Kilu von Prince, Düsseldorf

Word units in the Oceanic predicate complex
Project description

1 Starting point

The proposed project investigates word boundaries in the predicate complex in two closely related Oceanic languages. In particular, it focuses on frequency effects on variation in how specific morpheme sequences are realized in spoken and written modalities, and seeks to correlate language-internal with cross-linguistic variation.

Predicates in Oceanic There is great variation between Oceanic languages in terms of how functions are distributed over word units within the predicate complex and in the strength of boundaries between units. For example, in Nalögo, the predicate complex, including negation, subject agreement, tense/aspect/mood marking and the verb root forms one complex word:

(1) Nalögo (Santa-Cruz-Reefs)

te=lë-mno=lü *ma*
NEG1=3AUG.PFV-stay=NEG2 DEM1.PROX
"No, they were not here." (Alfarano, 2021: 249)

By contrast, the same functions are distributed over four (orthographic) word units in Tinrin:

(2) Tinrin (New Caledonia)

kevi *re* *see* *ta* *poka*
1PL.EXCL HAB NEG kill pig
"We did not hunt pigs in those days." (Osumi, 1990: 175)

While the differences in analysis of word units may be to some degree arbitrary, there are objective differences between the two structures exemplified here, which suggest that one forms a more tight-knit unit compared to the other. Thus, in example (1), the negation is realized as a circumfix (or circum-clitic), subject agreement and aspect are encoded by a single affix; in the second example, there is a clearer one-to-one correspondence between forms and functions. Since Oceanic languages, especially in the Melanesian and Micronesian parts of Oceania, are generally under-described (compare • von Prince et al., 2022), neither the range of variation observed here nor its theoretical significance has yet received much attention.

Daakaka and Dalkalaen The predicate complex of the two closely related Oceanic languages Daakaka and Dalkalaen is particularly interesting because of the exceptional range of variation within and across both languages. Daakaka and Dalkalaen are spoken in neighbouring regions on the island of Ambrym, Vanuatu. Each language has around 1000 speakers, many of them are multilingual and have some knowledge of the respective neighbour language. Being non-standardized, small, and mostly unwritten, they provide a rich context for the exploration of variation. The predicate complex in particular shows significant contextual,

	enclitic	proclitic	monosyllabic
Pos. Realis	*=m*	*mw=*	*mwe/mV*
Neg. Realis			*to*
Pos. Potential	*=p*	*w=*	*wV*
Distal	*=t*	*t=*	*tV*
Open Polarity			*doo*
Necessity /neg. Pot.	*=n*		*nV*
Change of State			*bwet*

Table 1: The system of Daakaka TAM markers. V: vowel; Pos: positive; Neg: negative; Pot: potential. From • von Prince (2015: 266)

and inter-speaker variation within each language. At the same time, the realization of TAM markers in the predicate complex is one of the few morpho-syntactic phenomena in which the two languages differ significantly.

Both Daakaka and Dalkalaen were undescribed prior to my documentary work starting in 2009. While my Grammar of Daakaka • von Prince (2015) covers the range of variation in predicate structures, the factors affecting their realization has never been fully explored. On Dalkalaen, I have so far published very little apart from primary data, and no one else has done research on the language so far. This project will provide the first account of the full range of variation within and across the two languages. More generally, the predicate complex of many Oceanic languages remains only vaguely described, especially with respect to word units and word-level phenomena such as vowel harmony and word stress.

Paradigms of Tense/Aspect/Mood markers (TAM) The nexus of variability in the Oceanic predicate complex are the TAM markers, which predominantly occur between the subject marker and the verb root. The syntagmatic dimension of this variation will be explored in more detail below. The paradigmatic variation of the Daakaka TAM system is introduced in table 1. As can be seen there, most markers in the system are highly variable, such that they can be realized as one-consonant enclitics to the preceding subject marker, as one-consonant proclitics to the subsequent verb root (if it starts with a vowel), or in syllablic shape with a vowel that is partially determined by context. There are however some markers which are always realized in the same way, such as the negative realis marker *to*. The fact that some items in the paradigm are more variable than others is crucial for examining the specific effects of language-internal variability on cross-linguistic differences (RQ3) and gives us a baseline to evaluate the phonetic and graphetic variability of TAM markers, as further explained in 2.3.2.

The observed variation presents challenges to rule-based theories about language architecture in several ways: The same functional units are distributed very unevenly onto phonological word units; the boundary strength between morphemic units appears to be gradient rather than categorical; some of the observed variation is amenable to an analysis in terms of rules and exceptions, but there remains a substantial amount of variation both across linguistic contexts and individuals that cannot easily be reduced to a strictly rule-based framework.

I will therefore adopt a usage-based perspective and operate under the assumptions that units such as words and morphemes, along with morpho-syntactic rules, emerge from gradient, probabilistic knowledge (cf. Kapatsinski, 2018: 2.2.4); and that boundaries between words and morphemes are gradient rather than categorical (Hay & Baayen, 2005).

