Tagging modality in Oceanic languages of Melanesia

Annika Tjuka, Lena Weißmann, and Kilu von Prince

Gefördert durch
DFG Deutsche Forschungsgemeinschaft

August 1st, 2019
The 13th Linguistic Annotation Workshop

The MelaTAMP project

Introduction

Figure 1: Subject languages of the MelaTAMP project.

The Languages

- Subject languages: Daakaka, Dalkalaen, Daakie, Mavea, Nafsan, Saliba-Logea, and North Ambrym.
- Speaker populations range from about 30 (Mavea) to around 6000 (Nafsan).
- So far, our understanding of the Oceanic languages of Melanesia is based mostly on descriptive accounts.

The MelaTAMP Project

- Comparative research
- Based on corpus data
  - Texts were recorded during fieldwork sessions with speakers of the respective language.
- Investigation of modality, aspect, tense, and polarity (TAMP) in Oceanic languages.

The focus of this talk is on our study on tagging modality in five of the seven subject languages.

Expressing TAMP

- TAM-related meanings are often expressed obligatorily within the verbal complex, sometimes in more than one place.

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

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

Expressing TAMP

- TAM-related meanings are often expressed obligatorily within the verbal complex, sometimes in more than one place.

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

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

- In contrast, Saliba-Logea only uses optional particles to express TAM-related meanings.

Data

Corpora

- Corpora of the following languages were considered in this study: **Daakaka**, **Dalkalaen**, **Mavea**, **Nafsan**, and **Saliba-Logea**.
- In comparison to previous approaches, we did not identify a specific target set of expressions to label (e.g., modal auxiliaries and adverbs).

Sub-Corpus

- Prioritizing of a comparable sub-corpus (**26 texts**).
  - Descriptions of wild-life behaviour, tales and fables about miraculous events including mysterious figures and animals native to the region.

Overview

Language	Total		Tagged	
	#Texts	#Tok.	#Texts	#Clauses
Daakaka	119	68k	5	141
Dalkalaen	114	34k	6	658
Mavea	61	45k	3	634
Nafsan	110	65k	6	363
Saliba-Logea	214	150k*	6	157
Total	618	362k	**26**	**1953**

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

Method

Previous Approaches to Tagging Modality

- Differentiation between **modal flavours** such as *deontic* and *epistemic* and **modal forces** such as *necessity* and *possibility*.
- These distinctions are notoriously difficult to tag (Rubinstein et al., 2013).

Our Tag Set

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, see `https://wikis.hu-berlin.de/melatamp/Main_page`.

Branching-times Framework

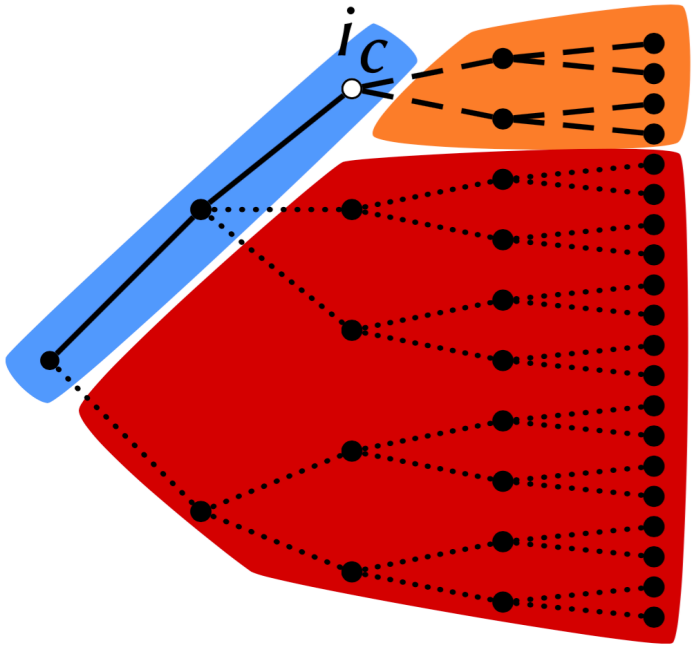

Figure 2: The three domains of the **factual** (solid line), the **counterfactual** (dotted lines), and the **possible** future (dashed lines). Vertically aligned indices are here taken to be simultaneous (von Prince, 2019).

