Applications of classification trees for endangered language description: Finite verb morphology in Kolyma Yukaghir

Albert Ventayol-Boada
University of California, Santa Barbara
aventayolboada@ucsb.edu

Abstract

This study investigates the use of the finite verb-focus forms *-jə* and *-mə* in Kolyma Yukaghir. These forms are described in terms of valency: *-jə* is the intransitive form, and *-mə* is the transitive. In this study, I revise this analysis by using decision tree modeling on five monologic texts annotated for different discourse and morphological factors. The results show that 1) valency alone cannot account for the distribution of the finite verb-focus forms in discourse, and 2) speakers are sensitive to different discourse factors and use *-jə* and *-mə* to achieve different communicative goals. In short, this study demonstrates how machine learning methods like classification decision trees can offer a more nuanced picture of the choices speakers make at the discourse level and help the language description process.

1 Introduction

Finite verb morphology in the endangered language of Kolyma Yukaghir (Yukaghiric; Russia) has a four-way distinction from which language users choose when making an assertion (Krejnovich, 1982; Maslova, 1997, 2003, 2008; Nagasaki, 2010, 2018). These four forms are said to represent the grammaticalization of the "information structure status" of core participants (Nikolaeva, 2005) and are classified according to valency (i.e., intransitive vs. transitive) and as to whether they highlight the event (i.e., verb-focus) or the event participants (i.e., subject-focus or object-focus). This description, however, fails to account for a significant number of utterances in spontaneous discourse, since examples of verbs with the 'intransitive' verb-focus form *-jə* are attested with two participants (1), and the 'transitive' verb-focus *-mə* also occurs with a single participant (2).

(1) *ополльэ мэткэлэ моннуӈи́й,*
opoll'ə met-kələ mon-nu-ŋi-**j**,
later 1SG-ACC say-IPFV-PL-**Jə**.3
'Later they told me' ("Tobacco," 34:13)

(2) *иӈэртэ**м*** *иӈэртэ**м***,
iŋer-tə-**m** iŋer-tə-**m**,
hole-?-**mə**.3SG hole-?-mə.3SG
'He dug, he dug' ("The felt boots," 13:21)

An alternative view is to consider the choice over the four finite forms as carrying out different discourse functions (Nagasaki, 2018), and thus being sensitive to different discourse and pragmatic factors. In this view, explanations for speakers' choices can be found in a more dynamic view of transitivity in which events do not have an inherent amount of participants (Hopper and Thompson, 1980; Diver et al., 2012), as well as the accessibility of the referents (Ariel, 2001), and the different degrees of attention-worthiness placed onto the participants and the event (Diver and Davis, 2012).

The goal of this study is to investigate these factors in concert in order to understand what communicative goals the two verb-focus forms (i.e., 'intransitive' *-jə* and 'transitive' *-mə*) fulfill, with a study of the less common subject-focus and object-focus forms to follow. Ultimately, the objective is to improve current descriptions of Kolyma Yukaghir through a multifactorial analysis of spontaneous discourse in a context where recording additional linguistic materials is difficult due to language endangerment.

2 Data

I analyzed 5 of the 40 monologic texts collected in the late 20th century (Nikolaeva and Mayer, 2004). These five texts were narrated by the same person, but differ slightly in terms of genre: there is a personal story ("Tobacco"), a fantastical story ("Elizar"), the description of a game played between two people with their hands ("A game"), the depiction of a yearly meeting among the Yukaghirs ("Yearly meetings"), and an account of the fortune telling practice to predict the ethnic group a woman will marry into ("Fortune telling").

I first divided each text into intonation units (IUs; Chafe, 1979, 1994; Du Bois et al., 1993) and then extracted all 135 verb-focus finite forms (88 -jə's and 47 -mə's). Afterwards, I manually annotated in a table each token for linguistic features that have been shown to correlate with accessibility and attention-worthiness: number of overt participants (Huffman, 2001) and their linguistic encoding (Ariel, 2009), case-marking and co-occurrence with the finite form in the same IU (Himmelmann, 2022), the grammatical persons involved in the event (Contini-Morava, 1983), polarity (Diver, 2012), and aspect (Reid, 1976; Gorup, 1987). Below I list the predictors I used in the study and their levels.