Contextually determined variation Some of the variation we find is illustrated by several different realizations of the Daakaka potential marker in the examples below. In example (3), we see the potential marker form one phonological word unit with the preceding subject marker. By contrast, in example (4) the potential marker combines with the subsequent verb root. In example (5), the potential marker has a CV shape instead of just a consonant. The vowel harmonizes with the (first) vowel of the subsequent verb, which also gives the sequence of TAM marker and verb a word-like appearance. But in example (6), the potential marker receives a vowel that does not harmonize with the subsequent verb, thus resembling an independent word. Finally, when the TAM marker is preceded by a subject agreement marker and the verb root starts with a bilabial consonant, the potential marker is not realized at all, as illustrated in (7).

(3) Enclitic /p/:
Da=p lyung vyan pyan!
1D.IN=POT bathe go under
"Let's dive!"

(4) Monoconsonantal proclitic /w/:
ko w=en we!
2SG POT=eat first
"Please eat!"

(5) Vowel-harmonic proclictic /wV/:
sam oko=an ka wu=vu
2SG.POSS travel=NMLZ ASR POT=be.good
"your journey will be successful."

(6) Syllabic with disharmonic vowel:
ka wa mini vyos
ASR POT drink coconut
"she will drink a coconut"

(7) Not realized before bilabial consonant:
Da=∅ vyan
1.INCL.DU=POT go
"Let's go!"

Note that orthographic representations of word units are based on the design process with speaker communities. The question of which role word boundaries play in the two target languages is one objective of the research project, and will be critically examined throughout. For Daakaka, the range of different realizations and contextual factors affecting them has been briefly been described in • von Prince (2015), but has not been explored in full detail. For Dalkalaen, such an account is missing entirely.

Frequency effects While some realizations are quite easily predictable from context, part of the observed variation cannot be described in terms of simple rules. This applies in particular to the decision between vowel-harmonic and vowel-disharmonic syllabic realizations. In this context, many observations can be phrased in terms of lexeme-specific rules, which however do not hold great explanatory value. For stronger, more explanatory generalizations, it is necessary to consider the effect of frequencies.

The factor of vowel harmony is particularly interesting here, because 1) it is not well described for Oceanic languages, and 2) it is thought to be an important indicator of wordhood (see below). Alderete & Finley (2016: 770) state that "very few [Oceanic languages] have morphophonemic vowel harmony". This might be true for the Polynesian languages these authors work on. In Melanesia, morpho-phonemic vowel harmony is quite possibly under-reported,[1] but it is by no means unheard of. In a very preliminary survey of Oceanic language descriptions, I found morphophonemic vowel harmony in more than 20 languages, including, for

[1]For example, Crowley (2004) reports that the vowel in the Bislama transitivizing suffix *-Vn* often depends on the preceding vowel of the verb root, but does not describe this in terms of vowel harmony.

example, Neve'ei (Musgrave, 2007: 23ff), Sakao (Crowley, 2002: 17ff), and also Proto-Oceanic (Ross, 1988). All of these languages are from Melanesia, except for Puluwatese, which is spoken in Micronesia (Elbert, 1974: 52f).

As Schiering (2006: 166) observes, "[v]owel harmony is especially strong as a means of word demarcation in languages which have not developed segmental effects of stress in the form of vowel reduction." Van Kampen et al. (2008) found that Turkish children use vowel harmony to segment words, while German children do not. This corresponds with a difference in the prominence of word stress, which is very pronounced in German, but plays a much more minor role in Turkish.

I have argued in • von Prince (2015) that Daakaka has no word stress, which suggests all the more that vowel harmony is central to marking word boundaries. At the time of designing the orthographies, I did not realize that vowel-harmonic realizations of TAM markers can signal that they form a word unit with the verb root. This project will explore the implications of this idea and shine a light on vowel harmony and its role in signalling word units in Oceanic languages.

One hypothesis I will explore in this project is that frequency-based preferences for specific word-shapes affect the choice of realization of the predicate complex. For example, the most frequent syllable pattern in lexical words is bisyllabic, so I hypothesize that TAM markers will cliticize more systematically to monosyllabic verb roots, yielding a bisyllabic phonological word, than to verb roots with more syllables.

For a preliminary proof of concept, I looked up sequences of TAM markers with monosyllabic vs. multisyllabic verb roots in my corpus. Figure 1 shows the correlation between vowels of TAM markers and vowels of verb roots, comparing monosyllabic verb roots with multi-syllabic verb roots in Daakaka. The graphs show a tendency for vowel harmony with both monosyllabic and multi-syllabic verb roots. In both conditions, /a/ and /e/ in particular, are used for disharmonic realizations. For multi-syllabic verb roots, TAM markers with /a/ are more often disharmonic than harmonic, but the same is not true for monosyllabic verb roots.

