

Category clustering: A probabilistic bias in the morphology of verbal agreement marking

Supplementary material 1: Homophony avoidance

A minor limitation of our Chintang prefix bigram analysis is the small number of tokens for the *ma-*, *mai-* 1NS.P prefixes in combination with other prefixes. We have only 14 observations of *ma-*, *mai-* combining with SUBJ prefixes, and just five observations combined with *mai-* NEG.¹ However this poverty of data in itself points to an interesting pattern: the number of *ma-*, *mai-* 1NS.P + *mai-* NEG tokens turns out to be unexpectedly small, when we compare this with overall corpus counts. The *ma-*, *mai-* prefixes, which encode exclusive and inclusive versions of 1NS.P respectively, are associated with speakers from the village of Mulgaū, while *kha-* 1NS.P, making no clusivity distinction, is associated with speakers from the village of Sambugaū. However this is not a rigidly maintained dialectal distinction, and the corpus includes a substantial number of speakers using both forms (N=72). Both variants are relatively recent innovations, drawing on distinct etymological sources (*kha-* < **khəl* ‘all’; *ma-*, *mai-* < **rak-mi* ‘person’) to provide a strategy for speaker effacement under the influence of Maithili politeness strategies (Bickel & Gaenszle 2015).

Counting the usage of these dialectal variants in the corpus shows that overall, *ma-*, *mai-* are about four times as frequent as *kha-*. Yet in combination with *mai*-NEG, we find that there are fewer tokens of *ma-*, *mai-* than of *kha-* (Table 1). According to expectations established by the general corpus frequency, the number of tokens of combined *ma-*, *mai-* 1NS.P + *mai-* NEG is unexpectedly small, and highly unlikely to be a result of chance ($\chi^2 p < 0.001$).

Table 1. Corpus tokens of *ma-*, *mai-* 1NS.P and *kha-* 1NS.P

	<i>kha-</i> 1NS.P	<i>ma-</i> , <i>mai-</i> 1NS.P
General corpus	372	1453
Combined with <i>mai</i> - NEG ***	16	5

Homophony avoidance is the obvious explanation for the unexpectedly rare usage of the *ma-*, *mai-* 1NS.P variant in combination with *mai*- NEG, following other instances of probabilistic morphological homophony avoidance attested in recent literature (Baerman 2011; Kaplan & Muratani 2015). This is however the only instance we know of that involves avoidance of homophony in potentially adjacent morphemes, as opposed to paradigmatic alternants. There is a further dimension to the homophony

¹ The counts of prefix combinations reported here include non-adjacent observations from triple-prefixed verbs, e.g. *ma-u-mai*. These are not included in the main prefix bigram analysis.

avoidance pattern, in that *ma-*, *mai-* 1NS.P further shows an unexpected tendency to reposition to an alternative hosting site. In the main paper we discuss prefix hosting only with respect the main stem. However there is another option, though one that is very rarely used in the corpus: prefixes may also be hosted on a subsequent verbal stem, which we refer to as a ‘V2’.

A brief detour into V2 prefix hosting helps us to contextualise the unexpected scarcity of *ma-*, *mai-* 1NS.P on multiply prefixed main stems. V2 prefix hosting in general occurs very infrequently ($N=132$), relative to the number of verbs in the corpus that have both V2 and at least one prefix ($N=2,858$). Within the minority where the V2 hosts the prefix, the most striking fact is that one particular V2, *-mett CAUS*, accounts for a large majority of the V2 hosting instances (109/132) (cf. Schikowski 2014: 77). There are about a dozen V2s occurring in this set, but all others have at most a handful of instances as prefix hosts. Aside from the dominance of *-mett CAUS* in V2 prefix hosting, the other notable pattern is that *ma-*, *mai-* 1NS.P are the prefixes most likely to be hosted on a V2. Figure 1 shows that for most prefixes, only a tiny proportion of instances on verbs with V2s attach to the V2. For *ma-*, *mai-* 1NS.P, V2 attachment is still in the minority at 16%, but it accounts for a significantly higher proportion of prefix hosting ($\chi^2 p < 0.001$). This is in stark contrast to *mai-* NEG, which has the very marginal V2 hosting typical of most prefixes.

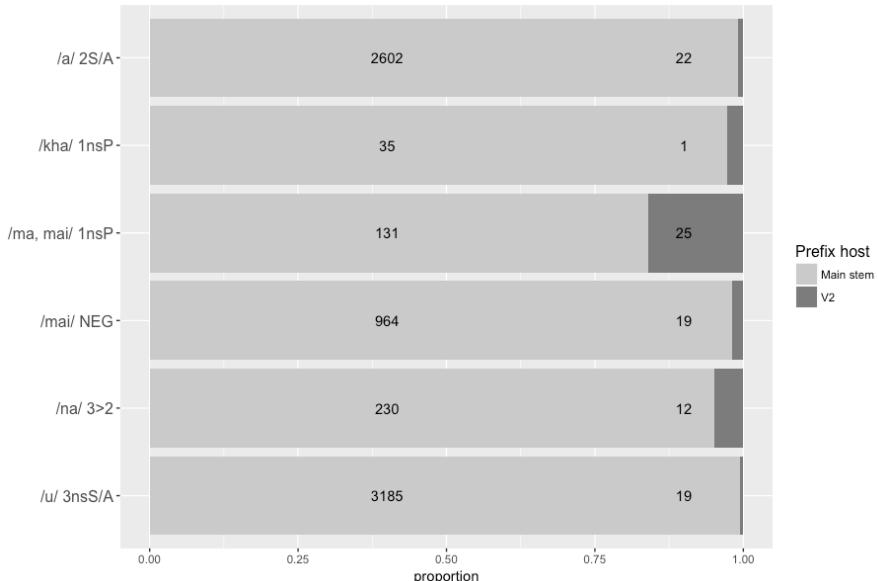


Figure 1. Hosting of prefixes on main stem or V2.

The pattern in V2 prefix hosting appears to be another dimension to probabilistic homophony avoidance. Though any prefix can be hosted on either main stem or V2, speakers show markedly different preferences in their positioning of the homophonous prefixes. While neither this nor the quasi-dialectal selection of *kha-* definitively resolve the homophony problem, they make it less likely to occur. The

homophony avoidance effects on *ma-*, *mai-* 1NS.P show how the principle of category clustering can be violated in an otherwise orderly system, due to an accident of linguistic history.

References

- Baerman, Matthew. 2011. Defectiveness and homophony avoidance. *Journal of Linguistics* 47. 1–29.
- Bickel, Balthasar & Gaenszle, Martin. 2015. First person objects, antipassives, and the political history of the Southern Kirant. *Journal of South Asian Languages and Linguistics* 2(1). 63–86.
- Kaplan, Abby & Muratani, Yuka. 2015. Categorical and gradient homophony avoidance: Evidence from Japanese. *Laboratory Phonology* 6(2). 167–195.
- Schikowski, Robert. 2014. Chintang sketch grammar. Unpublished ms.