UC Santa Barbara

UC Santa Barbara Previously Published Works

Title

Unsupervised part-of-speech induction for language description: Modeling documentation materials in Kolyma Yukaghir

Permalink

https://escholarship.org/uc/item/3rw9x3c2

Authors

Ventayol-Boada, Albert

Roll, Nathan

Todd, Simon

Publication Date

2023-12-14

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Unsupervised part-of-speech induction for language description: Modeling documentation materials in Kolyma Yukaghir

Albert Ventayol-Boada and **Nathan Roll** and **Simon Todd**
University of California, Santa Barbara
{aventayolboada, nroll, sjtodd}@ucsb.edu

Abstract

This study investigates the clustering of words into Part-of-Speech (POS) classes in Kolyma Yukaghir. In grammatical descriptions, lexical items are assigned to POS classes based on their morphological paradigms. Discursively, however, these classes share a fair amount of morphology. In this study, we turn to POS induction to evaluate if classes based on quantification of the distributions in which roots and affixes are used can be useful for language description purposes, and, if so, what those classes might be. We qualitatively compare clusters of roots and affixes based on four different definitions of their distributions. The results show that clustering is more reliable for words that typically bear more morphology. Additionally, the results suggest that the number of POS classes in Kolyma Yukaghir might be smaller than stated in current descriptions. This study thus demonstrates how unsupervised learning methods can provide insights for language description, particularly for highly inflectional languages.

1 Introduction

Many NLP applications and linguistic investigations are facilitated by having Part-of-Speech (POS) tags for words in context. Providing such tags flexibly and at scale for novel texts requires a POS tagger. When working with low-resource languages, it is often infeasibly labor-intensive to develop labeled data that would enable the training of a supervised tagger, or even to develop a lexicon that delimits the set of tags that may be appropriate for each word (Hasan and Ng, 2009) and thereby facilitates the training of an accurate unsupervised tagger (e.g. Goldwater and Griffiths, 2007). Working with such languages requires turning to POS *induction*, which clusters words according to the contexts in which they occur in an unannotated corpus. Following the *distributional hypothesis* (Harris, 1951, 1954), words that occur in similar contexts are assumed to belong to the same POS class.

POS induction is a potentially useful tool for the documentary linguist. However, its utility for endangered and underdocumented languages remains to be established, as it is not always clear what the POS class of a word should be in many such languages or even whether the notion of POS classes as established for high-resource European languages is appropriate (Bender, 2011; Finn et al., 2022), due to potentially high degrees of polyfunctionality (Mithun, 2017; Hieber, 2021; Carter, 2023). The goal of this paper is to evaluate the insights of POS induction for language documentation, through a case study on Kolyma Yukaghir (Yukaghiric), a highly inflectional endangered language of Northeastern Siberia (Republic of Sakha, Russia).

2 POS induction in highly inflectional languages

POS induction leverages distributional information by representing each word as a *co-occurrence vector*, which reflects how often it appears near each other word in a corpus, and clustering words with similar vectors. This approach is successful for languages that display fairly rigid word order and little inflection, like English, because the vectors are characterized by frequent function words that predominantly co-occur with words in certain POS classes, such as “the” and “to”. However, it is not so successful for highly inflectional languages, in which the corresponding function elements are bound morphemes (Dasgupta and Ng, 2007; Bender, 2011).

Successful unsupervised POS induction for highly inflectional languages requires building morphological information into the model. This approach leverages the fact that inflectional affixes are strongly associated with POS classes. If the POS classes of the affixes in a word are known, they can be used to delimit the set of possible POS classes for the root of that word (Hajič, 2000; Duh and Kirchhoff, 2006). If the POS classes of affixes are not known, the distributional hypothesis can be applied

at the morphological level: roots that have similar distributions of co-occurrence with affixes can be assumed to belong to the same POS class (Cucerzan and Yarowsky, 2000; Clark, 2003; Freitag, 2004; Dasgupta and Ng, 2007).

