech tagging (cf. Westpfahl & Schmidt 2013: 140), because there are no lexicon entries for spoken forms. It thus becomes necessary to normalize the transcribed forms before doing further steps of annotation. That is, every transcribed form is assigned to its equivalent in standard orthography which, at a later stage, offers expanded possibilities for searching the corpus. In order to speed up this manually time-consuming process, the tool OrthoNormal was developed at the Institute for German Language (IDS) in Mannheim (cf. Schmidt 2012: 239–240).

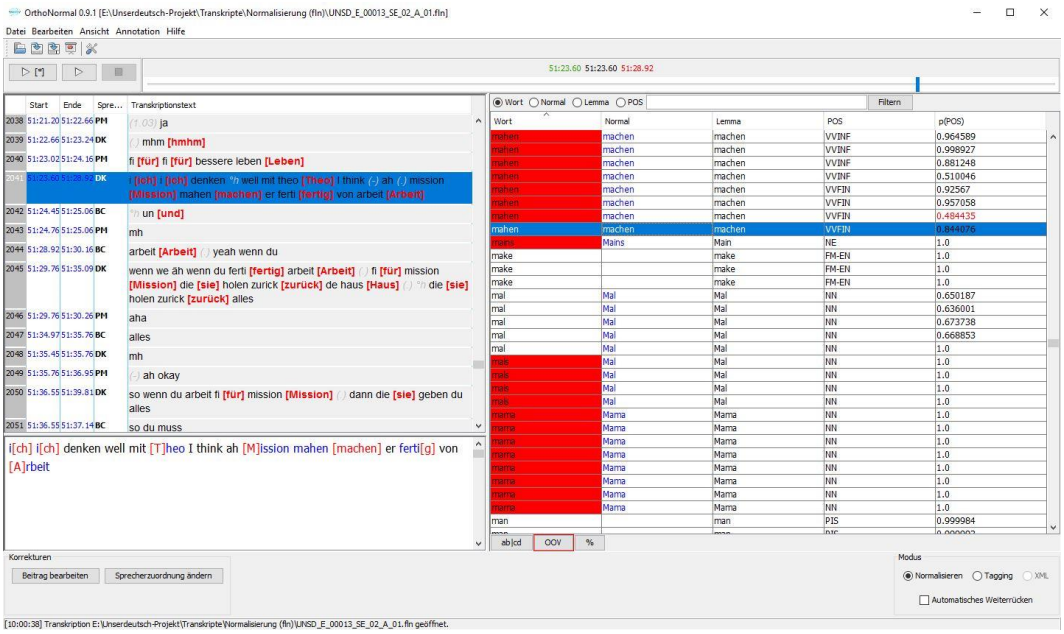

Figure 2. The work surface of OrthoNormal in normalization mode

The normalization with this tool is semi-automatic, as it suggests forms by itself, and identical forms can be assigned to their standard form in only one step in a whole transcript. Moreover, the program proposes equivalents on the basis of an integrated lexicon, which keeps growing by the assignment of further forms. By this means, fewer and fewer forms have to be typed in manually. The word suggestions are sorted by frequency of use as well. However, until now only a Standard German lexicon has been implemented: The lexicon for the substrate language Tok Pisin and the adstrat language English would have to be built as part of the normalization of the Unserdeutsch data in the first place. In this way a foundation for the use of the tool in English-speaking areas and in the Pacific could be laid. But since we only used standard spellings for Tok Pisin and English in the transcriptions, we can get around this task: The forms no longer need to be normalized, since there are no deviating spelling forms in the transcripts.

Phase 3 – Lemmatization

The normalized forms are the basis for an error-free automatic lemmatization. This happens by the already established tool TreeTagger. An interface for the TreeTagger functions has been integrated into OrthoNormal, so the lemmatization can easily be done there, too. The assignment of tokens to their lemmas provides further possibilities for an effective search of the corpus at a later stage.

Phase 4 – Annotation

In order to guarantee the searchability of the corpus by morphosyntactic categories, independent from individual word forms, we do a part-of-speech tagging of the data. This also happens automatized with TreeTagger with its interface in OrthoNormal. The Stuttgart-Tübingen-Tagset will be applied for this purpose, which is already an established standard for the German language (STTS, cf. Schiller et al. 1999). To be more precise, we use the expanded STTS 2.0, which has been designed for the annotation of spoken language in particular, optimizing the tagset for that purpose by introducing slight modifications and by adding further categories of spoken language (cf. Westpfahl 2014; Westpfahl et al. 2017).

