Frequency norms in Tashlhiyt, Part I

John Alderete, Jane Li, Rachid Ridouane

Frequency norms in Tashlhiyt, Part I

John Alderete, Jane Li, Simon Fraser University

Rachid Ridouane, CNRS & Sorbonne Nouvelle, Paris

Abstract. This article gives a preliminary account of lexical and sub-lexical frequency norms in Tashlhiyt, an Amazigh language of Morocco. A corpus of nine texts with roughly 19k words was analyzed as a list of structured representations for words that distinguish part of speech, open/closed class, morphological make-up, and surface versus lexical representations. A lexicon was built from the corpus to identify unique word types and document the frequencies of these types. The frequencies of sub-lexical phonological segments are also documented, including how phonological representation, morpho-syntactic categories, and the type/token distinction affect segment frequency. All of the frequency norms are available as open data sets.

Key words: frequency norms, word frequency, segment frequency, corpus linguistics, Tashlhiyt, Amazigh

1. Introduction

Frequency norms are critical to psycholinguistic research, natural language processing applications, corpus studies, and, increasingly, linguistic analysis (Bod et al., 2003; Gries, 2016). For example, our understanding of language comprehension and production in Indo-European languages like English and Dutch depends on well-constructed and validated data sets that document frequency effects in these languages (Baayen et al., 1996; Kessler & Treiman, 1997). For a variety of reasons, the creation of these high-quality data sets has been largely focused on a small number of majority languages with large numbers of speakers and socio-economic power. For example, Anand et al. (2011) reports that among the 550 corpora available from the Linguistic Data Consortium, just five languages account for 85% of the data sources. The goal of this paper is to address this gap by creating a core set of frequency norms for word and segmental patterns in Tashlhiyt, an Amazigh language of Morocco.

Past research on Tashlhiyt has documented the frequencies of selected phonological patterns. In particular, quantitative accounts have been given of phonotactic patterns (Boukous, 1987), consonant co-occurrence patterns (Bensoukas, 2015), and contact phenomena (Boukous, 2012), but a general account of segment frequencies and segment combinations has not been given. As a sign of the interest in this topic, Orzechowska and Ridouane (2018) have recently given a detailed quantitative account of the patterns in long strings of consonants in vowelless verbal roots. Analysis of these combinations using principle component analysis identified the key features shaping these sequences, providing important insight into the phonotactic constraints on these verbs. However, a comprehensive account of segment frequencies in other word classes, word frequency, and the differential impact of type and token frequency has yet to be developed.

In this article, we document segment and word frequency in a corpus of 9 texts of Tashlhiyt (Boukous, 1977). The corpus has 18,827 word tokens of 4,195 lexical items from all word classes. As the first general account of these frequencies, we focus here on a few

foundational questions and problems. After describing our methods (section 2), we document word frequency in the lexicon and investigate high frequency words (section 4). In section 5, we document the frequencies of consonants and vowels, and examine in particular the impact of word class (i.e., noun versus verb), the open class/closed class distinction, type versus token frequency, and the level of the phonological representation (i.e., surface versus lexical). Section 6 summarizes these results and sketches how frequency norms can be developed in future study.

2. Methods

The frequencies of words and sounds are drawn from Jebbour et al. (2021), an electronic corpus of nine texts from Boukous (1977). The transcript of each text was digitized and analyzed as a list of structured representations, or *n*-tuples. That is, each word was encoded as a 6-tuple that analyzes different aspects of the word, and then ordered in the list as it appears in the text, as shown in Table 1 for the sentence, *Imma l!bisara mnʃk at tga i kilu* 'and how much does the bean soup cost?'

Table 1. Illustration of structured representations of words.