However, building morphological information into POS induction may not be successful in all languages. There may be three major issues, which we illustrate with examples from Kolyma Yukaghir.

The first issue is that the same affix may attach to roots that would traditionally be considered to have distinct POS classes, and as a result are analyzed as homonyms. In our examples, the suffix *-n* attaches to nouns when they modify a noun (1) or encode the arguments of a postposition (2), in which case it is glossed as "genitive". An identical morpheme attaches to numerals (3) and "adjectives" (4), but in these cases it is often glossed as "attributive" or "adverbializer", respectively.

(1) *одун* *мархиль,*
odu-**n** marqil',
Yukaghir-**GEN** girl
'(The) Yukaghir girl' ("Yearly meetings")

(2) *таа нумөн* *ниŋиэлгэн*
taa numö-**n** niŋeel-gən
there house-**GEN** between-PROL
эйрэт,
ej-rə-t,
walk-NONIT-CVB.CTX
'Walking along the houses there'
("Tobacco")

(3) *иркин* *йалɣилгэ йахайэ,*
irk-i-**n** jalğil-gə jaqa-jə,
one-EP-**ATTR** lake-LOC reach-1SG
'I arrived at a lake' ("Tobacco")

(4) *чомоон* *йукоодьоон* *оодьэ,*
čom-oo-**n** juk-oo-d'oon oo-d'ə,
big-RES-**ADVZ** small-RES-NMLZ be-1SG
'I was very small'
[Lit. 'I was smalling greatly'] ("Tobacco")

This issue reflects a problem with traditional considerations: labels like "genitive", "attributive" or "adverbializer" reflect a view that tries to bend Kolyma Yukaghir to ill-fitting POS classes developed for other (European) languages. It is more fitting to characterize the grammatical relations in the language's own terms (Mithun, 2001; Epps, 2011). From this perspective, examples (1–4) display a single form that attaches to a modifier to grammatically mark its relationship with the modified. That relationship may be of a more attributive nature like in (1) or (3), but it may also be of another kind, as in (2) and (4).

The second issue is that the same root may appear with affixes that are prototypically associated with traditionally distinct POS classes (Maslova, 2003). Numerals and "adjectives" appear in (3–4) with a suffix that is indistinguishable from prototypically nominal case-marking in (1–2), but they are also attested as the main predication of a clause bearing prototypical verbal morphology, such as aspect, evidentiality, person and number (5–6).

(5) *иркидьэ* *мит йаал**ооиили***
irk-i-d'ə mit jaa-l-**oo-iili**
one-EP-PTCP 1PL three-EP-**RES-1PL**
'Once we were three'
[Lit. 'Once we threed'] ("The first lesson")

(6) *киндьэ,*
kind'ə
moon
'(The) moon'

ишлэмэдэ *чом**мунульэл,***
iilə-mə-də čom-**mu-nu-l'əl-0**
other-TEMP-UNK big-**IMPF-INCH-EV-3SG**
'Sometimes the moon becomes big'
[Lit. 'Sometimes the moon bigs'] ("The first lesson")

The third issue, which is a consequence of the first two, is that two roots may have highly similar affix co-occurrence distributions but nevertheless be considered as having distinct POS classes. Despite the similarities between numerals and "adjectives" in the examples (3-6), they are treated differently in grammars: "adjectives" are grouped with verbs (Krejnovič, 1982; Maslova, 2003; Nagasaki, 2010), while numerals are either considered as a separate POS class (Maslova, 2003) or classified simultaneously with adnominals and verbs (Nagasaki, 2010). These differences in conceptualization result in a different number of POS classes: 8 according to Maslova (2003), and 6 to Nagasaki (2010).

The large degree of shared morphology across roots in such highly inflectional languages raises the question of whether applying the distributional hypothesis at the morphological level is appropriate, as well as the question of what the relevant POS classes might be in such languages in the first place. We explore these questions from a bottom-up, data-driven approach, with a case study on Kolyma Yukaghir. Specifically, we seek to identify and evaluate the number of POS clusters through unsupervised induction, without specifying a predetermined value.