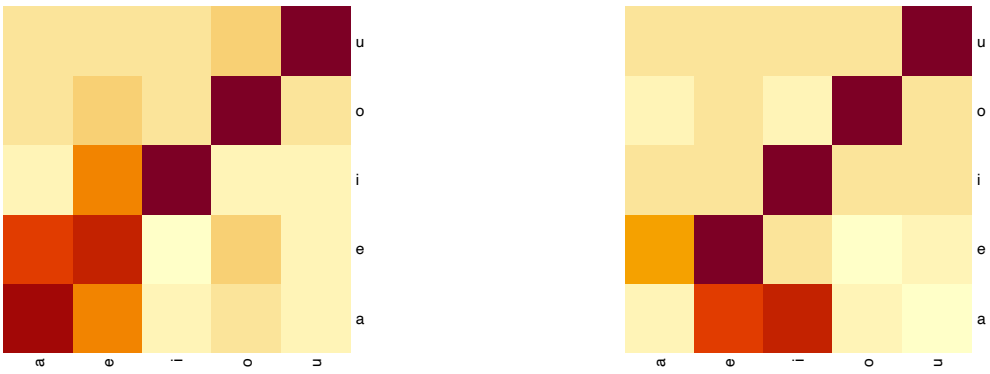

Figure 1: Correlations between vowels of syllabic TAM markers and vowels of verb roots; y-axis: Vowel of syllabic TAM marker; x-axis: Vowel of subsequent verb root. Left: monosyllabic verb roots (3345 tokens); Right: multi syllabic verb roots (2054 tokens).

Another potential example for word-shape preferences is the distal TAM marker *tV*, which is often realized as vowel-harmonic with the subsequent predicate *minyes* "to be different", while the realis TAM marker *m(w)V* is generally realized as a full syllable with a disharmonic vowel:

(8) Realis marker with disharmonic vowel:
ma minyes
REAL be.different
“it is different”

(9) Distal marker with harmonic vowel:
na ti=minyes
COMP DIST=be.different
“which is different”

The difference in assimilation between the two markers to this particular verb may well be accounted for in terms of how frequently each sequence occurs. The distal marker precedes this verb much more often than the realis marker, because this verb is used more often in relative clauses, for which the default TAM marker is the distal marker, than it occurs as the predicate of the main clause, where it would mostly be preceded by the realis marker. Another possible factor is that Daakaka has a general preference for bisyllabic words, but for those words that have more than two syllables, /t/ is the most frequent onset (• von Prince, 2015, 2017: 28).

Such observations suggest that vowel-harmonic realizations of syllabic markers are more likely when they result in a highly frequent word-shape, and when the sequence itself is highly frequent. This hypothesis has yet to be tested systematically, and the relative impact of the two factors will have to be assessed.

Variation in spontaneous orthographies Recent investigations into hand-written texts suggest that the morphological composition of a word impacts its graphemic realization (e. g. Berg, 2019). The two languages Daakaka and Dalkalaen were largely unwritten prior to my work with the speaker communities. There was some written use in the language in text messages, and individual written representations, for example on the local church building in Emyotungan. Speakers are usually literate in Bislama and either English or French. Their expectations about the written representation of their language can be assumed to be influenced by their knowledge of different orthographies. In both language communities, people established a language committee to design an orthography for each language with my support. In addition, I held writing competitions in both communities to explore speakers’ intuitions about orthographic conventions. While the results have fed into the design of the orthography, I have not yet analysed the range of variation between speakers with respect to word boundaries in the predicate complex.

The following examples illustrate some of the variation speakers produce in writing. The first line shows a direct transcription of their handwritten records, the second line a transliteration into the orthography later developed. The variation in how the predicate complex is separated into orthographic words corresponds to the variation in the spoken language, as illustrated below:

(10) Syllabic TAM marker with harmonic vowel forms a separate orthographic word unit from the verb root:
Bosu ma miny mo nok
bosu ma mini mo=nok
cat REAL drink REAL=finish
“the cat finished drinking”

(11) Syllabic TAM marker with harmonic vowel forms one orthographic word unit with the verb root:
yam tovasse, monok te titilie
ya=m towaase **mo=nok** te ti-tilye
3PL=REAL clean REAL=finish then REDUP-tear

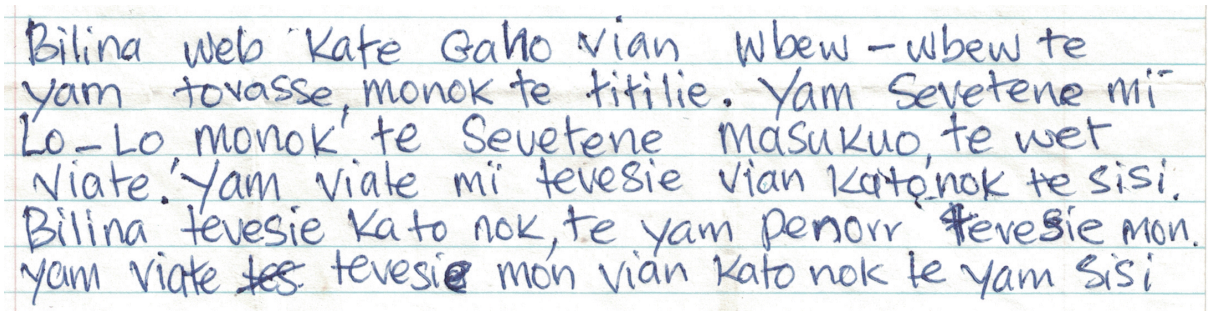

Figure 2: Excerpt from a handwritten text in Daakaka. See example ((11)) for a partial transcription.