Order	Texte	SOUND	MORPHEMES	M-GLOSS	CATEGORY
…					
443	lxdmt	New_Sentence	New_Sentence	New_Sentence	
444	lxdmt	imma	[imma]	[whereas]	
445	lxdmt	l!bisara	[l-!bisara]	[bean soup]	N
446	lxdmt	mnʃk	[mnʃk]	[how much]	
447	lxdmt	at	[ad]	[AD:COMP]	
448	lxdmt	tga	[t-ga]	[3FS-cost:pf]	V
449	lxdmt	i	[i]	[for]	
450	lxdmt	kilu	[kilu]	[one kilogram]	N
451	lxdmt	New_Sentence	New_Sentence	New_Sentence	

The first two elements of the 6-tuple give the order in the text and the text's title. The third (SOUND) and fourth (MORPHEMES) give the surface and lexical representation, respectively. The surface form is the phonetic form of a word as it is pronounced, and the lexical representation is the surface form before any phonological rules have been applied. The M-GLOSS (fifth element) gives a gloss of the morpheme and its grammatical functions (see the documentation of Jebbour et al. (2021) for the coding principles for grammatical morphemes), and CATEGORY specifies N for nouns, V for verbs, and Ø for all other word classes. The first word of every sentence is preceded by a 6-tuple containing New_Sentence, indicating the end of the previous sentence and the beginning of a new one.

The sound frequencies documented below were established with searches on different components of the 6-tuples. For example, counts of surface segments in nouns were made by searching in the SOUND column, and only in words with 'N' under CATEGORY. In this way, we were able to distinguish lexical versus surface phonological representations, nouns versus verbs, and content versus functional items (i.e., all words that are not nouns and verbs), etc.

We also distinguished type versus token frequencies in the following way. Token frequencies are simply counts of all words in the corpus. Word types are unique words identical in morphological make-up, regardless of their lemma status. Concretely, items in the dataset with the same morphological make-up (underlying form), gloss, and part of speech are treated as one type. This entails that items with differing surface forms can be joined as one type. In such contexts, we chose the most frequent surface variant for sound-level type tabulations, following

common practice (Connine et al., 2008; Pitt et al., 2011). The word /a-mʁar/ ‘community leader’, for example, has two surface forms [jamʁar] and [amʁar]. Since [amʁar] is the more frequent variant, our type frequencies reflect this by adding 2 counts of [a], 1 count of [ʁ], [r], and [m] respectively. We are aware that some studies of morphologically rich languages also specify root type based on shared roots but not shared grammatical morphemes (Aksan & Yaldır, 2012; Gerz et al., 2018), but we leave typing based on roots for future research (see below). A lexicon of word types was built for the entire corpus, and type frequencies were generated using this lexicon.

Two large data tables documenting the frequency norms and a set of scripts for extracting frequencies are available at: https://github.com/jane-lisy/tashfreq. The contents of data tables for words (wordfreq_master) and sounds (soundfreq_master) are explained in more detail in the Appendix.

3. Word frequency

The corpus contains 18,827 word tokens and, as explained in the methods above, these tokens represent 4,915 word types with distinct form or morphological make-up. The word frequency of specific lexical items (among other things) is given in the wordfreq_master data table (see Appendix), but we report of the salient facts of the corpus as a whole here. In general, the vast majority of words have very low frequency: 3,191 items (65%) occur one time in the corpus, and another 660 (13%) occur only twice. Furthermore, there are just 18 lexical items with a frequency of 100 or greater, listed in Table 2. As one might expect, these are dominated by grammatical morphemes: only two forms of *say* and two names that are common in these texts rank among these other high-frequency function words.

Table 2. All lexical items with frequency greater than 100.

Lexical Item	Part of speech	M-Gloss	Frequency	Probability
ad		AD:Comp	518	0.02842
ias		to-dat-3S	423	0.02321
a		Voc	301	0.01651
ar		AR	286	0.01569
ur		neg	270	0.01481
i		to	224	0.01229
as		dat-3S	206	0.01130
inna	V	3MS-say:pf	193	0.01059
iini	V	3MS-say:ao	187	0.01026
awa		so	168	0.00922
d		and	138	0.00757
ns		PossDet:3S	138	0.00757
iji		dat:1S	138	0.00757
gi-s		loc:3MS	120	0.00658
ħasan	Name	Hassan	118	0.00647
n		of	113	0.00620

!rbbi	Name	God	110	0.00604
nkki		1S:me	104	0.00571

The relative frequency of nouns, verbs, and closed class morphemes are reported in Table 3. While nouns and verbs together as a group have about the same frequency in the corpus as all other words, they greatly outnumber these words in the lexicon, which is expected, because these other words are by and large closed class items with higher corpus frequency.