3 Kolyma Yukaghir

Like other languages in the Siberian linguistic area (Anderson, 2006; Pakendorf, 2010), Kolyma Yukaghir is strongly head-final, and it displays SV/AOV constituent order with nominative-accusative alignment. Morphologically, Kolyma Yukaghir is a predominantly agglutinating, suffix-dominant language, with partially fusional morphology. Suffixes display some allomorphy due to residual vowel harmony and consonantal assimilation processes (Krejnovič, 1982; Maslova, 2003; Nagasaki, 2010).

In terms of morphological complexity, roots show differences in terms of the number and range of affixes they typically occur with. Roots used "verbally" (i.e., for predication) have the largest number of affixal slots, some of which can be filled by a wide range of possible items (e.g., aspect). Roots used "nominally" have fewer slots, which can typically be filled by fewer possible affixes, and sometimes occur without affixes at all (e.g., *kind'ə* in 6).

In this study, we analyzed 19 of the 40 monologic texts collected in the late 20th century (Nikolaeva and Mayer, 2004). These texts were narrated by five different speakers in the community and include a variety of genres: folktales, personal and fantastical stories, descriptions of games and competitions, an account of fortune telling, etc.

To prepare the data, we stripped the texts of glosses, transliterated them into Cyrillic orthography, and divided them into intonation units (IU; Chafe, 1979, 1992). IUs are defined as "a stretch of speech uttered under a single coherent intonation contour" (Du Bois et al., 1993:47) or the "spurts of language" in which speakers typically produce speech (Chafe, 1994:29). Affix boundary markers from the original transcriptions were maintained, so the choice of writing system did not impact the results. However, we removed root-internal boundary markers in compounds (13 words total). We also removed clitic boundary markers, replacing them with white space in the case of proclitics, and affix boundary markers in the case of enclitics. This treatment yielded texts that follow established written conventions as closely as possible. Additionally, it meant that every word presented the same structure: if one or more morphological boundaries were present, the left-most morpheme was the root, and any subsequent elements were suffixes.

After preprocessing, the data contained 3,513 word tokens (where a token was taken to be anything bounded by white space). These word tokens contained 3,513 root tokens of 663 types and 3,911 affix tokens of 138 types.

Definition	**Example**
ROOT(ROOTS; IU)	irk-*i-d'ə* mit **jaa**-*l-oo-iili*
ROOT(AFFIXES; WORD)	*irk-i-d'ə mit* **jaa**-l-oo-iili
AFFIX(ROOTS; WORD)	*irk-i-d'ə mit* jaa-*l-oo*-**iili**
AFFIX(AFFIXES; WORD)	*irk-i-d'ə mit* *jaa*-l-oo-**iili**

Table 1: Examples of co-occurrence vector definitions, based on (5). The vector for the target (bold) includes the counts of each underlined item in the box, summed across all occurrences of the target in the corpus.

4 Methods

We obtained co-occurrence vectors for roots and affixes under four distinct definitions. The first definition, ROOT(ROOTS; IU), yielded a vector for each root, based on the roots it co-occurs with in an IU, as shown in the first row of Table 1 based on example (5). We constructed a sparse matrix that counted how often each root in the corpus occurred in the same IU as each other root, as well as how often it occurred alone within an IU.

We removed rows corresponding to roots that only ever occurred alone within an IU, then applied truncated SVD to obtain a dense matrix with 40 columns, from which we extracted the rows. We obtained co-occurrence vectors by normalizing these rows to have unit length.