First, the number of overt (i.e., explicitly-mentioned) participants contains three levels: 0, 1, and 2. Additionally, I created a separate variable with the total number of participants (i.e., including covert, contextual-inferred participants that are not explicitly mentioned). For example, verbs of transfer like 'give' only appear with one or two explicit participants but never three, although these are often thought of involving three participants (Haspelmath, 2004).

Second, the linguistic encoding of participants generates two predictors, one for each participant. Both include five levels: lexical, pronoun, and quantifier (for overt participants), and mentioned and implicit (for covert participants). The last two categories capture the distinction between covert participants of a finite verb appearing as overt participants of a preceding non-finite verb (but with case-marking assigned by the finite verb), and covert participants of a finite verb appearing as overt participants of a preceding finite verb.

Third, case marking on overt first participants contains two levels (i.e., predicative and no case marking), and three levels for overt second participants (i.e., accusative, predicative and no case marking). Similarly, I included another predictor for the case markings of non-core arguments (i.e., ablative, dative, instrumental, lative, locative, and prolative). Fourth, the co-occurrence of overt participants with the finite form in the same IU generates two predictors, one for each participant. Both predictors have two levels: co-occurring in the same IU and not co-occurring in the same IU.

Fifth, the grammatical persons involved in the event is a single predictor with 14 levels: five with a single participant (1SG, 1PL, 2SG, 3NPL, 3PL), seven with two participants (1SG > 3NPL, 1SG > 3PL, 3NPL > 1SG, 3NPL > 3NPL, 3NPL > 3PL, 3PL > 1SG, 3PL > 3NPL), and two with three participants (1SG > 3NPL > 3NPL, 1SG > 3NPL > 3PL). Other combinations of grammatical persons are not attested in the data.

Sixth, polarity is operationalized as the co-occurrence of the negative proclitic *əl* with the finite verb form. This predictor has two levels: co-occurring and not co-occurring.

Finally, aspect in Kolyma Yukaghir includes several categories that occupy different slots in the finite verb template and can co-occur; these are: habitual, inchoative, imperfective, iterative, noniterative, resultative, and perfective. I also annotated the data for three additional morphemes that can appear in the morphological template and co-occur with aspect: future tense, causative and evidential. Each of these categories is treated as a separate predictor with a binary choice, i.e., whether they co-occur or not with the finite verb form.

Table 1 summarizes the list of predictors and their levels. In total, the data was annotated for 21 independent variables that include a variety of morphological and discourse factors, and I used the choice of the verb-focus form (i.e., *-jə* vs. *-mə*) as the dependent variable.

3 Methods

In order to investigate speakers' choices of the verb-focus forms at the discourse level, I used a classification decision tree. Tree-based methods have gained popularity in linguistics research over the past decade (Tagliamonte and Baayen 2012; Wiechmann and Kerz 2013; Bernaisch et al. 2014; Hundt 2018, among others), but studies applied to languages other than English have mostly focused on NLP applications rather than linguistic analyses (but see Klavan et al. 2015). Tree-based modeling, however, is particularly suitable for grammatical analyses in endangered languages, as it is applicable to small-*n* large-*p* (i.e., few data points, many predictors) scenarios, and it avoids problems of collinearity (Strobl et al., 2009; Gries, 2021).

A classification decision tree is an effective method to investigate the choice of verb-focus forms. As mentioned, current descriptions of Kolyma Yukaghir describe the use of *-jə* vs. *-mə* as depending only on valency (intransitive vs. transitive). In machine learning terms, this characterization can be formulated as a decision tree with a single split based on transitivity: if intransitive, pre-

Predictor	Levels
Number of overt ppts	0, 1, 2
Number of covert ppts	0, 1, 2, 3
Encoding of 1st & 2nd ppts	lexical, pronoun, quantifier, mentioned, implied
Case marking on 1st ppt	bare, predicative
Case marking on 2nd ppt	bare, predicative, accusative
Other referents	ablative, dative, instrumental, lative, locative, prolative
Co-occurrence with 1st ppt in IU	yes, no
Co-occurrence with 2nd ppt in IU in IU	yes, no
Grammatical persons	1SG, 1PL, 2SG, 3NPL, 3PL, 1SG > 3NPL, 1SG > 3PL, 3NPL > 1SG, 3NPL > 3NPL, 3NPL > 3PL, 3PL > 1SG, 3PL > 3NPL, 1SG > 3NPL > 3NPL, 1SG > 3NPL > 3PL
Co-occurrence with negative *əl*	yes, no
Co-occurrence with habitual, imperfective, evidential...	yes, no

Table 1: Predictors and their levels used to annotate each token of *-jə* and *-mə*

dict *-jə*; if not intransitive, predict *-mə*. Thus, any alternative decision tree configuration from the supervised model (i.e., either with more leaves or a single split with a different predictor) would suggest that speakers are sensitive to the morphological and discourse factors listed above.