Furthermore, a manual annotation of transfer phenomena on the lexical level will be performed, which can also be done with the tool OrthoNormal.

By this means, the occurrence of Tok Pisin and English lexemes will be marked. This will later facilitate research, for instance on the integration of lexical material from English or Tok Pisin into Unserdeutsch, e.g., when creating a dictionary.

Transcription	*i*	*hat*	*ein*	*sch (.) schtore*
Meaning	‘I‘	‘have‘	‘a‘	‘store‘
Normalization	[ich]	[habe]	[einen]	[Store]
Lemmatization	[ich]	[haben]	[ein]	[Store]
POS-Tagging	PPER	VVFIN	ART	NN

Figure 3. Normalization, lemmatization and POS-Tagging of transcribed word forms

Phase 5 – Database implementation

After completion and a period in which access will be restricted to selected direct collaborators of the project, the Unserdeutsch corpus will be made available for international research and teaching (i.e. non-commercial use) via the Database for Spoken German (URL: http://dgd.ids-mannheim.de). The DGD as a corpus management system digitally provides a part of the Archive of Spoken German (*Archiv für Gesprochenes Deutsch*, AGD) that is part of the IDS in Mannheim. The access is web-based and free of charge after a one-time registration. The database provides the transcripts aligned with the recordings so that they can be read and listened to in the browser. Furthermore, there is additional material and extensive metadata. The core of the database is an elaborated search function, which allows complex search queries within the metadata, the transcripts, and their annotation layers (cf. Schmidt 2014: 1453–1454). Here, different variables and wildcards can be used. The structure-sensitive token search provides the possibility to search on different annotation levels, including transcribed, normalized and lemmatized forms as well as POS tags. The context-sensitive search displays co-occurrences in a KWIC-view. An individual working area provides the possibility to compile one’s own virtual corpora. They can even be downloaded on request for offline work.

6 PRACTICAL ISSUES 1: ETHICO-JURISTIC ASPECTS

Particular project specific issues have arisen in this practical fieldwork-based data collection. One of them refers to the ethical and legal dimension of the project in terms of data protection (for general ethical aspects in language documentation refer to Dwyer 2006). The collection, storage, processing and publication of personal data are subject to the consent of the concerned people according to German law as well as to the regulation of data protection of the European Union guaranteeing the right of informational self-determination. This obviously includes the building of a corpus (cf. DFG 2013). A mutual agreement must therefore be established between researchers and informants. This contains data protection statements on the researchers' part and the written consent of the interviewed speakers, as was done in the Unserdeutsch project.

Normally, this must be guaranteed by anonymization and pseudonymization so that the identities of the speakers are, if at all, only possible with tremendous effort. Unserdeutsch is however a special case: The speakers care for the documentation of their biographic stories and their language, especially for their own descendants. Therefore, there is no need for them to conceal their identities. The reference persons explicitly agreed to have their names and pictures published in connection with the Unserdeutsch data on the homepage of the IDS and on the university project website. Therefore, the obligation for the anonymization and pseudo-nymization of transcripts and metadata, as well as the necessity to fade out or to blend the recordings at passages that are sensitive for data security is not applicable. Nonetheless, during transcription all occurring names are assigned to a separate tier used solely for names. This gives the possibility to fade out this tier or to blend it at a later stage, thus granting complete anonymization. This could be necessary if the declaration of consent in this respect was withdrawn, or if the data were to be passed on for (non-commercial) purposes that do not concern the project – the latter in consultation with the speakers of course.

Apart from such data that allow the identification of the informants (personal names, precise names of location), the handling of other personal data has to be considered carefully, as well, even more so, as, because of the arrangements just described, the full identification of the speakers is

achievable without any effort. According to the German laws, “information on racial or ethnic origin, political opinions, religious or philosophical convictions, […] health, sex life” (Federal Data Protection Act, Section 35, 2) are included. First, as a matter of principle, information about the “racial or ethnic origin” of the Unserdeutsch speakers is crucial for research questions on the social context of Unserdeutsch, its functions and its emergence. Apart from that, the speakers talk openly about their lives during the interviews. Thus, further sensitive information occurs in the recordings, not only about the speakers themselves but also about persons related to them.