Table 3. Tokens, types, and mean frequency by part of speech category.

	Tokens	Types	Mean frequency
Nouns	4363	1490	2.9282
Verbs	4567	2169	2.1056
Other	9297	1256	7.40207

The general skewing of words towards the low end of the frequency spectrum is shown in Figure 1, as is the Zipfian-like distribution in the lexicon (Zipf, 1949). There are a relatively small number of very high-frequency words, which falls very quickly in frequency (from left to right in the graph), until word frequency plateaus at the low end of the spectrum.

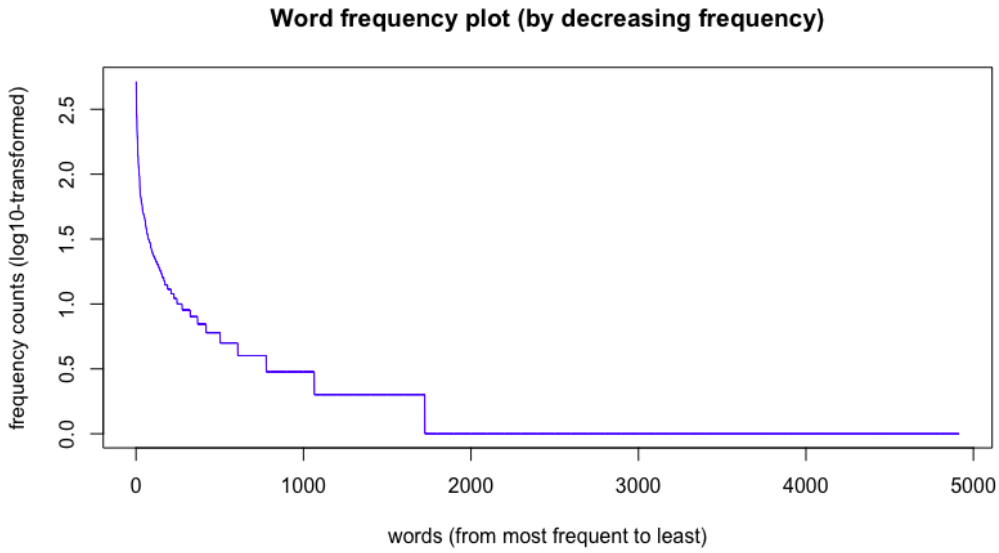

Figure 1. Distribution of word frequency in the corpus.

4. Consonant and vowel frequency

We report now the segment frequencies for all segments, including consonants and vowels in a variety of conditions of interest. We begin with frequency distributions for consonants (Table 4) in surface representations of all words. The counts of the six most frequent segments are color-coded (from highest to lowest: yellow, green, aqua, magenta, grey, red). Within this set of high-frequency segments, *t, s, n,* and *l* rank in the top six in both token and type frequencies, and in singleton and geminate sounds. The relative ranks, however, do differ in these classes. For example, *kk* and *qq* have higher relative frequency in geminates than in singletons. Furthermore,

while token and type frequency tend to be correlated, there are clear differences. For example, the relative rank of *s* drops from token to type frequency, but the opposite pattern is found for *l*. This table also establishes quantitatively some intuitions that many Amazigh linguists have for segment frequencies, namely that geminates and emphatics are far less frequent than their singleton and non-emphatic counterparts. Finally, some segments are exceedingly rare or non-existent in this corpus, including the velars with secondary labialization, the voiced geminate uvular ʁʁ, and the geminate pharyngeal ʕʕ.

