Tagging modality in Oceanic languages of Melanesia

Annika Tjuka
Humboldt-Universität zu Berlin
tjukanni@hu-berlin.de

Lena Weißmann
Freie Universität Berlin
lena.weissmann@fu-berlin.de

Kilu von Prince
Humboldt-Universität zu Berlin
kilu.von.prince@hu-berlin.de

Abstract

Primary data from small, low-resource languages of Oceania have only recently become available through language documentation. In our study, we explore corpus data of five Oceanic languages of Melanesia which are known to be mood-prominent (in the sense of Bhat, 1999). In order to find out more about tense, aspect, modality, and polarity, we tagged these categories in a subset of our corpora. For the category of modality, we developed a novel tag set (MelaTAMP, 2017), which categorizes clauses into *factual*, *possible*, and *counterfactual*. Based on an analysis of the inter-annotator consistency, we argue that our tag set for the modal domain is efficient for our subject languages and might be useful for other languages and purposes.

1 Introduction

Our understanding of the Oceanic languages of Melanesia has so far been based mostly on descriptive accounts rather than primary data, since no documentation existed until recently. For some of these languages, high-quality corpora have now become available, but their exploration is still in its infancy.

In our MelaTAMP research project, we carry out a comparative, corpus-based study on tense, aspect, and modality (TAM) categories in seven Oceanic languages: Daakaka, Dalkalaen, Daakie, Mavea, Nafsan, Saliba-Logea, North Ambrym (cf. MelaTAMP, 2017). Speaker populations range from about 30 (Mavea) to around 6000 (Nafsan). TAM-related meanings are often expressed obligatorily within the verbal complex, sometimes in more than one place. Thus, Mavea has three preverbal slots for expressing TAM values; in addition, some subject-agreement markers also express the difference between realis and irrealis modalities and reduplication can be used to express pluractionality (see Table 1). By contrast, Saliba-Logea only uses optional particles to express TAM-related meanings.

In this paper, we discuss our tag set and its application in a subset of texts in the corpora of five languages: Daakaka, Dalkalaen, Mavea, Nafsan, and Saliba-Logea. The focus of our paper is on the process of tagging modality.

Previous studies which tag modality in corpora have focused on differentiating between modal flavours such as deontic and epistemic, and modal forces such as necessity and possibility. Thus, the sentence in (1-a) expresses an epistemic possibility while (1-b) conveys a deontic necessity.

(1) a. *Naomi might be a surgeon.*
b. *Martha must hand in her assignment tomorrow.*

These distinctions are notoriously difficult to tag, with coarse-grained ontologies yielding better results than more fine-grained ones (Rubinstein et al., 2013). Most approaches focus on modal auxiliaries such as *must*, and modal adverbs such as *probably* (Cui and Chi, 2013; Quaresma et al., 2014).

In the languages of our project, however, modal auxiliaries and adverbs are rare, and do not play the same role in expressing modality as they do in many European languages. Instead, verb moods, such as *realis* and *irrealis*, are largely responsible for the modal interpretation of a clause. These expressions are usually under-specified for modal force and flavour. Instead of modal forces and flavours, we therefore differentiate three modal categories based on a branching-times framework (von Prince, 2019), which is explained in section 3.2.

The ontology of our modal tag set was primarily motivated by theoretical concerns and preliminary experiences with the driving factors in Oceanic TAM systems. The targets of our tags were individual clauses, regardless of the presence of spe-

Proceedings of the 13th Linguistic Annotation Workshop, pages 65–70
Florence, Italy, August 1, 2019.

SBJ.AGR	COND	NEG	IT/INCPT	NUM	IMPF	REDUP-	**Verb**	ADV	TR	OBJ
i-, ...	*mo-*	*sopo-*	*me-/pete-*	*r-/tol-*	*l(o)-*				*=i*	*=a*/NP

Table 1: The verbal complex in Mavea (Guérin, 2011).

cific modality-related expressions. Their TAM values were tagged according to their temporal-modal reference, irrespective of the presence of specific TAM markers (e. g., in *Emma wants [to eat ice cream]*, the infinitive complement clause would be tagged to refer to the (relative) possible future).

