The gender system of Tashlhiyt: Using supervised learning to predict noun gender

John Alderete, Simon Fraser University
Piyush Agarwal, Simon Fraser University
Kaye Holubowsky, Simon Fraser University
Abdelkrim Jebbour, Ibn Tofail University (Morocco)

Abstract: This article develops a comprehensive account of the system of gender assignment in Tashlhiyt and the first quantitative account of gender in an Amazigh language. From a corpus of 1914 noun paradigms, we establish the number of genders, gender morphology, and the masculine default in Tashlhiyt within contemporary theory of gender assignment. While much prior work focused on gender in wordforms, we explore form and meaning attributes of word lemmas and use them to predict Tashlhiyt gender with a set of computational classifiers. The resulting quantitative analysis reveals important roles for previously unacknowledged morphological and phonological attributes, and it also presents a broader argument for using machine learning techniques in the linguistic analysis of gender assignment. All data and models are open access.

Keywords: gender assignment, typology, Tashlhiyt, Amazigh, machine learning, classification, supervised learning

1. Introduction

Gender assignment, or the association of nouns to one of a fixed set of gender categories, is a surprisingly complex task. Adult native speakers have a strong purchase of the gender of thousands of words, retrieve these words in a matter of milliseconds, and use them in largely error-free fashion (Corbett 1991; Corbett 2014; Kramer 2020). Contemporary theory assumes that this success derives from the existence of an underlying system that draws on attributes of nouns, principally form and meaning, and uses them to predict noun gender. Though the task is daunting, language users can use this system in fast online production, as well as apply it productively in novel settings, such as loans and nonce words (Barkin 1980; Kupisch et al. 2022).

While at the macro-level, gender assignment exhibits properties of a system, it is often difficult to ascertain just how this system works. This is not to say that linguists lack the tools for predicting gender assignment. Coherent typologies exist for classifying and understanding the factors predicting gender that employ a range of semantic, morphological, and phonological variables (Corbett 1991; Corbett 2007; Payne 1998) as well as gender in online cognition and conflicting morpho-syntactic environments (Corbett & Fraser 2000; Gerdts 2011; Rice 2006). Rather, developing a linguistic analysis that pins down the important factors and their interaction can be quite a challenge, especially if a language is under-studied and lacks lexical resources. For example, many languages have been argued to have arbitrary gender and later found to be predictable, as in form-based gender in French (Tucker et al. 1977). Predictable systems also exhibit complexity that has required researchers to re-assess their assumptions and re-analyze gender based on new findings. Thus, Arapesh was originally assumed to have phonological gender assignment but was updated in later work to a morphological system based on inflection

classes (Aronoff 1991). Discovery of a gender assignment system therefore requires both a deep understanding of the relevant linguistic structures and the rigorous pursuit of how these factors predict gender.

This article gives a detailed investigation of the linguistic structures of gender assignment in Tashlhiyt, an Amazigh language of Morocco. On the first pass, Tashlhiyt seems to have a relatively simple two-category gender system based on phonology. A broad generalization that has guided prior analysis is that masculine citation forms by-and-large begin with a vowel, and feminine forms begin with *t* (for review in English, see Dell and Elmedlaoui (2002), p. 26 ff.). On deeper examination, however, this generalization is more a reflection of the inflectional rules responsible for gender than a hypothesis about the phonological basis for predicting gender. These rules supply masculine words with an initial vowel in bases that lack them, and feminine forms also receive a *t-* prefix by rule (Dell & Jebbour 1991). This analysis effectively nullifies the phonological generalization because the C/V status of the initial segment is the result of rules that assign the categories that the generalization is intended to predict. In other words, it is a system of describing gender in wordforms, not a system of predicting gender from word lemmas. In the investigation below, we examine all plausible morphological, phonological, and semantic factors, as well as their interaction, in an effort to explicate the underlying system of gender assignment. We conclude by documenting the important roles for both morphological and phonological factors that have not previously been explored.

Our aim is to contribute to theories of gender by giving a detailed empirical account of gender assignment in an under-studied language and by developing a methodology for investigating gender with machine learning techniques. In particular, we develop and analyze a corpus of 1914 noun paradigms in Tashlhiyt and investigate the potential of 14 noun attributes for predicting gender. Though prior research in Amazigh languages has led to consensus on the number of genders and the basic morphology (see e.g., El Moujahid (1997)), our account is the first comprehensive treatment of Tashlhiyt that brings these elements together in a coherent analysis. It is also the first to establish the key factors predicting gender using quantitative methods with a large data set in an Amazigh language. Following recent work using machine learning to discover and validate morphological generalizations (Ahlberg et al. 2015; Allassonnière-Tang et al. 2021; Quint & Allassonnière-Tang 2022), we approach the predictability of gender as a classification problem using labelled data. In particular, we train several computational classifiers on a portion of the data and then test their ability to predict gender in hold-out data. The resulting account is both a robust quantitative analysis of gender assignment and an argument for using machine learning methods to analyze complex gender data. These contributions support the larger theoretical enterprise of ascertaining the underlying factors predicting gender assignment in nouns and their interactions (Corbett 1991).

