Habitual aspect as a property of text spans

Annika Tjuka[1], **Lena Weißmann**[2], **Kilu von Prince**[1]
[1]Humboldt-Universität zu Berlin, [2]Freie Universität Berlin
tjukanni@hu-berlin.de
9th Language & Technology Conference 2019

The MelaTAMP project

- We investigate **TAM systems** ...
- ...in **seven Oceanic languages** ...
- ...of **Melanesia** (Vanuatu and Papua New Guinea) ...
- primarily based on **corpus data**.

SBJ.AGR | TAM | (AUX) | (REDUP-) | V

Figure: The typical, simplified anatomy of an Oceanic predicate

Method

1. Stage: Parallel Texts

- Looking for keywords in the English translations that are indicative of habitual contexts such as "used to/ would/ always/ usually/ often"
- Comparison of the use of verbal morphology in the matches to determine whether it occurred also in other aspectual contexts
- Utilizing pre-existing metadata on genres and content to identify texts that were likely to include habitual descriptions
- For example: the genre *explanation* often features descriptions of wild-life behaviour and thus typical generic statements, such as "the banded rail lives on the ground"
- Result: 26 parallel texts

2. Stage: Segmentation

- Segmentation of the texts into annotation units, which often correspond to sentences
- Further subdivision of these units into clauses for TAM annotation (1953 clauses in total)

Tagging

In the 26 parallel texts, each clause was annotated for clause type, temporal reference, modal reference, aspect, and polarity. These tags enabled us to compare certain contexts in the languages with different TAM systems.

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

The tagging was mainly based on the English translation of the texts although in some cases, the glosses were considered as well. An example of a annotation is given in (1).

(1) *Meerin yaapu nyoo ya=m du*
long.time big.man 3PL 3PL=REAL stay
"Long ago, there were great men" (Daakaka: 1388)

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

Data

The data of our project consists of corpora of seven languages in total, five of which were taken into account for this study. Each corpus contains a variety of texts which were recorded during fieldwork sessions with speakers of the respective language: Daakaka, Dalkalaen, Mavea, Nafsan, Saliba-Logea. Published versions of each corpus are available: von Prince (2013a,b); Krifka (2013); Guérin (2006); Thieberger (2006); Franjieh (2013); Margetts et al. (2017).

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

Certain stories and themes are widespread throughout the region and were present in more than one corpus, such as stories about the origin of the coconut. We thus created a sub-corpus of 26 texts in five of our seven subject languages.

Results and Discussion

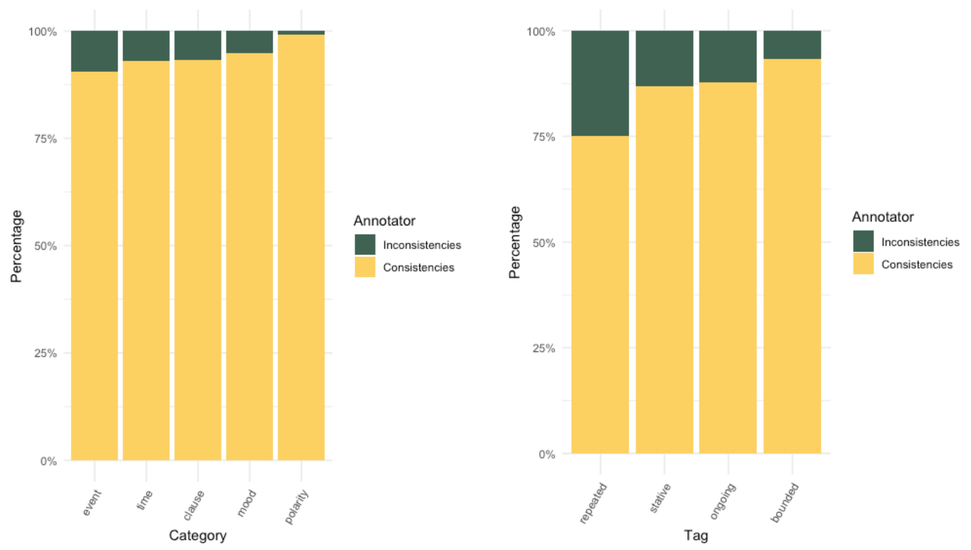

Figure: Percentages of total inter-annotator consistencies (yellow) and inconsistencies (green) in each TAM category of the tag set (left) and in each tag of the event category (right).