The analysis of inter-annotator consistency in the tagging process shows that our modal categories are reasonably easy to assign based on the translations into English. This suggests that the same ontology might be useful for other purposes and languages as well.

2 Data

The data of our study consists of a series of narrative and explanatory texts in corpora of five Oceanic languages. These corpora are the result of language documentation and are richly annotated, with morpheme-by-morpheme glosses, part-of-speech tags, translations into English, as well as metadata on speakers, text genre, and the circumstances of the recording. In addition, we enriched parts of the corpora with our own tag set for TAM values. For optimal facilities for searching and analysis, we imported all corpora to the ANNIS platform (Krause and Zeldes, 2016). We used Druskat (2018) to import them from their native SIL Toolbox format.

The corpora of the MelaTAMP project are held and versioned in a git repository (MelaTAMP, 2017). The repository itself is private and currently only accessible by members of the project team. Published versions of each corpus are available from various archives: von Prince (2013a,b); Krifka (2013); Guérin (2006); Thieberger (2006); Franjieh (2013); Margetts et al. (2017).

3 The Tag Set

3.1 Overview

In an initial stage of exploration, we identified comparable texts across the corpora (see Table 2). Each of the selected 26 texts was segmented into annotation units, which often correspond to a single sentence. These units were further segmented into clause-based subdivisions for TAM annotation (1953 clauses in total). Each clause was annotated for clause type, temporal reference, modal reference, aspect, and polarity. Our tag set which consisted of five categories with 21 tags is displayed in Table 3. Compared to some previous approaches, our ontology of clause types is richer than, e. g., Leech and Weisser (2003), but far less fine-grained than Twitchell and Nunamaker (2004); our tag set for tense is less fine-grained than, e. g., Zymla (2017). These differences are mostly due to different goals and data. We concentrated on those categories that were most likely to determine differential TAM marking in our subject languages. The tag set for clauses should be applicable for similar purposes to other languages. The tag sets for temporal and aspectual reference would have to be more fine-grained to accommodate graded tense systems, highly differentiated aspect systems, and similar.

3.2 The Modal Tag Set

We found that, for our subject languages, the distinction which is most useful and basic to the TAM systems is the distinction between *realis* and *irrealis*, as is often the case in Oceanic (compare Lichtenberk, 2016). At the same time, irrealis is a very large modal domain that is often subdivided by more specific markers. This can be modeled by the approach of von Prince (2019), which shows that a branching-times framework can be used to generate three different modal domains: the possible (future), the actual (past and present), and the counterfactual (past, present and future). This differs crucially from previous approaches to modality which were based on a binary distinction, without the option to exclusively quantify over counterfactual indices. It is this theoretical innovation which allows for a tag set that is more informative than a mere realis/irrealis distinction, without relying on the often elusive distinctions between modal flavors.

Given the assumptions in von Prince (2019), the three domains are defined as follows:

- The actual present i_0 and the actual past (predecessors of i_0).

Language	#Texts	#Tokens	#Texts taggged	#Clauses taggged
Daakaka	119	68k	5	143
Dalkalaen	114	34k	6	724
Mavea	61	45k	3	639
Nafsan	110	65k	6	364
Saliba-Logea	214	150k*	6	159
Total	618	362k	26	2029

Table 2: Corpora included in this study; *of the 150k tokens in this corpus, about 70k are fully annotated.

Category	Name	Tags
Clause type	clause	assertion, question, directive; embedded: proposition, conditional, e.question, temporal, adverbial, attributive
Temporal domain	time	past, future, present
Modal domain	mood	factual, counterfactual, possible
Aspectual domain	event	bounded, ongoing, repeated, stative
Polarity	polarity	positive, negative

Table 3: Tag set of the MelaTAMP project (MelaTAMP, 2017).

- The counterfactual past, present, and future: indices that are neither predecessors nor successors of i_0.
- The possible future(s): successors of i_0.

Figure 1 illustrates the three domains of modality.