4 Findings

Due to the imbalanced distribution of the two forms, the baseline/no-information rate accuracy of the classification model is already at 65.2%. Rather than a training-testing split, a leave-one-out cross-validation method was used instead, given that some predictors (e.g., grammatical persons) had too few observations for some levels to make predictions with a testing set. The model performs with an 81.5% true prediction accuracy. A test-is-training model, however, outperforms the cross-validation model with 89.6% accuracy ($p_{binomial\ test}$ = 0.006). Figure 1 shows the classification tree from the test-is-training model.

The results show that the choice of verb-focus forms is most sensitive to the grammatical persons involved in the event. Two-participant events with two third persons are favored by the form *-mə* (3), whereas single-participant events and two-participant events with a speech-act participant (i.e., first-person or second-person) are favored by *-jə* (4). Examples (3) and (4) both display a two-participant event with an implicit first participant and a mentioned second participant (i.e., 'pipe' and 'strap' appear as overt participants of a preceding non-finite verb), but they differ in the grammatical persons involved: (3) only involves third persons, whereas (4) involves a speech-act participant.

(3) *табаах пиэдэттэллэ хаŋсаа*
tabaaq peedə-t-təllə qaŋsaa
tobacco burn-?-CVB.SEQ pipe
*пиэдэттэллэ ооҕжаанунну**ŋaa**.*
peedə-t-təllə oož-aa-nun-nu-**ŋaa**.
burn-?-CVB.SEQ drink-INCH-HAB-IMPF-**MƏ.3PL**
'After kindling the tobacco, after kindling the pipe, they used to smoke (it)' ("Tobacco," 34:8)

(4) *льамкапки лончшиллэ*
l'amka-p-ki lon-čii-llə
strap-PL-3POSS take.down-ITER-CVB.SEQ
*иркилльэŋоон ултэ**сь**.*
irkill'ə-ŋoo-n ultə-**s'**.
together-COP-LNK tie-**JƏ.1SG**
'After taking down their straps, I tied (them) together' ("Tobacco," 34:34)

The exceptions to this pattern are single-participant events with a first-person plural, and three-participant events with a third-person non-plural recipient. These, however, only have 3 observations each, so the algorithm might be picking up on idiosyncrasies of these examples; in comparison, two-participant events with two third persons make up around 20% of the data—or 31 observations of the total 135.

Additionally, the results show that two-participant events with a speech-act participant are favored by *-mə* if there are two or more overt participants (5). The form *-mə* is also

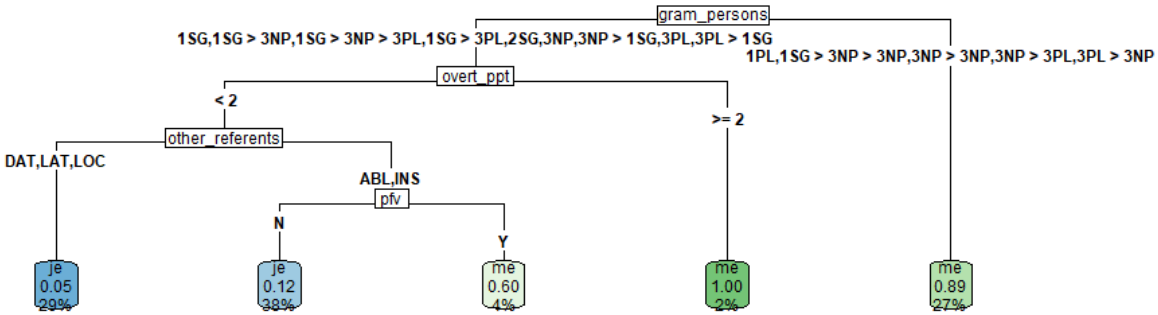

Figure 1: Classification tree from the test-is-training model

predicted for single-participant events and for two-participant events with a speech-act participant in co-occurrence with the perfective marker and an additional referent in ablative or instrumental (6).