"they clean it then tear it into stripes"

(12) Syllabic TAM marker with disharmonic vowel forms one orthographic word unit with the verb root:

dake	*laman*	*vyan*	*te*	*yase*	*wasukuo*
dange	laman	vyan	te	yaase	**wa**	**sukuo**
pour	lemon	go	then	turn	POT	be.together

"pour in the lemon juice and mix it together"

The spontaneous orthographies produced by speakers yield a rare opportunity to probe intuitions about word units and morpho-phonology, before spontaneous choices are obliterated by standardization. Moreover, since the primary records are handwritten, they not only allow for categorical differences in terms of word and morpheme boundaries, but can also yield quantitative, gradual measures of distances between orthographic word units and other graphetic correlates of morpho-syntactic boundaries.

Synchronic variation and diachronic change Daakaka and Dalkalaen are very closely related, with an estimated cognate rate of about 83% (• von Prince, 2015: 4). Even so, they are not mutually intelligible to individuals without much exposure to the other vernacular, which is why I describe them as different languages rather than dialects. Syntactically, too, the two languages are extremely close. These observations suggest that the two languages have only recently started to diverge from a common ancestor language. I believe that this allows us to form tentative hypotheses about the relation between intra-linguistic and cross-linguistic variation between the two languages. To be precise, I believe we should expect that greater variation within each language should correspond to a greater difference between the two languages. This expectation derives from the following assumptions:

1. Synchronic variation facilitates diachronic change. This assumption should be fairly uncontroversial. It has been developed in the seminal paper by Weinreich et al. (1968) and has shaped the study of language change ever since.
2. Given the great similarity between Daakaka and Dalkalaen, I assume that their ancestral language was also very similar to both languages. This entails that those grammatical features that are highly variable in one language, were probably also variable in the ancestral language; and those areas that are highly stable in one current language were, with some probability, also stable in the ancestral language. While it is of course possible that, even over short time spans, a language develops variability in a previously stable area of its grammar, or stabilizes a highly variable one, I assume that, over

the entire grammar, a similar distribution of variability is more likely than a dissimilar distribution.

3. The second assumption also allows us to generate a prediction about another closely related descendent of our ancestral language: It, too, should, by and large, show higher variability in those areas where S_1 is highly variable than in those areas where it is highly stable.
   So far, this reasoning has lead us to the following interim conclusion: From the observation that the two languages are very similar in general, we can conclude that they should also be similar in terms of how variability is distributed over their grammars, and they should also be similar to their closest common ancestor language in this respect. Taking it one small step further, we can also conclude that if both S_1 and S_2 are highly variable in area 1, then the ancestor language is even more likely to also have high variability in area 1.
4. We get yet more leverage for interesting expectations when we combine assumptions 1 and 2. If variation is a catalyst for change, and the next common ancestor of S_1 and S_2 should have a similar distribution of variation across its grammar as either language, we can expect that **variability within each language** should correlate with **differences between the two languages**. For those areas of the grammar which are highly variable in one language, we can expect to see greater cross-linguistic differences than for stable areas. We should see the greatest differences in those areas that are highly variable in both languages.
   Applied to the predicate complex, a very preliminary assessment lends initial plausibility to this idea. Among the TAM markers of Daakaka, the potential marker has an especially wide range of possible realizations. In Dalkalaen, the variability of the corresponding marker is apparently more limited, but still relatively high. At the same time, the realization of the potential marker is one of the most striking differences between the two languages. In Dalkalaen, depending on the environment, it even switches its position to the very start of the predicate complex, in contrast to Daakaka, where it occurs strictly between subject marker and verb root:

(13) Dalkalaen:
ba muju yan fyan
POT 2PC go down
"Go down!" (to several people)

(14) Daakaka:
ka=p kueli vyan
2DU=POT return go
"Go back!" (you two)

   By contrast, the negative realis marker *to*, which shows no allomorphy in either language, is used and realized in a very similar manner across the two languages.

In order to work on the questions introduced in the previous section, it will be necessary to quantify the variability of different TAM markers and longer sequences within the predicate complex. I have previously assessed word order variability with entropy-based measures (• von Prince & Demberg, 2018; • Berdicevskis et al., 2018). The same methodology may not be feasible for the dataset at hand, but an approximation of similar measures will be possible through the number of possible realizations for each TAM marker, the number of realizations for fixed sequences of TAM marker plus a given verb root, and the relative frequencies of each realization per marker.