The rest of this article is organized as follows. Section 2 provides the methods of constructing our noun corpus and an introduction to computational classification methods. Section 3 gives a comprehensive account of gender by establishing the agreement classes that underlie gender, a description of the morphology of gender and how it patterns in noun paradigms, as well as a short discussion of the masculine default assumed in this description. Section 4 gives the quantitative account of gender assignment, first reporting the descriptive statistics of how several linguistic attributes pattern with gender, and then uses these facts to build several classifiers to predict gender assignment using supervised learning techniques. Section 5 concludes the article with some questions for future research.

2. Methods

2.1 The database

The database of 1914 noun paradigms (rows) with 50 linguistic attributes (columns) is available at: https://github.com/aldo-git-bit/tashlhiyt-predicting-gender. The nouns reflect the native speaker intuitions of the fourth author, whose variety of Tashlhiyt is spoken in Tiznit. The database is in fact the culmination of data collection from several research projects by the first and fourth authors (Dell & Jebbour 1991; Dell & Jebbour 1995; Jebbour 1988; Jebbour 1996; Jebbour et al. 2021). Many of these projects centered on questions of plural formation and the prosodic structure of stems, and so the corpus may be skewed toward nouns that exhibit structures relevant to this research. To mitigate this problem, we also examined all of the nouns from nine Tashlhiyt texts (Boukous 1977). In matching up these two sources, we found substantial overlap between the nouns from texts and our initial noun list from prior research, but we also found and incorporated several hundred new noun paradigms in the database. The size of the corpus is more than twice the size of the corpus used in a comparable study, Allassonnière-Tang et al. (2021), which established clear results with 917 nouns, and our data set of underived nouns alone (n=1260) is a 37% increase in size from this study. Given these facts, we believe that the current database is representative of the noun paradigms in Tashlhiyt and suitable for our methods.

Each record in our database contains a complete noun paradigm with wordforms and a host of lexical and analytical attributes describing the paradigm. In addition to English and French glosses, semantic attributes include human/non-human, animate/inanimate (where animates include humans and animals), sex/social gender of animate nouns, and the semantic field that best characterizes the noun's meaning. For the latter, we used a simplified version of the semantic fields commonly used for Indo-European languages (Buck 2008). While this field list was developed for Indo-European languages, it is a standard reference for semantic fields, and our native speaker consultant had no trouble fitting noun meanings to these categories. Nouns were also classified as either native or loanwords, and the latter were classified as either assimilated and unassimilated Arabic loans or loans from French and Spanish (which were much smaller in number).

The record for each noun paradigm also analyzes its formal characteristics, including phonological and morphological attributes that we investigate in detail below. The noun's theme (i.e., wordform minus inflections for number, gender, and state) is given, analyzed phonologically for CV structure, segment length, and initial and final segments. Morphological attributes include the existence of initial vowel augments (i.e., the result of the R-AUG rule discussed below), secondary morphology for diminutive/augmentative distinctions, derivational morphological categories, loanword source, and plural formation pattern. We use all of these linguistic attributes in our analyses below as potential predictors of gender, which is also recorded for each paradigm.

2.2 Data analysis

To explore gender assignment quantitatively (section 4), we first start with descriptive statistics, tabulating the distributions of the above linguistic attributes and building contingency tables that cross-classify these attributes with gender categories. The results of these searches are used to test for confounds among variables and inform feature engineering to improve the predictive value of a feature.

In our quantitative analysis, we use a set of machine learning methods to investigate which factors predict gender. We start by examining how well variables from the traditional classes of gender systems predict Tashlhiyt gender, that is, semantic, morphological, and phonological factors, and then combine these groups of factors to get a sense of the predictability of the larger system. Current research examining a host of unrelated languages generally assumes that gender systems, however they fit in gender typology, are at least 85% predictable, meaning that gender can generally be predicted from the meaning and form of a noun most of the time (Corbett 1991; Corbett 2007). We use this benchmark to assess predictability in Tashlhiyt gender assignment.

We analyze gender predictability as a classification problem using supervised learning with labelled data (Hastie et al. 2009). After exploring the data and developing the features through feature engineering, we develop a host of models for predicting gender using standard cross-validation training, including Random Forest, Gradient Boosting, Adaboosting, Support Vector classifiers, and a voting classifier that takes the outputs of these models as inputs to standardize model predictions. Our goal here is not to develop the best model for predicting gender, though as our quantitative analysis demonstrates, many of them achieve the 85% predictability benchmark. Rather, we are interested in a composite analysis that takes into consideration different models composed of different groups of attributes to get a general assessment of what type of gender system Tashlhiyt has (e.g., semantic vs. formal), and within this system type, what features are particularly important. We conjecture that, while models may differ in their overall accuracy, the relative importance of linguistic factors will be broadly consistent. As demonstrated by the nascent research using these methods (e.g., Allassonnière-Tang et al. (2021)), machine learning provides powerful tools for answering these questions.