Table 4. Consonant frequencies, all words, surface representations (color coding for top six segments)

Singleton	Token	Type	Geminate	Token	Type
b	822	391	bb	237	55
t	3935	1608	tt	637	334
d	1979	643	dd	431	140
tˤ	155	106	ttˤ	189	92
dˤ	328	194	ddˤ	34	24
k	1617	570	kk	395	153
g	977	295	gg	103	68
kʷ	3	3	kkʷ	1	1
gʷ	2	2	ggʷ	0	0
q	179	101	qq	312	197
qʷ	1	1	qqʷ	0	0
f	894	404	ff	73	39
s	2831	705	ss	427	185
z	447	197	zz	183	97
sˤ	156	105	ssˤ	77	49
zˤ	169	106	zzˤ	19	16
ʃ	385	204	ʃʃ	100	62
ʒ	151	96	ʒʒ	77	42
ʃˤ	27	21	ʃʃˤ	28	14
ʒˤ	25	15	ʒʒˤ	49	9
χ	320	156	χχ	47	14
ʁ	1819	642	ʁʁ	1	1
χʷ	0	0	χχʷ	0	0
ʁʷ	4	3	ʁʁʷ	0	0
ħ	769	286	ħħ	7	4
h	685	191	hh	4	3
ʕ	562	296	ʕʕ	0	0
m	2214	866	mm	293	115
n	3859	1365	nn	937	265
nˤ	133	76	nnˤ	12	10
l	2340	966	ll	990	315
lˤ	152	88	llˤ	122	33
r	2737	818	rr	132	65
rˤ	660	333	rrˤ	70	43
w	1837	650	ww	24	16
y	2377	633	yy	308	61

These frequencies are aggregated into manner, place, and voicing classes in Table 5, which excludes glides because of their vowel-like status. Many of these frequencies are

expected. For example, coronals have the highest token and type frequency among the place classes, consistent with the fact that they dominate the consonant inventory with 36 members. Interestingly, however, the frequencies of labials and dorsals are comparable, but there are roughly three times as many dorsals as labials in Tashlhiyt. Likewise, voiced consonants are both more numerous in the inventory and frequency. Fricatives, however, outnumber stops in the inventory by a proportion of roughly 3 to 2, but stops are much more frequent than fricatives in both type and token frequencies. The descriptive accounts of these natural classes and many others can be explored further with the data supplement.

Table 5. Frequencies of natural classes.

Type	Natural class	Token	Type
Place	Labial	4533	1870
	Coronal	24983	9441
	Dorsal	5781	2398
	Pharyngeal	1338	586
	Glottal	898	407
Manner	Stop	12337	4978
	Fricative	10336	3962
	Nasal	7448	2697
	Liquid	7203	2661
Voicing	Voiced	23759	8884
	Voiceless	13565	5414

As for vowel frequencies (Table 6), this trio of segments decreases in frequency, *a* > *i* > *u*, in both token and type frequencies.

Table 6. Vowel frequencies, all words, surface representations.

	Token	Type
i	8175	2645
u	2488	868
a	11907	3270

As far as how vowels pattern with the rest of the inventory, the distribution of vowels relative to consonants is roughly 35/65% in tokens and 30/70% in types.

As explained in section 2, the structured representations in our corpus distinguish lexical and surface representations of words. How well are the frequencies in these representations correlated? Here and throughout we assess this by giving the correlation coefficients for consonants (including geminates) and all segments (all consonants and vowels). Thus, the frequency distribution of all consonants is a vector of 72 segment counts, and 75 counts for all segments. To compare two vectors, we assessed the relationship between two classes (e.g., frequencies in nouns versus verbs) with Pearson's correlation coefficient *r*. As shown in Table 7, the correlation between frequencies in lexical and surface representations is very strong in both token and type frequency. Also, in both token and type frequencies, these correlations improve with all segments, presumably because of the large counts of vowels. In sum, one can say that the representations differ slightly before and after phonological processes have applied, but not very much.

Table 7. Correlations between surface and lexical frequencies.