As the loss of essential context information has to be avoided, this ethical and legal issue has to be dealt with before the corpus goes online. If the obliteration of sensitive passages is inevitable, this has to be conducted manually. In case of doubt, the participants must be asked – or their descendants, in case of speakers having already passed away. Besides the speakers themselves and their personal environment, sensitive content might involve other parties as well, such as the MSC mission and its former staff. Some interviews may even contain politically delicate information, for example, when speakers talk about war crimes during the Japanese occupation of New Britain in the course of World War II.

7 PRACTICAL ISSUES 2: TRANSCRIPTION PROBLEMS

Besides the common challenges emerging when transcribing spoken language, some project-specific complications have arisen. Here four of these shall be described.

(1) Dealing with homography

Some of the speakers switch between Unserdeutsch, English, and Tok Pisin to a considerable extent, the latter depending on the speaker and generally occurring more sporadically. As the system of GAT 2 minimum transcript does not include the use of capital letters, capitalization is not an option to differentiate between German and non-German substantives. As a result, German-English homographs, such as <mission> for English [ˈmɪʃ(ə)n] and likewise Unserdeutsch [miˈsion], or <plan> for English [plæn] and likewise

Unserdeutsch [plan], do appear. Names are especially relevant here, as <angela> for English [ˈeɪndʒəla] could also be pronounced like in Unserdeutsch [ˈaŋgɛla]. One exceptional rule for two highly frequent homographs was set up from the beginning: the differentiation between the English first person singular nominative pronoun *I* (capital letter) and the equivalent pronoun in Unserdeutsch *i* (small letter, from Standard German *ich* 'I'). For a detailed study of code-switching phenomena, however, it would be inevitable to make additional use of the recording for the disambiguation of words like those mentioned above. A painless solution, unless one wants to give up the morphologically-oriented standard orthography, is the use of a comment tier for the notation of the English- or German-oriented pronunciation of homographs. The problem rarely appears with Tok Pisin because of its strict phonologically oriented orthography, which shows very few homographic equivalents to English or German words. Additionally, hybrid morphological structures can occur, such as in the combination of English verbal stems with German or Tok Pisin suffixes. For example, in a sentence like *er sackim i* ('he sacked me') it is not apparent from the transcript alone whether the stem of the verb *sackim* (English *to sack* + Tok Pisin suffix {-im}) is pronounced like in English (with [ɛ]) or perhaps like in German (with [a]). Again, a remark in the comment tier can solve the problem. Alternatively, a more phoneme-oriented transcription (eye dialect) could have been used in case of an English pronunciation (e.g., in the mentioned case: <säcken>). The issue becomes even more delicate with speakers sporadically pronouncing English words in a pseudo-German style: For example, words like *Japanese* or *Asian* are sometimes pronounced as [japanis] and [asian] etc. In such cases, a remark in the comment tier is inevitable.

(2) Dealing with homophony

In particular cases, the decision whether an expression derives from English or Tok Pisin is not easily made. This concerns quasi-homophone word pairs such as English *New Guinea* versus Tok Pisin *Niugini* or English *boy* versus Tok Pisin *boi*. The problem also occurs in cases where it is not recognizable if a word-final consonant occurs, such as in English *New Ireland* versus Tok Pisin *Niu Ailan*. For various other homophones a decision can be made easily from the context: in the phrase [lɪtl̩ˈmaʊntɛn], for example, the prenominal

adjective *little* (Tok Pisin: *liklik*) clearly indicates that the noun represents the English word *mountain* instead of the quasi-homophone Tok Pisin equivalent *maunten*. In case of proper nouns, however, the case is not always clear, as the donor language could be English as well as Tok Pisin (e.g., *New Guinea* vs. *Niugini*). The whole issue is further exacerbated by the existence of a "Tok Pisin-to-English continuum", resulting from decreolization processes currently ongoing in Tok Pisin (cf. Devette-Chee 2011). Regarding verb forms like *leasim, ownim, rentim, servim, takim, teachim* etc. which are used by some speakers, it is sometimes difficult to decide upon their linguistic status: Should they be dealt with as hybrid constructions (English loan + Tok Pisin transitive marker {-im}), favouring spellings like those chosen above, or as words already fully integrated into (urban-acrolectal, anglicized) Tok Pisin? The latter would suggest spellings like *tichim* (acrolectal form besides *tisim*) and so on. The decisions taken at this stage will influence the results of subsequent analyses anyhow.