Quantifying and correlating the variability of forms and the difference between the two languages will advance the emerging field of dialectology based on natural-language corpora (Grieve, 2015), as well as the empirical investigation of morpho-phonemic variability and its relation to language change.

2 Objectives and work program

2.1 Anticipated total duration of the project

The project requires funding for a duration of three years. The present proposal is a revision of a proposal for a four-year-long project within the context of the CRC 1675 "Using Complex Words" (project B02).

The CRC as a whole was not recommended for funding, but project B02 received the highest possible grade, with the German predicate "exzellent".

The original work plan included an ambitious work package on sociolinguistic variables and individual lexicons. The current proposal does not include this in order to make the project feasible within the duration of three instead of four years.

2.2 Objectives

This project sets out to investigate the following research questions:

RQ1: What is the full extent of variation in the predicate complex in each language? Documenting the full range of variation within each language is a central goal of this project and will significantly increase our knowledge about both languages, and about variation in small, non-standardized languages more generally.

Specifically, I will test the following hypotheses:

HYP1 Word-shape preferences and sequence frequencies affect preferences for realizations of the predicate complex.

HYP2 Vowel-harmonic versions of syllabic TAM markers are more likely when the resulting sequence resembles a highly frequent word-shape or if the sequence itself is highly frequent.

HYP3 In contexts that show considerable variation between utterances, the identity of a phoneme should be less clear compared to highly predictable contexts. The phonetic properties of a highly predictable phoneme should be less ambiguous than those of a phoneme that is hard to predict. For example, when a context strongly predicts that a TAM marker is realized with the vowel /u/, the corresponding vowel sounds should show a lesser degree of variation between utterances compared to contexts in which the TAM marker may be realized with /u/ or /e/ with roughly equal probability.

RQ2: What is the relation between spoken and written word units in spontaneous orthographies? I will use the hand-written records from the writing competitions I organized as a rich source of information on speaker intuitions about word units and vowel qualities. The following hypotheses will be investigated:

HYP1 Greater variation in the written language corresponds to greater variation in the spoken language.

HYP2 Vowel-harmonic TAM markers will more often be represented as one orthographic unit with the verb root than disharmonic syllabic TAM markers.

HYP3 Morpho-syntactic boundaries might be gradual rather than categorical.

RQ3: How does variation within each language relate to differences between the two languages?

HYP1 Greater variation in a certain area of the grammar within one language corresponds to greater variation in the same grammatical area in the other language.

HYP2 Those areas of the grammar that are the most variable within the two languages will show the greatest differences between the two languages.

2.3 Work programme and methods

2.3.1 Data

The main empirical basis for this project will be the data I have already collected during fieldwork between 2009-2013 and in 2017. Firstly, I have collected a substantial amount of recordings from both languages, which I have transcribed, translated and enriched with additional annotations and metadata. These will be used to investigate both the morpho-phonological and the fine-grained phonetic variation in the predicate complex of each language. They will also allow for a first exploration of how sociolinguistic factors correlate with speaker-specific preferences.

Secondly, I held writing competitions in each language community before developing an orthography for each language to probe for intuitions about written representations. The results of these competitions have fed into the orthography design but were never analysed in their own right. In this project, I will analyse the handwritten records with a particular focus on the orthographic representation of word units. These data have been transliterated into the standard orthography I designed, glossed, tagged for part-of-speech information (POS), and translated to English.

Ref	Language	Tokens	Modality	Transcription	Translation	Gloss	POS
Daa1	Daakaka	67k	spoken	yes	yes	yes	yes
Daa2	Daakaka	1k	written	yes	yes	yes	yes
Daa3	Daakaka	11k	spoken	yes	yes	no	no
Dal1	Dalkalaen	24k	spoken	yes	yes	yes	yes
Dal2	Dalkalaen	9k	written	yes	yes	yes	yes
Dal3	Dalkalaen	13k	spoken	yes	yes	no	no

Table 2: Data collection at the point of writing

Additional fieldwork will fill in any relevant gaps in the corpus data and test specific hypotheses about unattested or rarely attested forms. An overview of the available corpus data is given in table 2.

2.3.2 Work packages

WP1: Preprocessing The purpose of this WP is to prepare the corpus data for the specific needs of the project.

1. Existing annotations of the predicate complex will need to be re-checked for consistency.
2. The semi-parallel sub-corpus will be enriched with glosses and POS-tags.
3. The annotations for the spoken corpora are currently time-aligned at the utterance-level. I will collaborate with Ludger Paschen (ZAS Berlin) to explore a phoneme-level time alignment using forced alignment with MAUS (cf. Paschen et al., 2020). Minimally, TAM markers and their contexts will receive manual or semi-automatic phoneme-based alignments for fine-grained investigations of their phonetic properties.

4. Handwritten records from the writing competition will be digitized in high quality. I will consult with the archival team of HHU's library for best practices in digitizing these originals.

WP2: Quantitative analysis of spoken corpora This WP directly serves to provide a partial answer to RQ1 (*What is the full extent of variation in the predicate complex in each language?*). It also provides the basis for generating hypotheses that will be tested in other WPs.

