Investigating the productivity of Passamaquoddy medials: A computational approach

James Cooper Roberts
Massachusetts Institute of Technology
77 Massachusetts Ave
Cambridge, MA 02139
jcrobert@mit.edu

Abstract

Medials are a class of morphemes in the language Passamaquoddy that are involved in the construction of verbs. Members of this class have an unknown level of productivity. In this work, I investigate the matter by generating a comprehensive list of possible verb and compare it against a text corpus. Given the amount of time and energy traditional fieldwork/lexical decision tasks require, this methodology is adavntageous, particularly for Algonquianists working on similar topics. I ultimately find that 679 of approximately 15 million possible verbs are attested in said corpus. The distribution of medial type frequencies would suggest that a handful of medials are productive (under some metrics for productivity) while many medials are not. It is my hope that this data will inform future fieldwork research on the topic.

1 Introduction

This work is part of a larger research effort on a certain set of morphemes in the endangered language Passamaquoddy (Algonquian, ISO: pqm). In this language, three distinct classes of morphemes are involved in the construction of verb stems. In the literature, these classes are referred to as *initial*, *medial*, and *final* (Bloomfield, 1946; Goddard, 1990).

(1) ∅−[$_{stem}$ ut- ek- som] -on
3- [from sheet cut] -N
. [**initial medial final**] .
's/he cuts a slice from it'

[1]The focus of the current study lies with the second of these. Medials in Passamaquoddy and the Algonquian languages in general are linguistically interesting for a number of reasons; some recent works have investigated the (morpho)syntax and semantics of this morphological class (Brittain, 2003; Quinn, 2009; Branigan et al., 2005; Biedny et al., 2021; Whitney et al., 2022; Slavin, 2012), but several questions remain. Crucially, there is the issue of *productivity*—i.e., does the grammar of Passamaquoddy allow speakers to create new verbs with these medials? Claims of productive verb stem construction are sprinkled throughout the literature on Algonquian; Bakker (1997), for example, claims that it is possible "make new [verb stem] combinations [with initials/medials/finals] productively" in Cree. However, he also reports that Cree speakers are unable to "ascribe meanings" to them, which casts some doubt on his initial claim. In more recent work, Mazzoli (2023) argues that some finals can combine with a root in predictable ways (and are therefore productive), while other finals can not.

As far as I am aware, though, there has been no systematic work investigating medial productivity. The answer to this question are consequential not only to our understanding of Passamaquoddy, but also to those interested in the language's revitalization. If verb stem construction is a regular process, second language instructors should impart this knowledge to learners. However, going about a survey on the matter is not a trivial effort, considering that the current estimate on the number of medials in Passamaquoddy is 97. Evaluating productivity with a lexical decision task would require a linguist to run through thousands of possible verbs with a native speaker, which can be time-consuming for both the researcher and the consultant. In any effort to document a language of Passamaquoddy's vitality, time is of the essence. Can this research be streamlined?

In this project, I automate this effort by first generating a list of all possible medial-containing verb stems. I then determine which of them are attested by comparing said list against a text corpus. In 2, I

[1]In this article, Passamaquoddy words are orthographically represented with the Newell-Hale Alphabet. This writing system is largely phonemic, save for some characters. For more details on the orthography and the language's phonology, I direct the reader to Grishin (2023) for an overview.

Proceedings of the The Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 13–20

situate this work by providing some details on the Passamquoddy language and its speakers. In 3, I elaborate on the morphophonology of verb stems. This is followed by 4, where I discuss the procedure employed in generating the possible verb stems and running them against the dictionary. I provide a statistical analysis for the data in 5, and conclude with a discussion of my findings and plans for future work in 6.