We obtained vectors similarly for the remaining three definitions: ROOT(AFFIXES; WORD) yielded a vector for each root, based on the affixes that attach to it; AFFIX(ROOTS; WORD) yielded a vector for each affix, based on the roots it attaches to; and AFFIX(AFFIXES; WORD) yielded a vector for each affix, based on the affixes it co-occurs with in a word. Examples of these definitions are shown in Table 1. For ROOT(AFFIXES; WORD) and AFFIX(AFFIXES; WORD), our vectors also included counts for the number of times a root occurred without any attached affixes and the number of times an affix was the only affix attached to a word, respectively.

For each definition, we first removed elements that only ever occurred as isolates in the corpus (for ROOT(ROOTS; IU), roots that only ever occurred alone in an IU; for ROOT(AFFIXES; WORD), roots that only ever occurred without affixes; and for AFFIX(AFFIXES; WORD), affixes that never occurred alongside other affixes in a word). We then used k-means clus-

tering on the vectors of remaining elements under each definition to induce classes of roots/affixes that have similar distributions within the corpus. We picked the number of clusters using the elbow method, where cluster quality was measured by inertia. For qualitative interpretation, we identified the 20 roots/affixes with the highest degree of centrality from each cluster.

These four definitions represent different ways to approach POS induction. ROOT(ROOTS; IU) and ROOT(AFFIXES; WORD) assign each root to a class, as is typical for POS in European languages. We expect ROOT(AFFIXES; WORD) to be better than ROOT(ROOTS; IU) for Kolyma Yukaghir because it incorporates crucial morphological information; however, we do not expect it to be particularly useful, due to the large degree of shared morphology across roots. AFFIX(ROOTS; WORD) and AFFIX(AFFIXES; WORD) assign each affix to a class, which allows the POS of a root to be determined in context by the affixes that are attached to it. We expect these definitions to be more useful than the root-wise ones because they reflect the polyfunctional nature of the language. Given the potential that affixes may mark different functional roles in Kolyma Yukaghir than is typically assumed for European languages, and may therefore co-occur with each other broadly, we might expect AFFIX(AFFIXES; WORD) to be less useful than AFFIX(ROOTS; WORD); however, the utility of AFFIX(ROOTS; WORD) ultimately depends on the extent to which roots have (gradient) prototypical associations with traditional POS roles.

5 Results

As shown in Figure 1, the elbow method identified 2 clusters (of non-isolates) for 3 definitions, and 3 clusters for ROOT(AFFIXES; WORD). Figure 2 visualizes the clusters under each definition using t-SNE.

The qualitative analysis of the 20 words with the highest degree of centrality under ROOT(ROOTS; IU) shows a lot of variability. Words closest to the center in the small cluster ($n = 55$) include Russian adverb loanwords (e.g., 'later'), pronouns (e.g., 'y'all', 'who'), nouns (e.g., 'hoof'), verbs (e.g., 'blow') and "adjectives" (e.g., 'fast'). Similarly, the big cluster ($n = 192$) does not display a clear thread; we find the same categories as above.

The clusters under ROOT(AFFIXES; WORD) show more consistency, as expected. The 20 words in the smallest cluster ($n = 116$) are almost exclusively nominal roots (e.g., 'river'), with the exception of

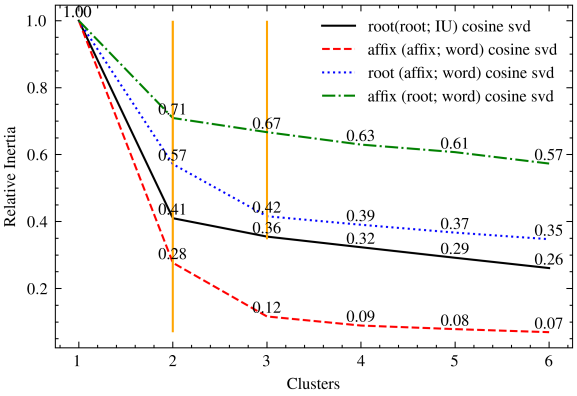
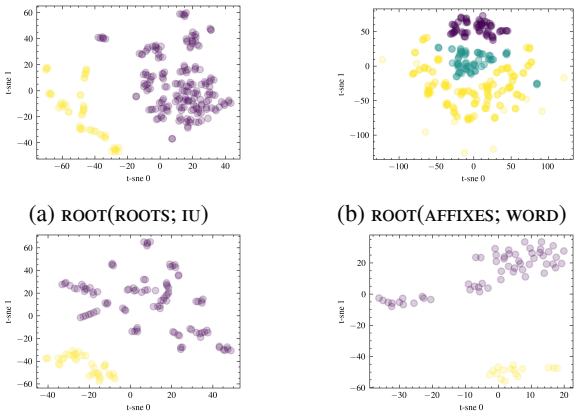