	Token	Type
Consonants	0.9856	0.9952
All segments	0.9935	0.9974

Though the differences between lexical and surface forms are minor, we nonetheless analyzed the mis-matches in case they might be relevant to a future study. Of the 4,915 items in the lexicon, only 503 (10.23%) mis-match with the surface form (though these mis-matches do not include changes in surface form due to emphatic spread, which is substantial). The average string edit distance of the mis-matches is 1.14. Therefore, while a non-trivial percentage of the lexicon has a mis-match, the mis-matches are typically in a single segment, consistent with the high correlations in Table 7. Among the 503 mis-matched words, roughly a fifth involve a deletion or addition, and the remaining phonological patterns involve well-known alternations in Tashlhiyt, including vowel-glide alternations and devoicing, as broken down in Table 8.

Table 8. Phonological patterns in mis-matched words.

Pattern	Example	*N* (of 503)
Changes segment count		92
Deletions	/ra-nt/ → ran	56
Additions	/ajt/ → jajt	36
No changes segment count		411
i ~ *j*	/i-ili/ → jili	271
t ~ *d*	/ard/ → art	38
u ~ *w*	/uarjal/ → warjal	31
ss ~ *zz*	/issnza/ → izzna	21
Other	/mad/ → maj	50

Next, we ask if and how segment frequencies are affected by morpho-syntax. It is well known that the distributions of segments are different in content and function words, often with function words having more restricted sound inventories (Beckman, 1999; Willerman, 1994). Furthermore, we expect large differences in segment distributions because of Tashlhiyt morphology. For example, approximately 38% of verbs in our corpus start with *i* ~ *j* due in part to the prevalence of the 3rd person masculine prefix, whereas only 11% of nouns start with these sounds, and verbs likewise introduce schematic patterns not found in nouns, like the occurrence of geminates in imperfectives. Table 9 gives the correlations between nouns and verbs, and content and functional items for both token and type frequencies. These results show that, while correlated, there are significant differences between nouns and verbs, on the one hand, and content and functional items, on the other. The correlations in consonant and segment frequencies are also stronger in type frequencies, as one would expect because the impact of high frequency words (skewed toward functional items) is reduced.

Table 9. Correlations based on morpho-syntactic structure.

	Token		Type	
	Consonants	All segments	Consonants	All segments
Noun vs. Verb	0.7917	0.8346	0.8503	0.8810
Content vs. Functional	0.7673	0.9021	0.8589	0.9373

Finally, are token frequencies correlated with type frequencies? That is, are the frequencies as they are reflected in the entire corpus well-correlated with the frequencies in the lexicon? Table 10 breaks these frequencies down by the four morpho-syntactic classes, and the distinction between consonants versus all segments. In general, token and type frequencies are highly correlated, and they become stronger when the entire set of segments is involved, as observed above. Nouns and verbs do not seem to differ very much, except with the slightly weaker correlation in verbs for all segments. Perhaps the most noticeable difference is in function words, which have a .07 drop in *r* in consonants relative to nouns and verbs. We expect this because a higher percentage of very high-frequency words are function words. Thus, when the effect of these words is lost in the lexicon, we expect changes in the frequency distributions of the sounds of these words.

Table 10. Correlations between token and type frequencies.

	Consonants	All segments
Nouns	0.9873	0.9952
Verbs	0.9831	0.9815
Function words	0.9188	0.9785
All words	0.9748	0.9873

5. Conclusion

This report gives a first approximation of the frequency distributions of words and segments in Tashlhiyt. We have examined a relatively large corpus of words, built a lexicon for this corpus, and analyzed the impact of the token/type distinction, lexical and surface representation, and morpho-syntactic classes on frequency norms. It turns out that the distribution of segments is affected very little by the lexical-surface distinction and the token-type distinction, though specific segments sometimes do exhibit non-trivial differences. Distributions based on different morpho-syntactic classes (i.e., noun versus verb and content word versus functional item) are also correlated, but the correlations are much weaker. The morphology of these words, and the basic distinction between open and closed class items, therefore affect the relative frequencies of sounds.