2 Language background

Passamaquoddy is a member of the Northeastern branch of Eastern Algonquian subfamily, which makes it a close relative of Mi'kmaq, Penobscot, and Abenaki (Oxford). There are two mutually-intelligible dialects of this language, namely Passamaquoddy and Maliseet (an exonym for Wolastoqey, alternatively spelled *Malecite*). The former of these is spoken in Eastern Maine, USA, while the latter is spoken in New Brunswick, Canada (Grishin, 2023). Given their similarities, the language is also referred to as *Passamaquoddy-Maliseet* or *Passamaquoddy-Wolastoqey*. While there are some differences between the two, none of them are consequential to this work as far as I am aware. For the sake of brevity, I refer to the language as simply Passamaquoddy in this article.

Like other Algonquian languages, Passamaquoddy is polysynthetic, which is reflected especially clearly in the domain of verbs. Interestingly, verbs do a large portion of the semantic legwork in the language; based on some preliminary investigations, there are no adjectives proper, no singular "group nouns" (e.g., *team*, *family*) (Peter Grishin, p.c.), and no non-verb measure functions (e.g., *height*, *weight*, *cost*). In Passamaquoddy, equivalent expressions are handled by verbs, which are subject to the tripartite structure introduced in 1. Other features include complex agreement patterns, proximate/obviative marking for third person noun phrases, animate/inanimate distinctions for nouns, and free word order (Grishin, 2023).

Recent estimates claim that there are approximately 500 native speakers of the language (Lewis and Fennig, 2016), the majority of which are over the age of sixty (Crockett-Current, 2020). The risk of language dormancy has spurred the creation of language courses at universities (Crockett-Current, 2020) and grade schools. Detailed knowledge of Passamaquoddy grammar is crucial for such programs, and could aid in non-native speakers achieving native-like fluency in the language.[2]

3 More on verb stems

In this section, I present some facts about Passamaquoddy verb stems and medials that are relevant to this research and/or may be of interest to the reader. Most of the following is based off of LeSourd (1988), Bloomfield (1946) and Goddard (1990).

To begin, some verbs do not include medials, such as that in (2). In Passamaquoddy, a well-formed verb stem consists of at minimum a final. Ergo, some verbs do not include an initial or medial. Note that the existence of a verb stem with a medial does not necessarily entail the existence of a similar verb stem without a medial. For example, (3) is an attested word in the dictionary (see 4), but a corresponding medial-less form *aluwahke* is unattested.

(2)
pehki-	kon
clean	be$_{II}$
initial	final

'it is clean'

(3)
aluw-	al-	ke
aluw	al	ahke
initial	medial	final

's/he goes around causing trouble'

Furthermore, many medials appear noun-like in their translations into other languages, such as *atpe* 'head' in (4). However, the syntax and morphology of Passamaquoddy do not treat these morphemes as though they were nouns.[3] As a result, predicates that may require a transitive verb with an object in other languages can be expressed with a single intransitive verb in Passamaquoddy.[4]

[2] A reviewer wonders whether this specific work presented in this article was requested by the Passamaquoddy and Maliseet communities. There is a desire to produce more speakers of the language, and many members of the community are involved in the aforementioned pedagogical efforts. While this research into medial productivity is beneficial to language pedagogy for the reasons mentioned in 1, this work was not directly requested.

[3] It is worth mentioning that medials often bear no obvious resemblance to their respective noun. For example, the medial translated as 'head' is *atpe*, but the independent noun for head is *woniyakon*.

[4] The argument structure of a verb (specifically, the transitivity of the verb and the animacy of one of its arguments) is determined by the final. The subscripts on the finals in (2), (4), and (5) stand for *inanimate intransitive*, *animate intransitive*, and *transitive inanimate*, respectively.

(4) mask- -atpe -mahsu
smelly head smell$_{AI}$
initial medial final
's/he has a smelly head/hair'

Medials can not be consistently interpreted as the theme of their respective verb. In (5) for example, *ocok* has an arguably instrumental interpretation.