Figure 1: Number of clusters identified by the elbow method



(a) ROOT(ROOTS; IU)

(b) ROOT(AFFIXES; WORD)

(c) AFFIX(ROOTS; WORD)

(d) AFFIX(AFFIXES; WORD)

Figure 2: *k*-means clusters under the 4 definitions

the two copulas (of which one also functions as a placeholder), one verbal root (e.g., '(be) outside') and a homonym (e.g., *аӣии-* 'shoot', 'only'). We find the opposite pattern in the slightly bigger cluster ($n = 123$), where all the roots are more verbal in nature (e.g., 'hear') but one (e.g., 'bell'). The third and biggest cluster ($n = 249$) displays some variability; we find nominal (e.g., 'old woman') and verbal roots (e.g., 'take'), along with "adjectives" (e.g., 'good'), pronouns (e.g., 'what'), and nouns that can function as postpositions (e.g., 'back').

As for the clusters of affixes, AFFIX(AFFIXES; WORD) yields similar behavior to ROOT(ROOTS; IU) with a very asymmetric split. All the affixes in the small cluster ($n = 16$) are verbal, and the 20 affixes returned for the big cluster ($n = 56$) are also predominantly verbal, with the exception of a plural and a genitive/attributive allomorph.

The results for AFFIX(ROOTS; WORD) are more insightful, as expected. All but 3 of the 20 affixes in the big cluster ($n = 102$) mark verbal functions

(e.g., inchoative); the exceptions are two case markers and the directional -ŋуɔ. In the small cluster ($n = 36$), two thirds of the affixes returned were nominal (e.g. 3rd person possessive), whereas the remaining third were verbal and predominantly associated with non-finiteness.

6 Discussion & Conclusion

The number of clusters identified by the elbow method is rather small. This could be because there is not enough data to make finer distinctions in the clustering process, beyond a coarse split into prototypically "nominal" and "verbal" POS classes (and a third mixed class in ROOT(AFFIXES; WORD)). Alternatively, it could be because Kolyma Yukaghir permits a given affix (or, to some extent, root) to be used in myriad ways, such that the treatment of each root/affix as monolithic (in terms of representing one feature in the vectors, and in terms of having only one POS class) obscures deeper complexity.

As for the qualitative analysis of the clusters, the results suggest that ROOT(AFFIXES; WORD) indeed offers a more informative clustering than ROOT(ROOTS; IU). The latter definition fails to find structure in the data, whereas the former returns two cohesive clusters (with a nominal and a verbal tendency) and a third cluster with some variability. This variability, however, reflects in part the polyfunctional nature of the language. Some roots that look prototypically nominal, like 'old woman', can bear verbal morphology to convey predicative possession (i.e., 'have a wife'), and thus their clustering with "adjectives", like 'good', that can also be marked with nominal and verbal morphology is coherent with their distributions. Overall, the smaller number of function words makes the incorporation of morphological information particularly useful as anticipated.