(3) Dealing with fluctuating realizations

The delabialization of the German umlaut vowels [y] → [i] and [ø] → [ɛ], being typical for Unserdeutsch, are often realized in a gradually different manner even by the same speaker. Therefore, an articulatory continuum emerges, where a binary decision is not always clear to make. The same issue occurs in basilectal hypercorrect depalatalization [ʃ] → [s] (e.g., Unserdeutsch partly *[s]warze* for German *[ʃ]warze* 'black people', *[s]wer* for German *[ʃ]wer* 'heavy', *[s]tation* for German *[ʃ]tation*), where various articulatory graduations can be found. In the end, a decision for either the one or the other sound is unavoidable for transcribers due to their being restricted to using the alphabet. Only the phonetic transcription of specific passages, which is still to be done, will take account of a finer mapping. Even more difficult seems the decision whether or not a simplification such as the deaffrication [ts] → [s], which is typical for Unserdeutsch, occurs in specific phonetic environments. In cases of an alveolar nasal or, even worse, an alveolar plosive preceding the sound in question it is almost impossible to reach a decision on a purely auditive basis: [gants] ('very') with German-oriented spelling <ganz> or, assuming [gans], spelled <gans>? Similarly, [gɛtsu] ('go to') may be spelled <geht zu> according to German orthography – or otherwise <geht su>, assuming deaffrication in the onset of

the preposition. In such cases, it is only possible to decide by looking at the pattern occurring in other phonetic environments. When in doubt, the standard orthography is to be preferred (here with <z>). The same applies to the potential loss of final consonants, when the consonant in question co-occurs in the onset of the following word or syllable. This is especially so in cases of cliticization: [unˈdan] (‘and then’) may be taken as <und dann> or <un dann>, the latter presupposing a coda cluster simplification.

(4) Dealing with local proper nouns

Various local proper nouns, which are part of the speakers’ personal history but are hard to identify for an outsider, are mentioned in the interviews. These include names of rarely recorded places (e.g., little plantations on the island New Britain), former local employers, or vernacular names for certain animals (e.g., particular fish species) or plants. Names for local food are not always easy to identify either. Most of these difficulties are related to the countless personal names that are brought up in the interviews. In this respect, the Facebook group created for the community as part of the project turned out to be especially useful. Some of the elder speakers did create a Facebook profile solely for the purpose of joining this group, so now they can easily be consulted by the project staff any time. In other instances, sons and daughters of the speakers can be asked for help, who are mostly available on Facebook anyway. This convenient opportunity prevents time-consuming searches and serves to clarify remaining terms not yet identified.

8 PRACTICAL ISSUES 3: ASSIGNMENT PROBLEMS

The further processing of the transcribed language data, i.e. the normalization, lemmatization and annotation of the data, has prompted some questions. These can be summarized in one main point: How far should we be oriented towards Standard German grammar? In particular, if we decide to move away from it, which is highly recommended from at least a typological point of view, how can we do these (semi-)automatized processing steps without too much manual correction effort and how can we make sure that researchers who are not familiar with the structure of Unserdeutsch and with our conventions will find what they are looking for?