1. Assessing the relative variability of different TAM markers;
   One way to assess variability consists in counting the number of different realizations. For example, for the Daakaka potential marker, we can count each of the different realizations illustrated in examples (3) through (7), giving this marker a variability count of five. Of course, other ways of counting are possible. For example, instead of counting "vowel-harmonic" and "vowel-disharmonic" realizations as two categories, one could count each vowel combination as one category as in {*wa, we, wo, wu*}, yielding four realizations instead of two, or a total of seven instead of five. We will start out with the most fine-grained classification, and reserve the option for more coarse classifications later on. This methodology is inspired by Witzlack-Makarevich et al. (2022).
   To further approximate relative variability and predictability, we will also consider how many realizations are attested for a fixed sequence of a specific TAM marker and a specific verb root; and how frequent different realizations of TAM markers and sequences are.
   This part of the project provides the empirical basis for much of the other WPs. It directly generates hypotheses for WP3 (addressing HYP1 of RQ1, *Word-shape preferences and sequence frequencies affect preferences for realizations of the predicate complex*); it also serves as input to the next step.
2. Assessing the impact of frequency effects on realizations to address HYP1, HYP2 of RQ1 (*Vowel-harmonic versions of syllabic TAM markers are more likely when the resulting sequence resembles a highly frequent word-shape or if the sequence itself is highly frequent*).
3. Quantitative analysis of phonetic variation in the predicate complex, with a focus on vowel qualities of TAM markers to assess HYP3 of RQ1 (*In contexts that show considerable variation between utterances, the identity of a phoneme should be less clear compared to highly predictable contexts.*).

The results of this WP will be tested through elicitations in WP4.

WP3: Analysis of written corpora This WP will investigate how speakers have chosen to write their language prior to a standardized orthography, based on the data I have collected during two writing competitions.

1. Graphemic analysis: Based on the transcriptions, we will test HYP1 and HYP2 of RQ2 (*Greater variation in the written language corresponds to greater variation in the spoken language*; *vowel-harmonic TAM markers will more often be represented as one orthographic unit with the verb root than disharmonic syllabic TAM markers*).
2. Graphetic analysis: We will collaborate with Kristian Berg (Uni Bonn) and Stefan Hartmann (HHU) to explore novel methodologies of semi-automatic extraction of graphetic information from scanned pictures of hand-written texts. We will explore methods to quantify distances between letters in handwritten texts. We will focus on bigrams which span the boundaries between morphemes. This project will investigate distances

between letters to address HYP3 (*Morpho-syntactic boundaries might be gradual rather than categorical*). In particular, we will test whether TAM markers with variable vowels are more variable in their distance to the verb root than TAM markers such as the negative realis marker *to*, which always have the same vowel.

WP4: Elicitations The results produced from WP2 will be tested and further refined through elicitations in the field. We will extend the existing collection of semi-parallel corpora based on storyboard stimuli to include both more stimuli and involve more speakers. New storyboards will be designed to target specific sequences of TAM markers and verb roots that are not well enough represented in the pre-existing corpus data to allow to differentiate between different hypotheses. For example, trisyllabic verbs are relatively rare, so to investigate whether TAM markers are vowel-harmonic with trisyllabic verbs with different initial vowels, we may want to create stimuli to elicit a corresponding set of verbs in combination with a range of different TAM markers.

WP5: Qualitative analysis of the Dalkalaen predicate complex Dalkalaen has never been described beyond a handful of observations (• von Prince et al., 2019; • von Prince & Margett, 2019). One important outcome of this project will be a complete descriptive account of variation of word units in the predicate complex. In particular, this WP contains the following components:

1. A sketch grammar of Dalkalaen that highlights its contrast to Daakaka and other neighbouring languages.
2. A full account of contextual factors such as the shape of the verb root which impact the realization of the predicate complex.

WP6: Quantifying cross-linguistic differences, correlations with language-internal variation In this WP, we will compare the language-specific measures obtained in WP2 and WP4 to assess the differences between the two languages. Directly addressing RQ3 (*How does variation within each language relate to differences between the two languages?*), we will assess whether the two languages have a similar distribution of variability over their grammars (HYP1); and whether they diverge most clearly from each other in the areas that show the highest level of variability (HYP2).

Table 3: Timeline

Month	WP1	WP2	WP3	WP4	WP5	WP6
1-6						
7-12						
13-18						
19-24						
25-30						
31-36						
Output		1-2 articles	1-2 articles	1 article, materials for speaker communities	PhD thesis	1-2 articles

2.4 Handling of research data

I have an excellent track record of sustainable research data management, having received an honorable mention for the 2019 DELAMAN Franz Boas Award for my archived corpus collection at The Language Archives. My elicitation materials are publicly available at zenodo.org and listed with the http://tulquest.huma-num.fr/ platform.

Using pre-existing corpus data The main empirical basis for this project consists in my previously collected corpora, which are archived at The Language Archives in Nijmegen, and currently being duplicated at the IDS in Mannheim. These will be revised and in part enriched with additional annotations.