The analysis of the inter-annotator consistency revealed that mismatches between the two annotators occurred in 817 tags of a total of 9765 tags (including typos and other easily resolved mismatches). The number of mismatches is greatest in the event category: $\kappa = 0.79$. In contrast, the polarity and mood categories, $\kappa = 0.91$ and $\kappa = 0.85$, respectively, have an almost perfect inter-annotator agreement (Carletta, 1996).

Habituality

Most inconsistencies in the *repeated* tag occur in passages which describe a habitual context, as in (2). While the context of the passage is habitual, individual clauses within the passage might differ with respect to their local aspectual values. Thus, in (2), the first part includes a bounded event description, while the second part is stative.

(2) a. *hinage ta dup-paisowa*
also 1INCL.SBJ DUP-work
"we work hard too" (Saliba: Tautolowaiya_01AG_0048)

b. *kamna-da te se yababa*
feeling-1INCL.POSS near.SP 3PL.SBJ bad
"and we feel tired" (Saliba: Tautolowaiya_01AG_0049)

The fact that habituality can be a property of passages has been previously discussed and it has been argued that the property of genericity or habituality ranges over an entire stretch of discourse, as in (3) (Carlson & Spejewski, 1997).

(3) a. My grandmother used to bake the most wonderful pies every Saturday.
b. She went to the orchard on Shady Lane early in the morning.
b'. The alarm clock would **have gone** off at 6 a.m.
c. She then would pick a basket each of apples and peaches.
c'. Cows would be in the orchard **mooing** at her.

Conclusion

The comparison of the annotation of parallel texts in corpora in different languages highlights the difference between **passage-level** aspect and **clause-level** aspect. In some cases, habituality can span over sequences of several clauses which indicates that it can be a property of passages, not only of individual clauses. Moreover, it is well possible that habitual passages are just one special case of a much more general situation: there are several well-documented cases in which mood, aspect or tense are a property of entire passages or texts that combine with partially independent, local TAM values at the clause level. These include:

- **modal subordination**: *A wolf might come in here. It would eat you first* (Roberts, 1989; Klecha, 2011);
- **indirect speech** in German *Konjunktiv I*:

(4) *Sie sagte, die Ergebnisse **seien** erfreulich. Es **habe** außergewöhnlich viele Einreichungen gegeben.*
3SG.F said the results be.KONJI pleasing it have.KONJI exceptionally many submissions give.PARTPERF
"She said the results were pleasing. There had been [according to her] exceptionally many submissions."

- **present-in-the-past** (*It is the year 1990. The cold war is coming to an end.*);
- and **sequence-of-tense** phenomena (*Esra was determined to win the race. She would train every day.*).

Depending on the intended scope and degree of granularity, one may have to take these passage-wide properties into account when tagging TAM contexts.

References

Carletta, Jean. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics* 22(2). 249–254.
Carlson, Greg N. & Beverly Spejewski. 1997. Generic passages. *Natural Language Semantics* 5(2). 101.
Franjieh, Michael. 2013. *A documentation of North Ambrym, a language of Vanuatu*. London: SOAS, ELAR.
Guérin, Valérie. 2006. *Documentation of Mavea*. London: SOAS, ELAR.
Klecha, Peter. 2011. Optional and obligatory modal subordination. In *Proceedings of sinn und bedeutung*, vol. 15, 365–379.
Krifka, Manfred. 2013. *Daakie, the language archive*. Nijmegen: MPI for Psycholinguistics.
Margetts, Anna, Andrew Margetts & Carmen Dawuda. 2017. *Saliba/Logea*. The Language Archive.
MelaTAMP. 2017. Primary data repository – MelaTAMP. `https://wikis.hu-berlin.de/melatamp`.
von Prince, Kilu. 2013a. *Daakaka, the language archive*. Nijmegen: MPI for Psycholinguistics.
von Prince, Kilu. 2013b. *Dalkalaen, the language archive*. Nijmegen: MPI for Psycholinguistics.
Roberts, Craige. 1989. Modal subordination and pronominal anaphora in discourse. *Linguistics and Philosophy* 12. 683–721.
Thieberger, Nick. 2006. *Dictionary and texts in South Efate*. Digital collection managed by PARADISEC.