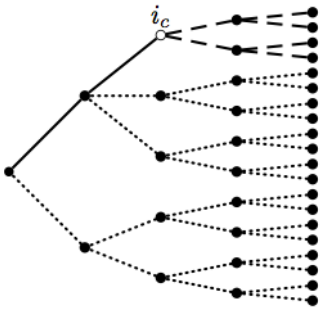

Figure 1: The three domains of the factual (solid line), the counterfactual (dotted lines), and the possible future (dashed lines). Vertically aligned indices are here taken to be simultaneous.

For the purposes of our tag set, we make a three-way distinction which builds on those domains, but is not entirely identical to them. The three values that we use are `factual`, `counterfactual`, and `possible`: the tags `factual` (*it rained*) and `counterfactual` (*she should have run faster*, *winning would have been hard*) coincide with the corresponding domains. The tag `possible` comprises several domains, depending on the temporal reference of the clause: the possible future (*it will rain*) and quantification over both the actual and the counterfactual domain (*it may have rained*).

Tagging was mainly based on the English translations of the texts although in some cases, the glosses were considered as well, when translations were unclear. Each clause was tagged manually by two annotators: Annika Tjuka and Lena Weißmann. There were no discontinuous clauses. The sentence in (2) was tagged as follows:

(2) *tenem iya Gesila stoli-na*
that.DIST 3SG Place.Name story-3SG.POSS
"that's the story of Gesila" (Saliba-Logea: Gesila_01BC_0265)

- clause: `assertion`
- time: `present`
- mood: `factual`
- event: `stative`
- polarity: `positive`

After a text was tagged by the two annotators independently, the tags of both versions were compared by one of the annotators and the inconsistencies were noted in a table and discussed. If the decision for either one of the tags was clear, the correct tag was inserted in the final document. Many early sources of disagreement were clarified by guidelines in the documentation of the tag set (MelaTAMP, 2017). In doubtful cases, the tags were discussed with the principal investiga-

tor of the project: Kilu von Prince. The inter-annotator agreement was calculated on the basis of the inconsistencies in each tag which were detected through the initial comparison.

In addition to corpus work, we and our collaborators also carried out field work in Vanuatu to elicit modal-temporal contexts that were rarely attested in the corpora. We report on this work in von Prince et al. (2018).

4 Analysis of Inter-Annotator Consistency

A total number of 9765 tags in 1953 clauses (five tags per clause) were assigned by the two annotators. In 817 tags, inconsistencies between the annotation of the annotators were present. Figure 2 illustrates the inter-annotator consistency and inconsistency in each category of the tag set.

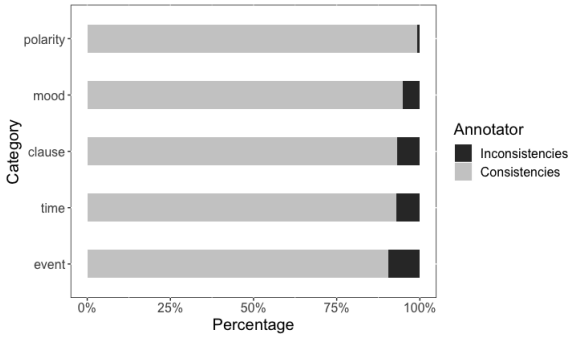

Figure 2: Percentages of inter-annotator consistencies (light) and inconsistencies (dark) in each TAM category of the tag set.

The graph shows that the percentages of inconsistencies between the categories differ. Mismatches are especially prone to arise in the event category. This category has the lowest inter-annotator agreement with $\alpha = 0.79$.[1] In contrast, the polarity category had the lowest inconsistency percentage with 0.82%. The α score in this category is $\alpha = 0.91$.

The analysis of each tag in the mood category reveals differences between the percentage of inconsistencies, as illustrated in Figure 3.

The 12.7% of the inter-annotator inconsistency in the `possible` tag is based on 496 clauses which are tagged as `possible`. Most of these inconsistencies result from mismatches in tagging

[1]The Krippendorff's alpha coefficient measures the statistical agreement between two annotators (Krippendorff, 1980).