As for affixes, the clustering under AFFIX(AFFIXES; WORD) is less useful than that under AFFIX(ROOTS; WORD). These results suggest that, to a certain degree, some roots might be prototypically associated with noun and verb POS roles. In addition, the homogeneity of the bigger cluster in AFFIX(ROOTS; WORD) with verbal functions indicates that verbal affixes might be a more reliable source of information. This probably results from finite, assertion-making words being more morphologically complex: a "verbal" root can carry several affixes simultaneously – marking it for tense, aspect, evidentiality, and person/number – whereas "nominal" roots tend not to carry many affixes at once. Nominal stems can be marked for possession, case, and evaluatives, but rarely do all co-occur. Thus, our removal of isolates – affixes that never occurred alongside other affixes in a word – is likely to have affected nominal affixes more than verbal affixes.

Taken together, the results suggest that applying the distributional hypothesis at the morphological level in a context with significant shared morphology can yield successful results, especially when clustering roots and affixes each on the basis of the other. Clustering might be more reliable for words that typically bear more morphology. However, the results can be fairly coarse-grained; to obtain finer-grained insights, more data and/or a more complex (mixture-based) approach may be necessary.

Additionally, the results also provide some insight into what the relevant POS classes in Kolyma Yukaghir might be. Rather than the eight and six POS classes listed in grammatical descriptions (Maslova, 2003 and Nagasaki, 2010, respectively), the clustering suggests a binary split at the morphological level centering around nominal and verbal functions, with the possibility of a third mixed class. Further research is needed to investigate the degree to which this third distinction is categorical or represents a cline with nouns and verbs on opposite ends.

Limitations

An important aspect of this study is the use of spoken data for the analysis, which might have had some effect on the results for ROOT(ROOTS; IU). The average IU length is 2.06 words, which effectively removes one neighbor for this definition.

Similarly, it is possible the different text genres may present different frequencies of words and constructions, which would influence the distributions underpinning POS induction. Addressing the effect of genre for POS induction is beyond the scope of this paper and remains an issue for future research.

In addition, we used morphologically segmented data rather than unsegmented data, which other POS induction studies use. Using morphologically segmented words requires some pre-existing knowledge and understanding of word structure and morphological paradigms in the language.

Finally, we treated all suffixation equally, since signs of derivation are not always clear. For highly inflectional languages with productive derivation, our approach might need a different operationalization of distributional information.

Ethics Statement

This study stems from a wider project to collect various documentation materials for Kolyma Yukaghir, and its close relative Tundra Yukaghir, and standardize them in the practical orthographies to make them more accessible to community members. With these materials, different studies are being carried out using machine learning methods in order to deepen our understanding of the grammatical structure of the languages. Ultimately, the goal is to use this knowledge to support language revitalization initiatives under way in the community.

Additionally, in this article we refrain from engaging in a "numbers game" to characterize the context of language endangerment in the Yukaghir community, as numbers are not well equipped to describe, explain or contextualize the factors that cause processes of language shift (Dobrin et al., 2009; Moore et al., 2010; Davis, 2017).

Abbreviations

1	first person	LOC	locative
3	third person	NMLZ	nominalizer
ADVZ	adverbializer	NONIT	noniterative
ATTR	attributive	PL	plural
CTX	contextual	PROL	prolative
CVB	converb	PTCP	participle
EP	epenthesis	RES	resultative
EV	evidential	SG	singular
GEN	genitive	TEMP	temporal
IMPF	imperfective	UNK	unknown/unclear
INCH	inchoative		

References

Gregory D.S. Anderson. 2006. Towards a typology of the Siberian linguistic area. In Yaron Matras, April McMahon, and Nigel Vincent, editors, *Linguistic Areas: Convergence in Historical and Typological Perspective*, pages 266–300. Palgrave Macmillan, London.

Emily M. Bender. 2011. On Achieving and Evaluating Language-Independence in NLP. *Linguistic Issues in Language Technology*, 6(3):1–26.

Matthew Carter. 2023. Polyfunctional argument markers in Ket: Implicative structure within the word. *Morphology*.

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. 1992. Intonation Units and prominences in English natural discourse. In *Proceedings of the IRCS Workshop on Prosody in Natural Speech*, pages 41–52, Philadelphia. University of Pennsylvania Press.