(5) kopp- ocok- ahm
close mushy by.tool$_{TI}$
's/he seals it with a sticky substance'

There are a number of (semi-regular) phonological alternations that occur within verb stems. In the remainder of this section, I will briefly summarize these alternations; specifically, I present those that have an overt effect on a verb stem's orthographic representation, as these are most relevant to the current study.

If the synthesis of a word stem creates a consonant cluster at a morpheme boundary, an epenthetic [i] is inserted between the two consonants.

(6) kin naqot
initial final
kininaqot 'it looks big'

Some vowels (*o* in particular) will drop at morpheme boundaries in certain environments. If *o* would otherwise be the first segment of the verb stem, it is dropped. It will also drop to resolve vowel hiatus (7) or to bring two identical sonorants together (8). It is also lost before obstruents (9).

(7) mace olan
initial final
macelan 'it starts raining'

(8) otol olan
initial final
tollan 'it is raining'

(9) kin okil
initial final
kinkil 's/he is big'

[i] and [a] will occasionally drop if they are preceded by a consonant and followed by an hC cluster.

(10) con ahte
initial final
conte 'it is stoped in place'

If an hC cluster ends up before a consonant, [h] is dropped.

(11) ehq ihtahal
initial final
eqtahal 's/he stops sitting'

If a [t] is preceded by an [i] at a morpheme boundary, it undergoes palatilization. This includes the aforementioned epenthetic [i]. Some morphemes such as *essi* and *eyi* "lexically" palatilize a preceding [t] despite not starting with [i].

(12) wikuwat eyi
initial final
wikuwaceyu 'it is fun'

An underlying /i/ is realized as [u] when it is word final.

(13) ahq al omi
initial medial final
ahqalomu 's/he is shy'

In the case of vowel hiatus, a glide (or sometimes [h]) is inserted. The complete list of alternations is presented in Table 1.

	first vowel				
		a	e	i	u
second vowel	a	aya	iya	iya	uwa
	e	aye	iye	iye	uwe
	i	ayi	ihi		uwi
	o	a	e	i	u
	u	ayu	iyu		uwu

Table 1: Vowel hiatus repairs

4 Procedure

Using a Python script, I begin by generating a comprehensive list of medial-bearing possible verb stems. This is produced by exhaustively combining every initial, medial, and final from a machine-readable list.[5] For each verb stem, the phonological rules sketched in 3 are applied. These are modeled computationally using base Python functions (if/elif/else, find-replace functions, truncation, etc.). The phonological alternations are broken into two separate series, which I have dubbed *pre-compounding phonology* and *post-compounding phonology*. As the nomenclature implies, pre-compounding rules act on individual morphemes prior to synthesis into a verb stem, while post-compounding rules apply after. The former houses

[5] See 6.

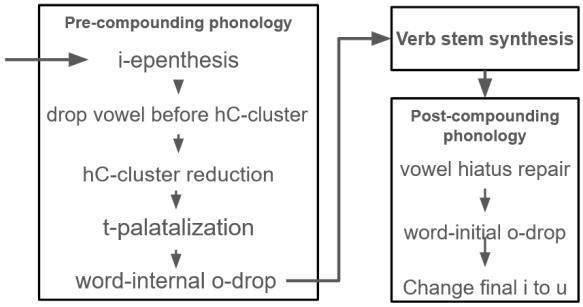

Figure 1: Diagram of simulated phonology rules applied to verb stem.

phonological rules that apply at morpheme boundaries, while the latter is for alternations that apply to the entire verb stem or at the word boundary.

A sketch of the rules and their application order is provided in Figure 1. The ordering is not trivial as some rules appear to feed others (e.g., i-epenthesis and t-palatilization).

This process yields 14,949,058 possible verb stems (502 initials * 97 medials * 307 finals). This list is then compared against a list of verbs from the Passamaquoddy-Maliseet Language Portal (pmp), an online dictionary and language resource. Over a hundred hours of transcribed conversation between native speakers are also available on the language portal. According to the website, at least 85 speakers are represented in the corpus. The transcriptions are particular helpful for finding token frequencies, which is relevant to the quantitative analysis of productivity presented in 5.