Collecting new data through fieldwork New data will be collected through fieldwork. In collecting and publishing these data, the project will adhere closely to FAIR and CARE standards (Carroll et al., 2020; Wilkinson et al., 2016). All data will be extensively documented with metadata to ensure maximal sustainability.

All newly generated storyboards and other elicitation materials will also be published and documented.

Publication and archiving of new data Data will be archived and published in accordance with best practices for language documentation (• Seyfeddinipur et al., 2019). Subsets of the data used as the basis of publications will be published along with the respective articles on platforms provided by the publisher or osf.io. All newly generated data will be archived with Language Archives Cologne (LAC).

Software The project will use open-source software such as R, Praat and ELAN. Scripts for processing will be published along with the corresponding articles and data.

Storage and sharing For the duration of the project, the data will be stored on local HHU servers and shared through safe platforms such as sciebo for collaborative work. The project will consult with the HHU RDM Competence Center and the Universitäts- und Landesbibliothek for all matters concerning research data and software management.

2.5 Relevance of sex, gender and/or diversity

In working with speakers, we will try to create a diverse sample, including different genders and generations. I will make sure that speaker communities receive tangible benefits resulting from our work.

In hiring staff, I always seek to promote qualified candidates whose identities are underrepresented in Western academia, concerning, for example, the dimensions of ability, neurodivergence, gender, as well as social, cultural and linguistic backgrounds. I understand that this implies that I provide

1. accommodations to make our work flow accessible to my staff,
2. support to compensate for a lack of accessibility in academic infrastructure,
3. and opportunities to learn navigate a tier of society they weren't born into.

3 Project- and subject-related list of publications

Alderete, J. & S. Finley. 2016. Gradient vowel harmony in Oceanic. *Language and Linguistics* 17(6). 769–796. doi:https://doi.org/10.1177/1606822X16660960.

Alfarano, V. 2021. *A grammar of Nalögo, an Oceanic language of Santa Cruz Island*. Paris: INALCO dissertation. https://theses.hal.science/tel-03421587.

- Berdicevskis, A., Ç. Çöltekin, K. Ehret, K. von Prince, D. Ross, B. Thompson, C. Yan, V. Demberg, G. Lupyan, T. Rama & C. Bentz. 2018. Using Universal Dependencies in cross-linguistic complexity research. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, 8–17. Brussels, Belgium: Association for Computational Linguistics. https://www.aclweb.org/anthology/W18-6002.

Berg, K. 2019. *Die Graphematik der Morpheme im Deutschen und Englischen*. Berlin & Boston: de Gruyter. doi:https://doi.org/10.1515/9783110604856.

Carroll, S. R., I. Garba, O. L. Figueroa-Rodríguez, J. Holbrook, R. Lovett, S. Materechera, M. Parsons, K. Raseroka, D. Rodriguez-Lonebear, R. Rowe et al. 2020. The CARE principles for indigenous data governance. *Data Science Journal* doi:http://doi.org/10.5334/dsj-2020-043.

Crowley, T. 2002. Sakao. In J. Lynch, M. Ross & T. Crowley (eds.), *The Oceanic Languages* Curzon Language Family Series, 599–607. Richmond, Surrey: Curzon.

Crowley, T. 2004. *Bislama reference grammar*, vol. 31 Oceanic Linguistics Special Publication. Honolulu: University of Hawai'i Press.

Elbert, S. H. 1974. *Puluwat grammar*. The Australian National University: Pacific Linguistics.

Grieve, J. 2015. Dialect variation. In D. Biber & R. Reppen (eds.), *The Cambridge handbook of English corpus linguistics*, 362–380. Cambridge University Press. doi:https://doi.org/10.1017/CBO9781139764377.021.

Hay, J. B. & R. H. Baayen. 2005. Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences* 9(7). 342–348. doi:https://doi.org/10.1016/j.tics.2005.04.002.

van Kampen, A., G. Parmaksiz, R. van de Vijver & B. Höhle. 2008. Metrical and Statistical Cues for Word Segmentation: Vowel harmony and Word Stress as Cues to Word Boundaries by 6- and 9-Month-Old Turkish Learners. In A. Gavarró & M. J. Freitas (eds.), *Language Acquisition and Development*, 313–324. Newcastle: Cambridge Scholars Publishing.

Kapatsinski, V. 2018. *Changing Minds Changing Tools: From Learning Theory to Language Acquisition to Language Change*. MIT Press. https://mitpress.mit.edu/9780262346320/changing-minds-changing-tools/.

Musgrave, J. 2007. *A grammar of Neve'ei, Vanuatu*. The Australian National University: Pacific Linguistics.

Osumi, M. 1990. *A grammar of Tinrin (New Caledonia)*: Australian National University dissertation.

Paschen, L., F. Delafontaine, C. Draxler, S. Fuchs, M. Stave & F. Seifart. 2020. Building a Time-Aligned Cross-Linguistic Reference Corpus from Language Documentation Data (DoReCo). In *Proceedings of the 12th Language Resources and Evaluation Conference*, 2657–2666. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.324.