Example: `factual`

(1) *mwe liye an bosi*
REAL take 3SG.POSS copra.chisel
"He took his copra chisel." (Daakaka)

- clause: `assertion`
- time: `past`
- **mood:** `factual`
- event: `bounded`
- polarity: `positive`

Example: `counterfactual`

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

- clause: `proposition`
- time: `past`
- **mood: `counterfactual`**
- event: `bounded`
- polarity: `positive`

Example: `possible`

(3) *ka* *na-p* *pwer-pwer* *yen* *or*
MOD 1SG-POT REDUP-stay in bush
"I will live in the bush." (Daakaka: 1348)

- clause: `assertion`
- time: `future`
- **mood:** `possible`
- event: `stative`
- polarity: `positive`

Results of Inter-Annotator Agreement

Results in each Category

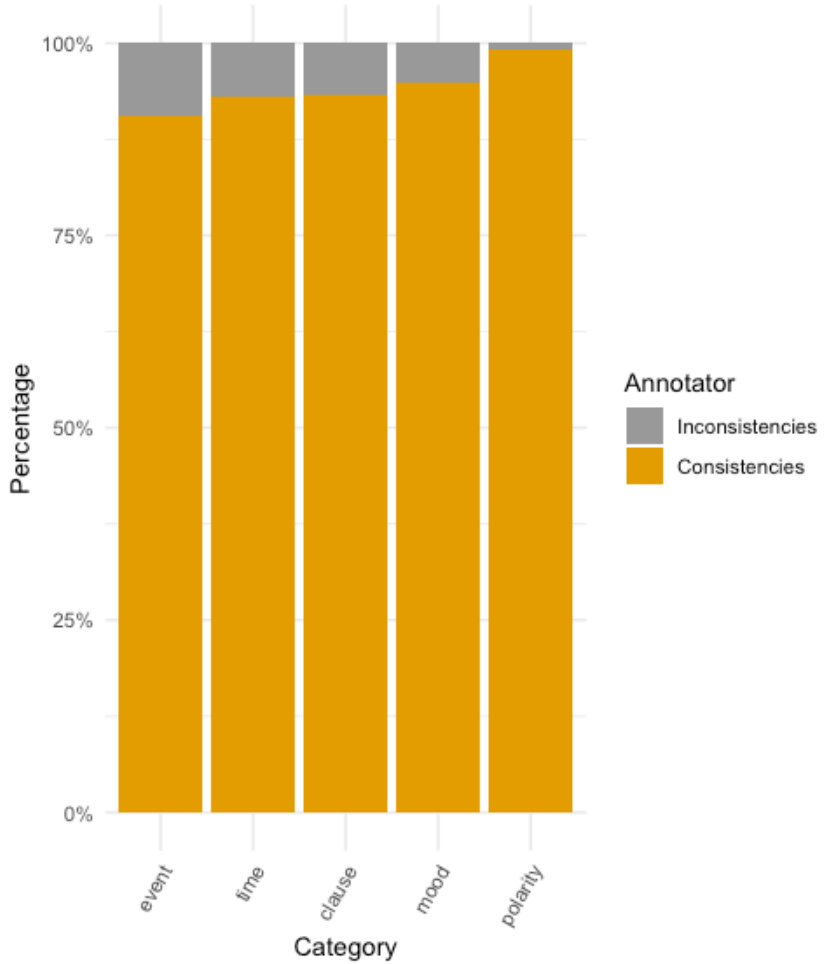

Figure 3: Percentages of total inter-annotator consistencies (orange) and inconsistencies (grey) in each TAMP category of the tag set.

Inter-Annotator Agreement Score for each Category

- Polarity: $\alpha^1 = 0.91$
- **Mood: $\alpha = 0.86$**
- Clause: $\alpha = 0.85$
- Time: $\alpha = 0.85$
- Event: $\alpha = 0.79$

[1] Krippendorff's alpha coefficient (Krippendorff, 1980).

Results in the Mood Category

- The high α score in this category indicates that our three-way distinction (`factual`, `counterfactual`, `possible`) seems to be efficient.

Conclusion

Conclusion

- The overall tag set that we used to annotate the TAM categories exhibits a high percentage of inter-annotator consistency throughout different categories.

Conclusion

- The overall tag set that we used to annotate the TAM categories exhibits a high percentage of inter-annotator consistency throughout different categories.
- Our modal tag set has been proven useful for our purposes.

Conclusion

- The overall tag set that we used to annotate the TAM categories exhibits a high percentage of inter-annotator consistency throughout different categories.
- Our modal tag set has been proven useful for our purposes.
- Depending on the languages and the goals of tagging modality, our tag set may be an interesting alternative to other models.

Conclusion

- The overall tag set that we used to annotate the TAM categories exhibits a high percentage of inter-annotator consistency throughout different categories.
- Our modal tag set has been proven useful for our purposes.
- Depending on the languages and the goals of tagging modality, our tag set may be an interesting alternative to other models.

Thank you!

References

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

Druskat, Stephan, 2018. ToolboxTextModules (Version 1.1.0).

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

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

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

Klecha, Peter, 2011. Optional and obligatory modal subordination. In Proceedings of Sinn und Bedeutung, volume 15.

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

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

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

Margetts, Anna, Andrew Margetts, and Carmen Dawuda, 2017. Saliba/Logea. The Language Archive.

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

von Prince, Kilu 2019. Counterfactuality and Past. Linguistics and Philosophy.

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

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

Rubinstein, Aynat et al. 2013. Toward fine-grained annotation of modality in text. In Proceedings of IWCS 10, WAMM, Potsdam.

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