It should be noted here that the list of verbs from the dictionary contains only 14,941 elements, so just a fraction of the possible verb stems in the generated list can be attested. This asymmetry in sizes may be jarring, but this analysis will nevertheless provide some insight into the behavior of Passamaquoddy medials.

5 Data & analysis

Of the ≈15 million possible verb stems generated, exactly 697 of them had exact matches in the dictionary. 84 of the 97 medials have attested forms, with Figure 2 (see A) showing a Zipf-like distribution in frequency. The highest frequency of any medial was 77 (*atok*), while the lowest was 1 (*akomite*, *altoqe*, many others). A similar graph with additional information on mean token frequency is provided in Figure 3.

With this data, there are a number of ways to calculate the productivity of each medial. One established method in the morphological productivity literature is Baayen's criterion (Baayen and Lieber, 1991). This method defines the productiveness of a morphological process by the function $\frac{n_1}{N}$, where N is the total number of words formed by said process and n_1 is the number of hapax legomena (words that only occur once in a corpus) derived by that process. The greater the number of hapax legomena in a set of derived words, the more productive a process is said to be. Unfortunately, none of the verbs in the corpus used in this study are hapax legomena, so this is not an acceptable option.[6]

Another way of quantifying productivity is Yang's Sufficiency Principle (Yang, 2018), which states that a process is productive if it meets the following criterion:

(14) $(N - M) \leq \theta_N$ where $\theta_N = \frac{N}{lnN}$

...where N is the number of items to which the process can possibly apply, and M is the number of items known to undergo that process. This method seems particularly well-suited to this research, as it does not presuppose knowledge on the number of exceptions to a process. For any given medial, M is the number of attested verbs that contain said medial. It seems reasonable to conclude that every possible pairing of initial and final is the set to which a medial can apply, so N = 154,114 ($n_{initial}$ * n_{final}) and $\theta_N \approx$ 12,901.5. These calculations predict that none of the medials are productive, but they would predict this even if every verb in the corpus were an attested form with a single shared medial (15).

(15) (154,114 - 15,000) $\not\leq$ 12,901.5

One final method I consider in this work compares type frequency (number of forms derived by a process) with the mean token frequency of words formed by that process. Under this method, a construction is said to be more productive when the type frequency is high and the mean token frequency is low (Baayen, 1991). In figure 3, we see evidence for the productivity of some medials under this view. Specifically, *atok*, *amk*, *ek*, *ak*, *apsk*, and *alok* are plausibly so. For medials of lower type frequencies, though, there is no evidence to indicate that they are productive.

[6]The only hapax legomena in the corpus is *u*, an interjection analogous to English *oh*.

6 Concluding discussion

Considering several different techniques for operationalizing productivity, I find evidence to support the productivity of a small handful of medials. At the very least, figure 3 is a helpful indicator of which morphemes are more productive than others. Yang's Sufficiency Principle has a much less charitable view of the data, but this is arguably a case of the method not fitting the data we have. The amount of things a single medial can combine with is massive, and it alone trumps the size of the dictionary many times over.

It is curious why so few of the possible verb stems had attested forms in the dictionary. Of course, there are many things outside of *initial* + *medial* + *final* in the Passamaquoddy verbal lexicon. As previously stated, some verbs lack a medial and even an initial. So, it makes sense that some verbs in the dictionary do not have a companion in the comprehensive verb stem list. More puzzling are the missing medials from the attested forms list. I mention in 5 that only 84 medials are represented in the 697 attested forms; this indicates to me that there may have been an error in generating the possible verb stems. I am uncertain whether this was due to an unforeseen consequence of my own code or the result of a phonology rule I neglected to add, but it warrants further investigation.