(5) *мэт табаах өнмэгэ эйтэ,*
met tabaaq önmə-gə ej-tə-**Ø**,
1SG tobacco mind-LOC get-?-**MƏ.1SG**
'I remembered the tobacco' ("Tobacco," 34:44)

(6) *пиэдисьэ**лэ**.*
peedis'ə-**lə**.
finger-**INS**
'With (his) finger'

*льэнуль**э**лум*
l'ə-nu-l'əl-u-m
PH-IMPF-EV-(EP)-MƏ.3SG
*ньум**шэ**шнульэлу**м**.*
ɲum-**šə**-š-nu-l'əl-u-**m**.
press-**PFV**-CAUS-IMPF-EV-(EP)-**MƏ.3SG**
'He whatchamacallit, he presses' ("A game," 43:10)

5 Discussion & Conclusion

The findings suggest that valency alone cannot account for the choices speakers make between the two verb-focus forms at the discourse level, as it is argued in current descriptions of Kolyma Yukaghir. The form *-mə* is used in events that might look "transitive," but the configuration of the participants involved is relevant: *-mə* is overwhelmingly preferred with two third-person participants. These events differ from other events (i.e., single-participant events and two-participant events with a speech-act participant) in that they involve more discourse referents. Thus, a potential interpretation of this skewing is that *-mə* might be cuing speakers to a higher potential for reference tracking problems, while *-jə* might signal a lower probability of problems in reference tracking.

Marginally, *-mə* is also preferred in two additional contexts: in two-participant events with a speech-act participant when both are overtly specified (i.e., lexically or pronominally), and in perfective events with one or two participants and a referent in locative or instrumental. These configurations are in line with the idea that 1) perfective aspect is used in discourse to foreground important events (Hopper and Thompson, 1980), and 2) the higher the number of explicitly-mentioned participants, the more thematic importance of the event (Diver and Davis, 2012). As a result, *-mə* can be seen as highlighting events worthy of more attention, whereas *-jə* is used for events with lower attention-worthiness.

Overall, the results suggest that speakers choose the verb-focus forms depending on different discourse factors and may use *-jə* and *-mə* to achieve different communicative goals: cuing the addressee to a higher potential of reference problems, and highlighting important events in the discourse. In order to validate these results, a follow-up study with random forests will also be carried out. In sum, this study demonstrates how machine learning methods like decision trees can offer a more accurate picture of the choices speakers make at the discourse level and can help documentary efforts in the description process. Tree-based approaches are especially well-suited for endangered languages, as they can model linguistic input and make predictions with a relatively small number of examples.

Acknowledgements

This publication resulted (in part) from research supported by the Columbia School Linguistic Society. I want to thank Prof. Ellen Contini-Morava for her in-

valuable insight and continuous feedback from the early stages of this project. I am also thankful to Prof. Joseph Davis and Prof. Ricardo Otheguy for their comments and questions to refine the ideas presented here.

Abbreviations

1	first person	IPFV	imperfective
2	second person	ITER	iterative
3	third person	JƏ	-jə
ACC	accusative	LNK	linker
CAUS	causative	LOC	locative
COP	copula	MƏ	-mə
CVB	converb	NPL	nonplural
EP	epenthesis	PFV	perfective
EV	evidential	PH	placeholder
HAB	habitual	PL	plural
IMPF	imperfective	POSS	possessive
INCH	inchoative	SEQ	sequential
INS	instrumental	SG	singular

References

Mira Ariel. 2001. Accessibility theory: an overview. In Ted Sanders, Joost Schilperoord, and Wilbert Spooren, editors, *Text Representation: Linguistic and Psycholinguistic Aspects*, pages 29–87. John Benjamins Publishing Company, Amsterdam, Philadelphia.

Mira Ariel. 2009. Discourse, grammar, discourse. *Discourse Studies*, 11(1):5–36.

Tobias Bernaisch, Stefan Th. Gries, and Joybrato Mukherjee. 2014. The dative alternation in South Asian English(es). *English World-Wide. A Journal of Varieties of English*, 35(1):7–31.

Wallace L. Chafe. 1979. The Flow of Thought and the Flow of Language. In Talmy Givón, editor, *Discourse and Syntax*, pages 159–181. Brill, New York.

Wallace L. Chafe. 1994. *Discourse, Consciousness, and Time*. The University of Chicago Press, Chicago, London.

Ellen Contini-Morava. 1983. Ranking of Participants in Kinyarwanda: The Limitations of Arbitrariness in Language. *Anthropological Linguistics*, 25(4):425–435.