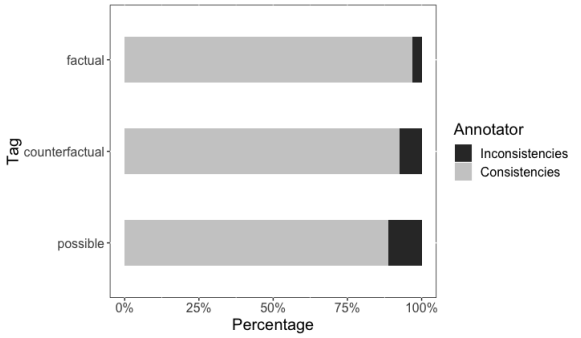

Figure 3: Percentages of inter-annotator consistencies (light) and inconsistencies (dark) in each tag of the mood category.

temporal sentences, see (3). Thus, in the following example, it is hard to tell whether the sentence implies that the agents did reach their destination or whether it only implies that they were headed there:

(3) *. . . panpan na ra=pak nagis*
until PURP 3D.RS=to point
"[they went] until they got to the [next] point" (Nafsan: 036.017)

Among the small number of clauses which had the `counterfactual` tag (37 clauses), there were 8.11% inconsistencies. In general, counterfactual sentences are rare and are not easy to detect. A prominent context for counterfactual modality is false-belief-reports (compare Van Linden and Verstraete, 2008), as the embedded clause in example (4); or conditional clauses referring to situations in the past that did not occur, as the two clauses in example (5).

(4) *ru=mroki [na ruk=fan sol tete*
3PL.RS=think COMP 3PL.IR=go get some
mane emrom st]o.
money inside shop
"they thought [someone had taken money from inside the shop]." (Nafsan: 030.048)

(5) *[taba lahi ya mwamwayauma]*
IRR yesterday 1SG.SBJ quick-to.SP
[kabo ya kai]
then 1SG.SBJ eat
"If I **had hurriedly come here** yesterday then I **would have eaten**." (Saliba: Boneyawa_05BC_0020)

The `factual` tag is the most consistent tag in the mood category with 3.17% inconsistencies in 1481 clauses. The tag is based on the factual do-

main of the branching-times framework and was assigned to clauses expressing the actual present and past, as in (6).

(6) *mwe liye an bosi*
REAL take 3S.POSS copra.chisel
"He took his copra chisel." (Daakaka: 0139)

The evaluation of the mood category results in an α score of $\alpha = 0.85$ which can be considered acceptable (cf. Carletta, 1996). This result reveals how efficient the tag set in this category seems to be.

5 Discussion

In this paper, we explored the tagging of TAM categories in corpora of five Oceanic languages with a focus on the modal domain. Selected texts were divided into clause-based annotation units which were then tagged by two annotators based on the previously established tag set. The two versions of the tagged texts were then compared manually in order to identify and resolve mismatches in certain cases. The results of the inter-annotator consistency show that our tag set works especially well in the mood category.

In comparison to more fine-grained distinctions, e.g., as proposed in Rubinstein et al. (2013), the differentiation between the tags `factual`, `counterfactual`, and `possible` seems to be less prone to inter-annotator inconsistencies. Their basic score of $\alpha = 0.49$ in the Modality Type (Rubinstein et al., 2013) was much lower than our overall result ($\alpha = 0.85$). Only when they collapsed priority types (i.e., bouletic, teleological, bouletic/teleological, deontic, and priority) and non-priority types (i.e., epistemic, circumstantial, ability, epistemic/circumstantial, ability/circumstantial), they achieved an α score of 0.89. This indicates that the distinction in more than three levels results in an unreliable annotation compared to a coarse-grained approach.

Our methodology also differs from previous approaches to tagging modality in that we did not identify a specific target set of expressions to label – such as modal auxiliaries and adverbs – but tagged all clauses within a selected set of texts. We believe that this approach is particularly useful for languages that rely more on verb moods such as irrealis and subjunctive, as opposed to lexical expressions such as auxiliaries, for the expression of modality. Depending on the languages and the goals of tagging modality, our tag set may therefore be an interesting alternative to other models.

6 Conclusion

We presented a novel approach for tagging the modal domain in mood-prominent languages (cf. Bhat, 1999) which contributes to a more stable inter-annotator consistency. The overall tag set that we used to annotate the TAM categories exhibits a high percentage of inter-annotator consistency throughout different categories. In addition, our modal tag set has been proven useful for our purposes and provides an alternative to previous distinctions based on modal flavours.