Regardless, the results of this study provide a number of interesting avenues for future research. For the seemingly-productive medials, one question is their semantic import. Are their semantics consistent across every verb stem? Do they consistently have instrumental/theme/classificatory interpretations, or are the meanings of their respective verb stems more opaque or idiomatic? Conversely, for medials with less attested forms, are their semantics consistent? Returning to the question of productivity, I plan to continue investigating this matter through lexical decision tasks with native speakers. It would make sense to start with medials that already have a large number of attested cases, then moving on to medials with fewer.

In this work, I present a simple yet (as far as I am aware) novel approach to investigating morphological productivity in an endangered polysynthetic language. While the findings of this study are interesting, Passamaquoddy is only a small part of the full story concerning Algonquian verbal morphology; I am hopeful that the methodology I introduce here will be employed by others working on Algonquian languages with similar questions.

In conclusion, this work proposes a computational approach to to investigating the productivity of medials in Passamaquoddy. For languages with low vitality, such methods are especially valuable for research and revitalization efforts. While only a small number of generated verbs were actually attested in the dictionary, there is evidence for some medials being productive. Regardless, this work provides a foundation for future fieldwork and more "traditional" linguistic inquiry.

Acknowledgments

This work is indebted to Peter Grishin, Jonathan Rawski, Norvin Richards, and three anonymous reviewers. I thank them for their helpful insight and feedback. The lists of initials, medials, and finals used in this project was compiled by Norvin Richards. The PM Portal entry frequency data was compiled by Yadav Gowda. Errors are my own.

7 Data availability statement

For access to the data and code used in this study, please contact me at `jcrobert@mit.edu`.

References

Passamaquoddy-maliseet language portal; language keepers and passamaquoddy-maliseet dictionary project. http://www.pmportal.org.

Harald Baayen. 1991. Quantitative aspects of morphological productivity. In *Yearbook of morphology 1991*, pages 109–149. Springer.

Harald Baayen and Rochelle Lieber. 1991. Productivity and english derivation: A corpus-based study.

Peter Bakker. 1997. *A language of our own: The genesis of Michif, the mixed Cree-French language of the Canadian Métis*, volume 10. Oxford University Press.

Jerome Biedny, Matthew Burner, Andrea Cudworth, and Monica Macaulay. 2021. Classifier medials across algonquian: A first look. *International Journal of American Linguistics*, 87(1):1–47.

Leonard Bloomfield. 1946. Algonquian. In *Linguistic structures of Native America*, pages 85–129. Vinking Fund.

Phil Branigan, Julie Brittain, and Carrie Dyck. 2005. Balancing syntax and prosody in the algonquian verb complex. *Algonquian Papers-Archive*, 36.

Julie Brittain. 2003. A distributed morphology account of the syntax of the algonquian verb. In *Proceedings of the 2003 annual conference of the Canadian Linguistic Association*, pages 25–39.

Sophia Crockett-Current. 2020. Pursuing passamaquoddy-maliseet language revitalization through song.

Ives Goddard. 1990. Primary and secondary stem derivation in algonquian. *International Journal of American Linguistics*, 56(4):449–483.

Peter Grishin. 2023. Lessons from cp in passamaquoddy and beyond. *Dissertation, MIT*. https://ling.auf.net/lingbuzz/007567.

Philip S LeSourd. 1988. Accent and syllable structure in passamaquoddy. phd diss.

Simons Gary F. Lewis, M. Paul and Charles D. Fennig. 2016. *Ethnologue: Languages of the World*. SIL International.

Maria Mazzoli. 2023. Productivity, polysynthesis, and the algonquian verb.

Will Oxford. Algonquian language maps. http://home.cc.umanitoba.ca/~oxfordwr/algling/maps.html#cite. Accessed: 2024-01-28.

Conor McDonough Quinn. 2009. Medials in the northeast. *Unpublished ms.< http://www. conormquinn. com/MedialsInTheNortheast-AC40writeup. pdf*.

Tanya Slavin. 2012. *The syntax and semantics of stem composition in Ojicree*. University of Toronto (Canada).