William Diver. 2012. The system of Relevance of the Homeric Verb. In Alan Huffman and Joseph Davis, editors, *Language: Communication and Human Behavior*, pages 135–159. Brill, Leiden.

William Diver and Joseph Davis. 2012. Latin voice and case. In Alan Huffman and Joseph Davis, editors, *Language: Communication and Human Behavior*, pages 194–245. Brill, Leiden.

William Diver, Joseph Davis, and Wallis Reid. 2012. Traditional Grammar and Its Legacy in Twentieth-century Linguistics. In Alan Huffman and Joseph Davis, editors, *Language: Communication and Human Behavior. The Linguistic Essays of William Diver*, pages 371–443. Brill, Leiden.

John W. Du Bois, Stephan Schuetze-Coburn, Susanna Cumming, and Danae Paolino. 1993. Outline of Discourse Transcription. In Jane A. Edwards and Martin D. Lampert, editors, *Talking data: Transcription and coding in discourse research*, pages 45–89. Lawrence Erlbaum Associates Publishers, Hillsdale.

Radmila J. Gorup. 1987. *The Semantic Organization of the Serbo-Croatian Verb*. Verlag Otto Sagner, Munich.

Stefan Th. Gries. 2021. *Statistics for linguistics with R: A practical introduction*, 3rd editio edition. De Gruyter Mouton, Berlin.

Martin Haspelmath. 2004. Explaining the Ditransitive Person-Role Constraint: A usage-based approach. *Constructions*, 2:1–71.

Nikolaus P. Himmelmann. 2022. Prosodic phrasing and the emergence of phrase structure. *Linguistics*, 60(3):715–743.

Paul J. Hopper and Sandra A. Thompson. 1980. Transitivity in Grammar and Discourse. *Language*, 56(2):251–299.

Alan Huffman. 2001. The linguistics of William Diver and the Columbia school. *Word*, 52(1):29–68.

Marianne Hundt. 2018. It is time that this (should) be studied across a broader range of Englishes. In Sandra C. Deshors, editor, *Modeling World Englishes. Assessing the interplay of emancipation and globalization of ESL varieties*, pages 217–244. John Benjamins Publishing Company, Amsterdam.

Jane Klavan, Maarja Liisa Pilvik, and Kristel Uiboaed. 2015. The use of multivariate statistical classification models for predicting constructional choice in spoken, non-standard varieties of Estonian. *SKY Journal of Linguistics*, 28(2015):187–224.

Eruxim A. Krejnovich. 1982. *Issledovanija i materialy po jukagirskomu jazyku*. Akademia Nauk SSSR, Moscow.

Elena S. Maslova. 1997. Yukagir focus in a typological perspective. *Journal of Pragmatics*, 27(4):457–475.

Elena S. Maslova. 2003. *A Grammar of Kolyma Yukaghir*. Mouton de Gruyter, Berlin, New York.

Elena S. Maslova. 2008. Case in Yukaghir languages. In Andrej L. Malchukov and Andrew Spencer, editors, *The Oxford Handbook of Case*, pages 789–796. Oxford University Press, Oxford.

Iku Nagasaki. 2010. Kolyma Yukaghir. In Yasuhiro Yamakoshi, editor, *Grammatical Sketches from the Field*. Research Institute for Languages and Cultures of Asia and Africa (ILCAA), Tokyo.

Iku Nagasaki. 2018. The Focus Construction in Early Modern Kolyma Yukaghir. *Gengo Kenkyu (Journal of the Linguistic Society of Japan)*, 154(C):123–152.

Irina Nikolaeva. 2005. Review of *A grammar of Kolyma Yukaghir* by Elena S. Maslova. *Linguistic Typology*, 9:299–325.

Irina Nikolaeva and Thomas Mayer. 2004. Online Documentation of Kolyma Yukaghir.

Wallis Reid. 1976. *The human factor in linguistic analysis: The passé simple and the Imparfait*. Phd dissertation, Columbia University.

Carolin Strobl, James Malley, and Gerhard Tutz. 2009. An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological Methods*, 14(4):323–348.

Sali A. Tagliamonte and R. Harald Baayen. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*, 24(2):135–178.

Daniel Wiechmann and Elma Kerz. 2013. The positioning of concessive adverbial clauses in English: Assessing the importance of discourse-pragmatic and processing-based constraints. *English Language and Linguistics*, 17(1):1–23.