(1) Dealing with normalization issues

Our guiding principle for the normalization process – with a few justified exceptions – is that we only do phonological and no morphological normalization. That means we trace back all Unserdeutsch words to their Standard German equivalents, thereby reversing substitutions and deletions of Standard German sounds, but we do not reconstitute Standard German inflectional paradigms, which have been eliminated in Unserdeutsch. Verbs for example, which are generally uninflected in Unserdeutsch, thus do not receive inflected normalized forms.[3] Conversely, of course, we preserve all inflected forms, which do occur in the data (most often used by meso- and acrolectal speakers or with high-frequency lexemes). Another aspect is the capitalization of nouns in Standard German. Since the POS tagger, which is trained on German word material, relies on capitalization in assigning word classes, we adopt the capitalized forms recommended by the program. Toponyms (like *Vunapope, Rabaul*) often have to be corrected by hand, as with Unserdeutsch morphosyntax, the tagger does not recognize them as proper nouns. Another source of error are all German-English homographs in the transcripts (like *mission, station*): The algorithm assumes that these are German words and thus capitalizes them regardless of the cotext (or even the pronunciation). Most errors, however, occur with the automatic normalization of short word forms (with only two or three letters), be they in English, Tok Pisin or Unserdeutsch. The algorithm does not know any of these languages; it only has the lexical entries for different varieties of spoken German. Many of these short words now happen to coincide with reduced regional word forms of spoken German, and consequently, the tagger "recognizes" them and assigns the wrong normalized form (e.g., UD *ma* 'do' → **mal* [correct: *mach*], UD *nu* 'only' → **nun* [correct: *nur*], UD *son* 'already' → **so ein* [correct: *schon*] – or EN *no* → **noch;* EN *we* → **wenn;* EN *is* → **ist*). The only possible solution here is the manual correction of all errors. In the mid-term, such errors could be avoided or reduced by overriding existing entries in the normalization lexicon with entries derived from the Unserdeutsch data.

[3] However, we assign the SG *lemma* to all UD verb forms during the process of lemmatization for reasons of searchability (this concerns verbs, whose uninflected default form deviates from the SG infinitive).

(2) Dealing with annotation issues

Basically, there are four main points that have to be dealt with in the process of the POS tagging. The first issue is the assignment of tags to foreign-language lexemes (mostly English, to a lesser extent Tok Pisin and some single words in other languages such as Japanese or Kuanua). The STTS provides the tag FM (foreign-language material) to avoid the problem of word classification in foreign languages on the basis of a tagset, which has been developed for another language (and, above all, it would be unnecessary work to tag the foreign material, since the researcher is interested in Unserdeutsch). What we do is to extend the FM-tag with a language-specific abbreviation (based on the ISO 639 codes), resulting in FM-EN (English), FM-TP (Tok Pisin) etc. This way later on it will become easy to exclude foreign-language material from searches or to, for example, look for the proportion of Tok Pisin lexemes in Unserdeutsch.[4]

The second issue that comes along with the POS tagging is that we have to modify the STTS tagset for the annotation of Unserdeutsch in some respects: a) Introduction of new categories: This concerns the word class "plural word", resulting in a tag PL, which is assigned to the plural word *alle* 'all' in prenominal position. But since *alle* may also bear the original Standard German meaning 'all' in the same position (and then functions as an indefinite pronoun), it seems difficult to create a rule for the tagger. Thus, this has to be done by hand. (b) Elimination of existing categories: Some grammatical distinctions provided by the tagger are not reasonably applicable to Unserdeutsch. One example is the distinction between finite forms (VVFIN) and infinite forms (VVINF) of verbs, since there is – at least among basilectal speakers – no paradigm of verbal inflection in Unserdeutsch. Therefore, we will eliminate this distinction everywhere, where it is not relevant. (c) Divergent interpretations: in some cases, the tagger recognizes forms and constructions correctly (at least in light of a grammar of spoken German), but, based on the Unserdeutsch grammar, we have to choose another way of

[4] The closed classes of function words and discourse elements from English and Tok Pisin that are fully integrated into Unserdeutsch, i.e. occur frequently across speakers, are, however, tagged according to their grammatical function, as far as they are embedded in Unserdeutsch cotext. This particularly refers to pragmatic elements from Tok Pisin (*orait, maski, nogat* and the question tag *a*) as well as junctions from English (e.g., *whether, cause*). In this way we can assume that future subsequent grammatical analyses will be facilitated.

analyzing them. An example is the tagging of *am* in progressive or habitual constructions, which in Standard German is a contraction (portmanteau) of the preposition *an* and the inflected form *dem* of the definite article. In Unserdeutsch, that analysis is in no way transparent and *am* has a much more specific function (i.e. marking aspectual use of verbs), so we cannot use the STTS tag APPRART (preposition + article) for it. Another example would be the pronoun *die,* which functions as the 3PL personal pronoun in Unserdeutsch (SG: *sie*), but which the tagger can only recognize as a demonstrative pronoun, its function in spoken German.