Anna Whitney, Garrett Johnson, and Cherry Meyer. 2022. A survey of "classificatory medials" in ojibwe: Classifiers versus incorporation.

Charles Yang. 2018. A formalist perspective on language acquisition. *Linguistic Approaches to Bilingualism*, 8(6):665–706.

A Appendix: Figures

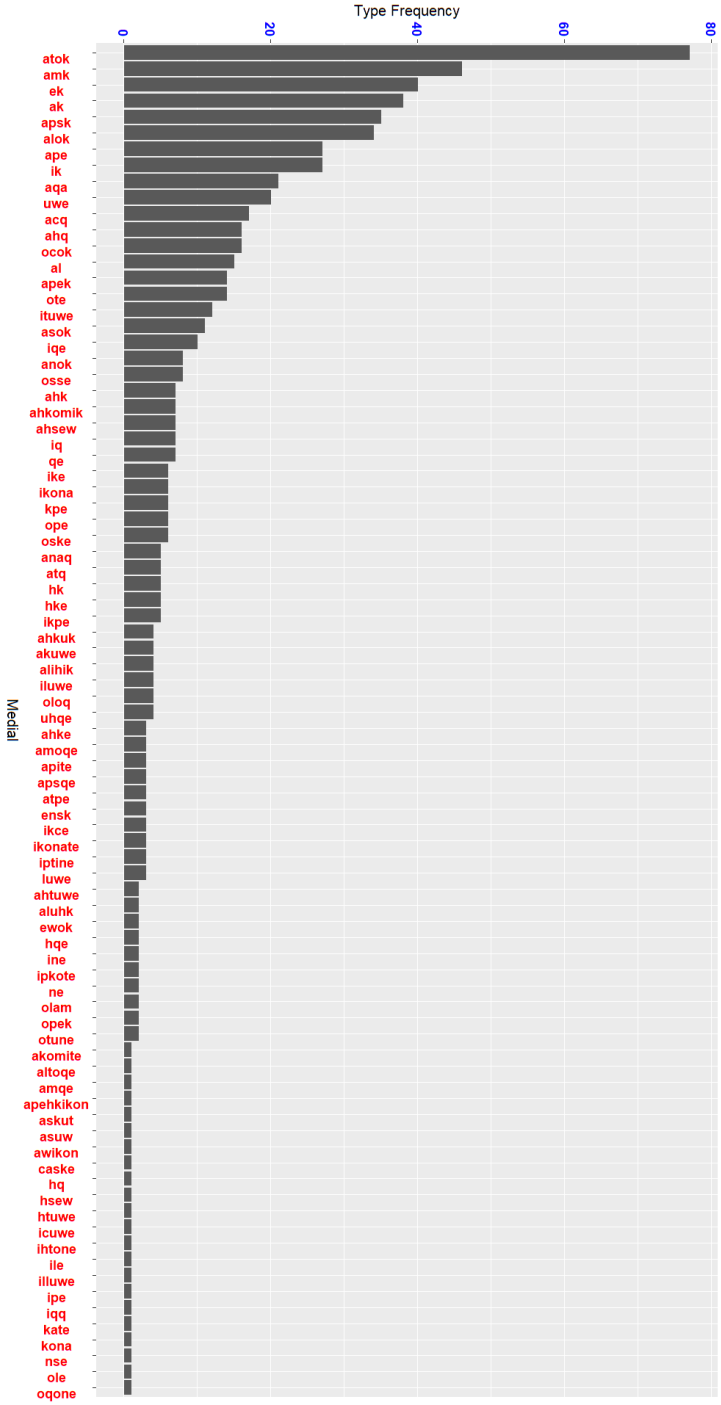

Figure 2: Frequency of medials for attested forms.