- von Prince, K. 2015. *A grammar of Daakaka*. Berlin & Boston: De Gruyter Mouton. doi:https://doi.org/10.1515/9783110766318.

- von Prince, K. 2017. Daakaka dictionary. *Dictionaria* (1). 1–2171. https://dictionaria.clld.org/contributions/daakaka.

- von Prince, K. & V. Demberg. 2018. POS tag Perplexity as a measure of syntactic complexity. In *Online Proceedings of MLC2018 (Measuring Linguistic Complexity, sattelite workshop of EVOLANG XII in Torun, Poland)*, http://www.christianbentz.de/MLC2018/Prince_Demberg.pdf.

- von Prince, K., A. Krajinović & M. Krifka. 2022. Irrealis is real. *Language* 98(2). 221–249.

- von Prince, K., A. Krajinović, M. Krifka, V. Guérin & M. Franjieh. 2019. Mapping irreality: Storyboards for eliciting TAM contexts. In A. Gattnar, R. Hörnig, M. Störzer & S. Featherston (eds.), *Proceedings of linguistic evidence 2018: Experimental data drives linguistic theory*, 187–207. Tübingen, Germany: University of Tübingen. doi:http://dx.doi.org/10.15496/publikation-32623.

- von Prince, K. & A. Margett. 2019. Expressing possibility in Daakaka and Saliba-Logea. *Studies in Language* 43(3). 628–667.

Ross, M. 1988. *Proto-Oceanic and the Austronesian languages of western Melanesia*, vol. 98 Pacific Linguistics: Series C. Canberra: Research School of Pacific and Asian Studies, Australian National University. Publication of PhD, ANU 1987.

Schiering, R. 2006. *Cliticization and the evolution of morphology: A cross-linguistic study on phonology in grammaticalization*. Konstanz: Universität Konstanz Phd dissertation. http://nbn-resolving.de/urn:nbn:de:bsz:352-opus-18728.

- Seyfeddinipur, M., F. Ameka, L. Bolton, J. Blumtritt, B. Carpenter, H. Cruz, S. Drude, P. L. Epps, V. Ferreira, A. V. Galucio, B. Hellwig, O. Hinte, G. Holton, D. Jung, I. K. Buddeberg, M. Krifka, S. Kung, M. Monroig, A. N. Neba, S. Nordhoff, B. Pakendorf, K. v. Prince, F. Rau, K. Rice, M. Riessler, V. Szoelloesi Brenig, N. Thieberger, P. Trilsbeek, H. van der Voort & T. Woodbur. 2019. Public access to research data in language documentation: Challenges and possible strategies. *Language Documentation & Conservation* 13. 545–563. http://hdl.handle.net/10125/24901.

Weinreich, U., W. Labov & M. I. Herzog. 1968. Empirical foundations for a theory of language change. In *Directions for Historical Linguistics: A Symposium*, Austin, Texas: University of Texas Press.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3. doi:https://doi.org/10.1038/sdata.2016.18.

Witzlack-Makarevich, A., J. Nichols, K. Hildebrandt, T. Zakharko, B. Bickel, A. L. Berez-Kroeker, B. McDonnell, E. Koller & L. B. Collister. 2022. Managing AUTOTYP data: Design principles and implementation. In A. L. Berez-Kroeker, B. McDonnell, E. Koller & L. B. Collister (eds.), *The Open Handbook of Linguistic Data Management*, Cambridge: MIT Press. doi:https://doi.org/10.7551/mitpress/12200.003.0061.

4 Supplementary information on the research context

4.1 Ethical and/or legal aspects of the project

4.1.1 General ethical aspects

We will collaborate with speakers of Dalkalaen and Daakaka to obtain more data about the corresponding languages. As such, we will

- take care to obtain fully informed consent on the modalities of publishing and archiving the data;
- ensure that the data are accessible to the speaker communities;
- protect speaker-specific metadata;
- avoid the recording of sensitive content;
- negotiate solutions with speaker communities such that they benefit from our research.

I have good experiences with both speaker communities and have created literacy materials and written records of stories as well as local digital collections before. I'm looking forward to exploring these options further. I have obtained clearance from the HHU ethics committee for a more elaborate version of this project.

4.1.2 Descriptions of proposed investigations involving humans, human materials or identifiable data

We will primarily elicit specific linguistic contexts through storyboards. We will collect metadata on speakers about certain criteria such as age, language background, and sex, to the extent that speakers consent. These data will not be publicly available. They will typically be archived along with the speech data, but only accessible to registered users and anonymized as well as possible given the small speaker communities.

4.1.3 Descriptions of proposed investigations involving experiments on animals

n/a

4.1.4 Descriptions of projects involving genetic resources (or associated traditional knowledge) from a foreign country

n/a

4.1.5 Explanations regarding any possible safety-related aspects ("Dual Use Research of Concern; foreign trade law)

n/a