The third issue in the process of annotating Unserdeutsch is the lack of some features that the tagger requires to assign word classes correctly, based on its rule knowledge from spoken German in Europe. This mainly results from the following conditions: (a) the broad absence of inflectional elements in Unserdeutsch, and (b) special word order rules in Unserdeutsch, especially the absence of a formal distinction between sentence types in Unserdeutsch (e.g., resulting in the problem of recognizing some subjunctives correctly).

Finally, the fourth issue is, obviously, the occurrence of words that are unknown to the tagger because of their language-specific form, such as the gender-indefinite article *de* or the polyfunctional lexeme *fi*. In each of the cases mentioned above, there are three possibilities: (a) manual correction, (b) the creation of a new rule for the tagger, or (c) the training of the tagger by means of a manually annotated Unserdeutsch test corpus to enable it to decide based on likelihood factors. Option (a) of course shows the lowest susceptibility to errors, but is, on the other hand, far more time-consuming. Option (b) only seems reasonable if a clear rule can be formulated without too many exceptions. Option (c) again is a question of cost-benefit ratios, as it takes time to create the test corpus, and afterwards, a manual correction process still seems to be inevitable. For each case, this decision has to be taken individually.

9 OUTLOOK

Upon completion of the Unserdeutsch corpus, it will provide the basis for the linguistic description of the last remaining undocumented (partly) Germanic language. Therefore, a funding for a follow-up project will need to be applied

for at the German Research Foundation (DFG) after having completed the corpus development. A detailed reference grammar for Unserdeutsch should be written as well within this future phase. The unexpected high interest in Germany and beyond – including Australia and Papua New Guinea – obliges us to present results soon; not to mention the relevance of Unserdeutsch for international research far beyond German linguistics. Unserdeutsch provides a rich source for research in language contact and language change, for evolutionary linguistics and sociolinguistics, as well as for creole studies, linguistic typology and other linguistic subdisciplines. The chance to get a profound insight into the history and structure of the only German-based creole language was made possible first and foremost thanks to the cooperativeness and commitment of the last remaining Unserdeutsch speakers.

ABBREVIATIONS

1PL	first person plural	NEG	negation
1SG	first person singular	NN	noun
2SG	second person singular	PL	plural
3PL	third person plural	POSS	possessive
APPRART	preposition + article	PPER	personal pronoun
ART	article	SG	Standard German
ATTR	attributive	TP	Tok Pisin
AUX	auxiliary	TR	transitive
DEF	definite	UD	Unserdeutsch
EN	English	V	Verb
FM	foreign-lang. material	VVFIN	full finite verb
FUT	future	VVINF	full infinite verb

REFERENCES

Austin, Peter. 2014. Language documentation in the 21st century. *JournaLIPP* 3. 57–71.

Austin, Peter & Lenore Grenoble. 2007. Current trends in language documentation. In Peter Austin (ed.), *Language Documentation and Description,* Vol. 4, 12–25. London: SOAS.

Devette-Chee, Kilala. 2011. Decreolization of Tok Pisin: Is there a Tok Pisin-to-English continuum? *Language and Linguistics in Melanesia* 29. 95–103.

DFG. 2013. *Informationen zu rechtlichen Aspekten bei der Handhabung von Sprachkorpora.* URL: <http://bit.ly/1PG4Gq6>; Date accessed: 14.11.2017.

Dwyer, Arienne. 2006. Ethics and practicalities of cooperative fieldwork and analysis. In Jost Gippert, Nikolaus Himmelmann & Ulrike Mosel (eds.), *Essentials of Language Documentation,* 31–66. Berlin: de Gruyter.

Ehlich, Konrad & Jochen Rehbein. 1976. Halbinterpretative Arbeitstranskriptionen (HIAT). *Linguistische Berichte* 45. 21–41.

Gippert, Jost, Nikolaus Himmelmann & Ulrike Mosel (eds.). 2006. *Essentials of Language Documentation.* Berlin: de Gruyter.