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

Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 59–66, Budapest, Hungary. Association for Computational Linguistics.

Silviu Cucerzan and David Yarowsky. 2000. Language Independent, Minimally Supervised Induction of Lexical Probabilities. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 270–277, Hong Kong. Association for Computational Linguistics.

Sajib Dasgupta and Vincent Ng. 2007. Unsupervised part-of-speech acquisition for resource-scarce languages. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 218–227, Prague, Czech Republic. Association for Computational Linguistics.

Jenny L. Davis. 2017. Resisting rhetorics of language endangerment: Reclamation through Indigenous language survivance. *Language Documentation and Description*, 14:37–58.

Lise M. Dobrin, Peter K. Austin, and David Nathan. 2009. Dying to be counted: the commodification of endangered languages in documentary linguistics. *Language Documentation and Description*, 6:37–52.

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.

Kevin Duh and Katrin Kirchhoff. 2006. Lexicon acquisition for dialectal Arabic using transductive learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 399–407, Sydney, Australia. Association for Computational Linguistics.

Patience Epps. 2011. Linguistic Typology and Language Documentation. In Jae Jung Song, editor, *The Oxford Handbook of Linguistic Typology*, pages 634–649. Oxford University Press, Oxford.

Aoife Finn, Peter-Lucas Jones, Keoni Mahelona, Suzanne Duncan, and Gianna Leoni. 2022. Developing a part-of-speech tagger for te reo Māori. In

Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 93–98, Dublin, Ireland. Association for Computational Linguistics.

Dayne Freitag. 2004. Toward Unsupervised Whole-Corpus Tagging. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 357–363, Morristown, United States. Association for Computational Linguistics.

Sharon Goldwater and Thomas L. Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 744–751, Prague, Czech Republic. Association for Computational Linguistics.

Jan Hajič. 2000. Morphological tagging: data vs. dictionaries. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 94–101, Seattle, United States. Association for Computational Linguistics.

Zellig S. Harris. 1951. *Structural linguistics*. The University of Chicago Press, Chicago, London.

Zellig S. Harris. 1954. Distributional Structure. *Word*, 10(2-3):146–162.

Kazi Saidul Hasan and Vincent Ng. 2009. Weakly supervised part-of-speech tagging for morphologically-rich, resource-scarce languages. In *Proceedings of the 2th Conference of the European Chapter of the Association for Computational Linguistics*, pages 363–371, Athens, Greece. Association for Computational Linguistics.

Daniel W. Hieber. 2021. *Lexical polyfunctionality in discourse: A quantitative corpus-based approach*. Phd dissertation, University of California, Santa Barbara.

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

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

Marianne Mithun. 2001. Who shapes the record: the speaker and the linguist. In Paul Newman and Martha Ratliff, editors, *Linguistic fieldwork*, pages 34–54. Cambridge University Press, Cambridge.

Marianne Mithun. 2017. Polycategoriality and zero derivation: Insights from Central Alaskan Yup'ik Eskimo. In Valentina Vapnarsky and Edy Veneziano, editors, *Lexical Polycategoriality. Cross-linguistic, cross-theoretical and language acquisition approaches*, pages 155–174. John Benjamins Publishing Company, Philadelphia.

Robert E. Moore, Sari Pietikäinen, and Jan Blommaert. 2010. Counting the losses: Numbers as the language of language endangerment. *Sociolinguistic Studies*, 4(1):1–26.

Iku Nagasaki. 2010. Kolyma Yukaghir. In Yasuhiro Yamakoshi, editor, *Grammatical Sketches from the Field*, pages 213–256. Research Institute for Languages and Cultures of Asia and Africa (ILCAA), Tokyo.

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

Brigitte Pakendorf. 2010. Contact and Siberian languages. In Raymond Hickey, editor, *The Handbook of Language Contact*, pages 714–737. Wiley-Blackwell, Oxford.