These results, and the data they are built from, can be employed in a variety of ways. They can be used in psycholinguistic experiments where word and segment frequency need to be balanced or manipulated. For example, one can use the word frequency data supplement to find stimuli from high versus low frequency classes. Likewise, experimental items can be balanced for type frequency by using the frequency information from the sub-lexical data supplement. In addition, all of these frequency calculations can be tailored to the experimental task, selecting frequencies from a particular kind of representation (lexical or surface), morpho-syntactic context (noun, verb, functional item), and frequency type (token or type). Beyond these applications, the frequency distributions documented here can inform linguistic analyses that

require frequency to determine core ingredients of phonological analysis (Blevins, 2004; Bybee, 2001; Hayes & Wilson, 2008).

While we believe the frequency norms documented here are of some use, this report is just the beginning of documenting frequency norms. In particular, we envision several ways of improving them so that they can better inform other kinds of investigations. First, our lexicon uses a word-based typing system, but perhaps a better way to investigate the token/type distinction is by recognizing root-based types. Though Tashlhiyt does not have distinct inflectional classes, it has rich verb and noun morphology, and so it makes sense to distinguish word types on the basis of a shared root or stem, as advocated by Aksan and Yaldır (2012) for Turkish. Doing so will reduce the number of word types and, as a result, it will change the type frequencies of segments and word frequencies. We do not think this will affect segment frequencies very much, but it will affect word frequencies, as a considerable amount of word types in our current lexicon will be collapsed into a single type.

Another way to further develop the project is to create a syllabified surface representation in the corpus and use them to derive frequencies for whole syllables and segments cross-classified by syllabic role (e.g., *n* in an onset as opposed to acting as a syllable nucleus). The algorithms for syllabifying segmental strings are well-known (Dell & Elmedlaoui, 2002; Jebbour, 1996), and so enriching the structured representations in this way is a tractable problem. Furthermore, Tashlhiyt is typologically unusual in that it allows all consonants, including non-sonorants, to occupy the syllable nucleus. An understanding of the frequencies of these different syllable nuclei, and the frequencies of syllables containing them, is of potential interest to investigations that seek to understand how syllable composition affects language behavior. For example, how does the syllable inventory of Tashlhiyt differ from well-known Indo-European languages, and are the syllables that are unattested in these languages, but attested in Tashlhiyt, spoken and perceived in different ways?

Finally, while the frequency norms we document here can inform new investigations, they have not yet been validated in the sense that they have been shown to predict certain well-known linguistic behaviors. For example, we expect high-frequency words to be classified as words of the language by native speakers faster than low-frequency words in lexical-decision tasks, and we expect words with high frequency syllables to be spoken faster than other words (Levelt & Wheeldon, 1994; Whaley, 1978). In terms of psycholinguistic behavior, Tashlhiyt is an under-studied language, and so it is difficult to find relevant behavioral data to correlate with the frequency distributions we have established here. There is no substitute to collecting behavioral data that can validate our norms. But in lieu of such data, perhaps the first step in data validation should be to examine the well-studied effect of frequency on reduction at the word and segment level (Aylett & Turk, 2006; Gahl, 2008). Indeed, we are currently working on a spoken version of the corpus used for this study, and so appropriate data will soon be available for probing the effects of frequency on reduction patterns. We hope that with this, and other efforts, future research will expand the empirical base for investigating frequency in Amazigh languages.

Appendix

All of the frequency data and scripts for generating them are available at the GitHub page for this project: https://github.com/jane-lisy/tashfreq. Two data tables are especially useful. The file <wordfreq_master> (in various formats) gives all of the information relevant to word frequency and word structure documented in section 3. This file is the lexicon, in effect, listing all words as

rows. The seven columns in the document analyze words for the following attributes: lexical representation, phonetic representations (including all variant forms), morphological gloss, part of speech category, the number of morphemes contained in the word, word frequency and word probability.

The file <soundfreq_master> contains all of the information for investigating sound frequencies discussed in section 4. It has 75 rows (for 36 singleton consonants, 36 geminates, and 3 vowels) and the 20 columns that give the frequencies for these 75 segments cross-classified by all of the categories discussed here (i.e., level of representation, frequency type, morpho-syntactic categories). Many levels of structure can be investigated: for example, token frequency in nouns, or open class words (both nouns and verbs), or all words can be easily extracted.