Gumperz, John. 1982. *Discourse Strategies.* Cambridge: Cambridge University Press.

Himmelmann, Nikolaus. 2006. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus Himmelmann & Ulrike Mosel (eds.), *Essentials of Language Documentation,*1–30. Berlin: de Gruyter.

Labov, William 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Maitz, Péter & Werner König, & Craig Alan Volker. 2016. Unserdeutsch (Rabaul Creole German). Dokumentation einer stark gefährdeten Kreolsprache in Papua-Neuguinea. *Zeitschrift für Germanistische Linguistik* 44. 93–96.

Maitz, Péter & Craig Alan Volker. 2017. Documenting Unserdeutsch: Reversing colonial amnesia. *Journal of Pidgin and Creole Languages* 32. 365–397.

Maitz, Péter & Craig Alan Volker (forthc.). Unserdeutsch (Rabaul Creole German). In Hans Boas, Ana Deumert, Mark L. Louden & Péter Maitz (eds.), *Varieties of German Worldwide.* Oxford: Oxford University Press.

Schiller, Anne, Simone Teufel & Christine Stöckert. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset)*. URL: <http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf>; Date accessed: 14.11.2017.

Schmidt, Thomas. 2012. EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language. In Nocoletta Calzolari et al. (eds.), *Proceedings of the Eighth Eonference on International Language Resources and Evaluation (LREC'12),* 236–240. Istanbul: European Language Resources Association (ELRA).

Schmidt, Thomas. 2014. The Database for Spoken German – DGD2. In Nocoletta Calzolari et al. (eds.), *Proceedings of the Ninth Conference on International Language Resources and Evaluation (LREC'14)*, 1451–1457. Reykjavik: European Language Resources Association (ELRA).

Schmidt, Thomas & Kai Wörner. 2014. EXMARaLDA. In Jacques Durand, Ulrike Gut, & Gjert Kristofferson (eds.), *Oxford Handbook of Corpus Phonology,* 402–419. Oxford: Oxford University Press.

Schmidt, Thomas, Wilfried Schütte & Jenny Winterscheid. 2015. cGat. *Konventionen für das computergestützte Transkribieren in Anlehnung an das Gesprächsanalytische Transkriptionssystem 2 (GAT 2)*. AGD, IDS Mannheim. URL: <http://bit.ly/2gJbapP>; Date accessed: 14.11.2017.

Selting, Margret et al. 2009. Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung. Online-Zeitschrift zur verbalen Interaktion* 10. 353–402.

Thomason, Sarah. 2015. *Endangered Languages: An Introduction.* Cambridge: Cambridge University Press.

UNESCO. 2003. *Language Vitality and Endangerment. Document submitted to the International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages.* Paris, 10.–12. März 2003. URL: <http://bit.ly/1GDzwra>; Date accessed: 14.11.2017.

Volker, Craig Alan. 1982. *An Introduction to Rabaul Creole German (Unserdeutsch)*. MLitSt thesis, University of Queensland.

Whalen, Douglas. 2004. How the study of endangered languages will revolutionize linguistics. In Piet van Sterkenburg (ed.), *Linguistics Today: Facing a Greater Challenge,* 321–342. Amsterdam & Philadelphia: Benjamins.

Westpfahl, Swantje & Thomas Schmidt. 2013. POS für(s) FOLK: Part of Speech Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch. In Heike Zinsmeister, Ulrich Heid & Kathrin Beck, (eds.), *Das Stuttgart-Tübingen Wortarten-Tagset: Standard und Perspektiven = Journal for Language Technology and Computational Linguistics* 28(1), 139–153. Regensburg: Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL).

Westpfahl, Swantje. 2014. STTS 2.0? Improving the tagset for the part-of-speech-tagging of German spoken data. In Lori Levin & Manfred Stede (eds.), *Proceedings of LAW VIII – The 8th Linguistic Annotation Workshop,* 1–10. Dublin, Ireland: Association for Computational Linguistics and Dublin City University.

Westpfahl, Swantje, Thomas Schmidt, Jasmin Jonietz & Anton Borlinghaus. 2017. *STTS 2.0. Guidelines für die Annotation von POS-Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS).* Version 1, 1. März 2017. URL: <https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/6063>; Date accessed: 14.11.2017.