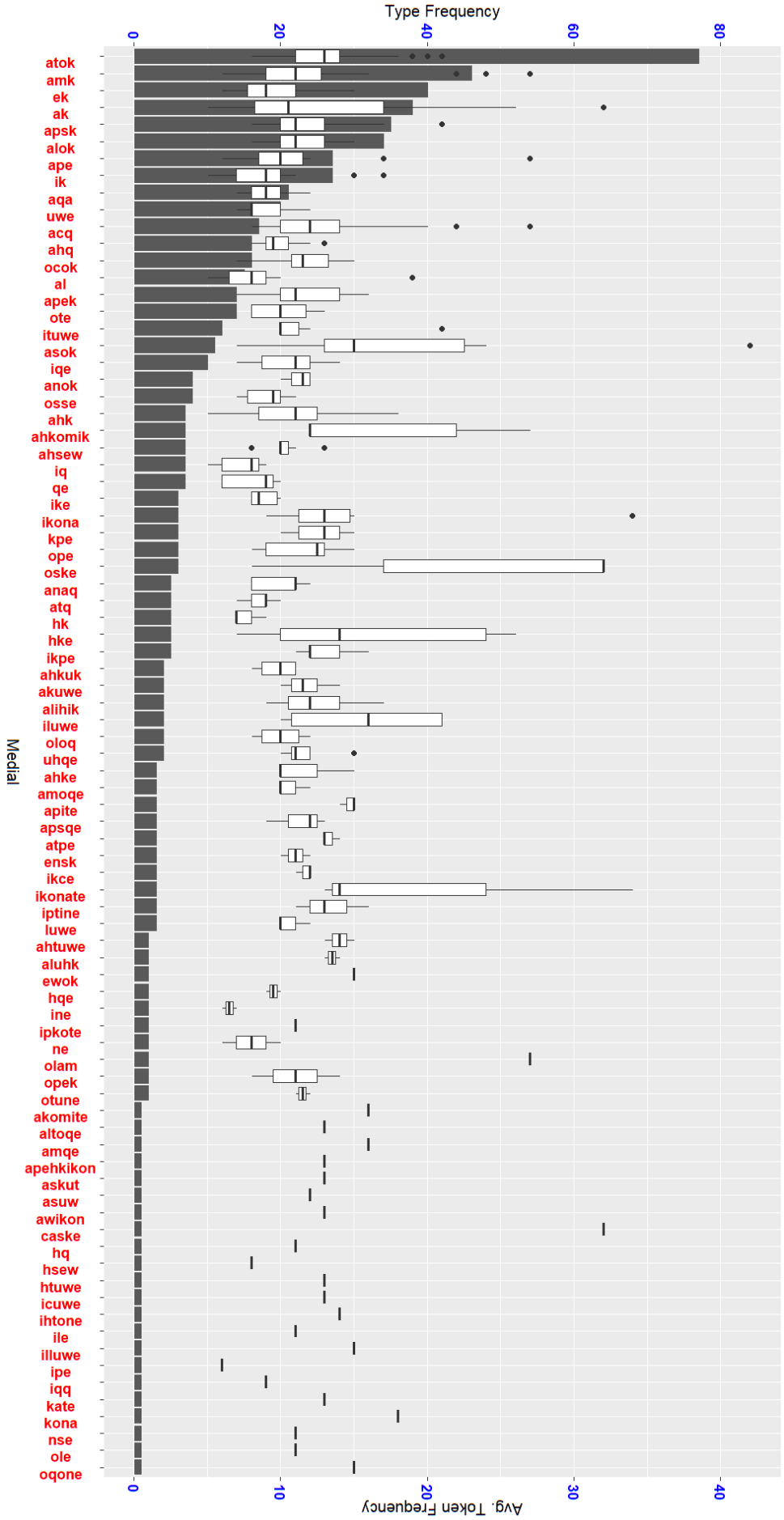

Figure 3: Frequency of medials and average frequency of tokens for attested forms. Columns are associated with the left y-axis, while boxplots are associated with the right.