Acknowledgments

We would like to thank three anonymous reviewers for their valuable feedback and suggestions. And we would like to thank the DFG (German Research Foundation) for funding this research (MelaTAMP project, 273640553).

References

D. N. S. Bhat. 1999. *The Prominence of Tense, Aspects, and Mood*. John Benjamins.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254.

Xiaomeng Cui and Ting Chi. 2013. Annotating Modal Expressions in the Chinese Treebank Yanyan.

Stephan Druskat. 2018. ToolboxTextModules (Version 1.1.0).

Michael Franjieh. 2013. *A documentation of North Ambrym, a language of Vanuatu*. SOAS, ELAR., London.

Valérie Guérin. 2006. *Documentation of Mavea*. SOAS, ELAR, London.

Valérie Guérin. 2011. *A grammar of Maŭea: An Oceanic language of Vanuatu*. University of Hawai'i Press, Honolulu.

Thomas Krause and Amir Zeldes. 2016. ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1):118–139.

Manfred Krifka. 2013. *Daakie, The Language Archive*. MPI for Psycholinguistics, Nijmegen.

Klaus Krippendorff. 1980. *Content analysis: An introduction to its methodology*. Sage publications.

Geoffrey Leech and Martin Weisser. 2003. Generic speech act annotation for task-oriented dialogues. In *Proceedings of the Corpus Linguistics 2003 conference*, volume 16, Lancaster. Lancaster University.

Frantisek Lichtenberk. 2016. Modality and Mood in Oceanic. In Jan Nuyts and Johan van der Auwera, editors, *The Oxford Handbook of Mood and Modality*, chapter 14, pages 330–361. Oxford University Press, Oxford.

An Van Linden and Jean-Christophe Verstraete. 2008. The nature and origins of counterfactuality in simple clauses: Cross-linguistic evidence. *Journal of Pragmatics*, 40:1865–1895.

Anna Margetts, Andrew Margetts, and Carmen Dawuda. 2017. *Saliba/Logea, The Language Archive*. MPI for Psycholinguistics, Nijmegen.

MelaTAMP. 2017. Primary data repository – MelaTAMP. https://wikis.hu-berlin.de/melatamp.

Kilu von Prince. 2019. Counterfactuality and past. *Linguistics and Philosophy*.

Kilu von Prince. 2013a. *Daakaka, The Language Archive*. MPI for Psycholinguistics, Nijmegen.

Kilu von Prince. 2013b. *Dalkalaen, The Language Archive*. MPI for Psycholinguistics, Nijmegen.

Kilu von Prince, Ana Krajinović, Manfred Krifka, Valérie Guérin, and Michael Franjieh. 2018. Mapping Irreality: Storyboards for Eliciting TAM contexts. In *Proceedings of Linguistic Evidence 2018*.

Paulo Quaresma, Amália Mendes, Iris Hendrickx, and Teresa Gonçalves. 2014. Tagging and Labelling Portuguese Modal Verbs. In J. Baptista, N. Mamede, S. Candeias, I. Paraboni, T.A.S. Pardo, and M.G. Volpe Nunes, editors, *Computational Processing of the Portuguese Language*, volume 8775 of *PROPOR 2014. Lecture Notes in Computer Science*, pages 70–81. Springer, Cham.

Aynat Rubinstein, Hillary Harner, Elizabeth Krawczyk, Daniel Simonson, Graham Katz, and Paul Portner. 2013. Toward fine-grained annotation of modality in text. In *Proceedings of IWCS 10, WAMM*, Potsdam.

Nick Thieberger. 2006. *Dictionary and texts in South Efate*. Digital collection managed by PARADISEC.

Douglas P. Twitchell and Jay F. Nunamaker. 2004. Speech act profiling: A probabilistic method for analyzing persistent conversations and their participants. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*, volume 37, pages 1713–1722.

Mark-Matthias Zymla. 2017. Comprehensive annotation of cross-linguistic variation in tense and aspect categories. In *12th International Conference on Computational Semantics*.