The Python scripts for investigating the corpus are also available from the project page. The Python notebooks for these scripts are fully commented and enable users to replicate the results reported here and extend them to new projects.

References

Aksan, Y., & Yaldır, Y. (2012). A corpus-based word frequency list of Turkish: Evidence from the subcorpora of Turkish National Corpus project. *Studia Uralo-altaica, 49*, 47-57.

Anand, P., Chung, S., & Wagers, M. (2011). Widening the net: Challenges for gathering linguistic data in the digital age. *NSF SBE 2020: Future research in the social, behavioral, & economic sciences*.

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of the Acoustical Society of America, 119*, 3048-3058.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). *CELEX2*. Philadelphia: Linguistic Data Consortium.

Beckman, J. N. (1999). *Positional faithfulness: An optimality theoretic treatment of phonological asymmetries*. New York: Garland.

Bensoukas, K. (2015). *Feature dissimilation in Tashlhit*. Rabat: University of Mohammed V, Publications of the Faculty of Letters and Social Sciences, These and Memoirs series, no. 72.

Blevins, J. (2004). *Evolutionary Phonology: The emergence of sound patterns*. Cambridge: Cambridge University Press.

Bod, R., Hay, J., & Jannedy, S. (2003). *Probablistic linguistics*. Cambridge, MA: The MIT Press.

Boukous, A. (1977). *Langage et culture populaire au Maroc, essai de sociolinguistique*. Casablanca: Dar El-kitab.

Boukous, A. (1987). *Phonotactique et domaines prosodiques en berbère (parler tachelhit d'Agadir, Maroc).* (Doctoral dissertation). Université Paris 8,

Boukous, A. (2012). *Revitalisation de la langue Amazighe. Defis, enjeux et strategies*. Rabat: Top Press.

Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Connine, C. M., Ranbom, L. J., & Patterson, D. J. (2008). Processing variant forms in spoken word recognition: The role of variant frequency. *Perception and Psychophysics, 70*, 403-411.

Dell, F., & Elmedlaoui, M. (2002). *Syllables in Tashlhiyt Berber and in Moroccan Arabic*. Dordrecht: Kluwer.
Gahl, S. (2008). Time and Thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language, 84*, 474-496.
Gerz, D., Vulić, I., Ponti, E., Naradowsky, J., Reichart, R., & Korhonen, A. (2018). Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction. *Transactions of the Association for Computational Linguistics, 6*, 451-465.
Gries, S. T. (2016). *Quantitative corpus linguistics with R: A practical introduction*. New York: Routledge.
Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry, 39*, 379-440.
Jebbour, A. (1996). *Contraintes prosodiques et morphologie en berbère (tachelhit de Tiznit - Maroc). Analyse linguistique et traitement automatique.* (These de Doctorat d'Etat). Université Mohammed V, Rabat, Morocco,
Jebbour, A., Boukous, A., El Alami, A., Li, J. S., Ridouane, R., & Alderete, J. (2021). *Nine texts of Tashlhiyt Berber: structured representations of 18,000 words (First Release).*
Kessler, B., & Treiman, R. (1997). Syllable structure and the distribution of phonemes in English syllables. *Journal of Memory and Language, 37*, 295-311.
Levelt, W. J. M., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary. *Cognition, 50*, 239-269.
Orzechowska, P., & Ridouane, R. (2018). *The structure of vowelless verbal roots in Tashlhiyt Berber.* Paper presented at the 9th International Conference on Speech Prosody 2018, Poznan, Poland.
Pitt, M. A., Dilley, L., & Tat, M. (2011). Exploring the role of exposure frequency in recognizing pronunciation variants. *Journal of Phonetics, 39*, 304-311.
Whaley, C. P. (1978). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior, 17*, 143-154.
Willerman, R. (1994). *The phonetics of pronouns: articulatory bases of markedness.* (Doctoral dissertation). University of Texas at Austin,
Zipf, G. K. (1949). *Human behavior and the principle of least effort* Reading, MA: Addison-Wesley.