While our use of machine learning is new to Tashlhiyt linguistics, we believe that a successful case for computational classifiers for gender can contribute to the flourishing field of computation linguistics applied to Amazigh languages. This field has made rapid advancements in computer-assisted linguistic analysis (Jebbour 1996; Taghbalout et al. 2015), automatic language detection (Adouane et al. 2016a; Adouane et al. 2016b; Lafkioui 2008), discovery in low resource languages (Allah & Boulaknadel 2012), machine translation (Taghbalout et al. 2015), and the development of annotated data sets (Amri et al. 2017; Jebbour et al. 2021). It has also supported the development of practical NLP tools for Amazigh languages, such as the TALAM portal (https://tal.ircam.ma/talam/). We hope that the computational analyses we develop, as well as our structured noun data set, can extend these efforts in new directions.

3. The gender system of Tashlhiyt

3.1 Language background

Tashlhiyt is an Afro-Asiatic language spoken in Southwestern Morocco. It has basic VSO word order, rich agreement, and a complex system of post-verbal clitics for marking objects, prepositional phrases, adverbials, and deictics. The morphology of Tashlhiyt includes both concatenation and non-concatenative morphology (Bensoukas 2001; Dell & Elmedlaoui 1992; El Hamdi 2018). Though the nature of bases for morphological processes is controversial, compelling arguments have been made, contra consonantal root theories (e.g., McCarthy (1979)), that words are based on stems that may include vowels (see Dell and Elmedlaoui (2002) for review). For more detailed descriptions of Tashlhiyt, Dell and Elmedlaoui (1989, 1991) give an introduction to the morphology and morpho-syntax of Tashlhiyt. In addition, Aspinion (1953),

though it is a pedagogical grammar with some archaic forms, gives an excellent description of the linguistic structures of the larger language, and Alderete et al. (2015) gives a grammatical overview by summarizing both theoretical and descriptive research on all aspects of the language.

Prior research on gender in Amazigh languages establishes a consistent account of the core elements of gender. In particular, there is general consensus on the number of genders (two, masculine and feminine), the basic morphology involved (with some language-particular differences), and the importance of gender in marking secondary morphology like diminutives. Aspinion (1953) gives the most complete account of Tashlhiyt, describing the markers of gender, the allomorphy of the *t-* prefix, and secondary morphology. These facts and analytical assumptions of the larger system have been confirmed in many studies, including encyclopedic accounts of gender in the language family (Chaker 1998), generative analysis of gender (El Moujahid 1997), gender in Kabyle (Mettouchi 1999), the Ait Atta variety of Amazigh (Boussayer 2021), and Zénage (Taine-Cheikh 2002) (an Amazigh language of Mauritania), gender in secondary morphology in the larger language family (Vycichl 1961), and the semantic categories associated with gender in Ayt Wirra (Oussikoum 2019), an Amazigh language of the Middle Atlas.

These studies provide tremendously valuable cross-linguistic information and establish the broad outlines of the gender system. However, taken as a whole, they do not provide the formal evidence for gender categories by establishing agreement classes, the standard in the field (Corbett 1991). Past research has also not investigated the overall predictability of gender and related it to typologies of gender assignment, nor has it investigated gender in any Amazigh language using large data sets. The rest of this section investigates gender assignment in Tashlhiyt to address the first question, and section 4 addresses the remaining questions with quantitative methods.

3.2 Agreement domains and classes

The first step in understanding gender in Corbett's (1991) framework is to establish the number of genders a system has. This analysis draws on morpho-syntactic evidence from agreement rules and their domains. An agreement domain is a morpho-syntactic environment in which a noun head controller requires agreement on a target sentence element. The number of genders, following Zaliznjak (1964), is equal to the number of agreement classes, which are defined as the set of noun controllers that, in the same agreement domain, take the same agreement and produce the same agreement on the target. Tashlhiyt has three agreement rules, shown below in Table 1, all of which point to the existence of two agreement classes, and therefore two genders, namely masculine and feminine.

The morphology of these rules is fleshed out in more detail below, but we summarize some key facts here to establish the number of genders. Verbs agree in the person, number, and gender of the subject noun phrase, and exhibit a two-way gender contrast in 3.sg, 2.pl, and 3.pl (Table 1a).[1] For example, if the 3.m.sg prefix *i-* is used instead of *t-* to form *i-lsa* in Table 1a (second example), the result is ungrammatical. Though the morphology is less transparent, determiners (i.e., indefinites, cardinal numbers, and ordinal numbers) agree in number and

[1] We use the following abbreviations for the grammatical categories in our glosses here and throughout: m = masculine, f = feminine, sg = singular, pl = plural, bs = bound state, fs = free state, pf = perfect.

gender with the head noun (Table 1b). Finally, certain "adjectives" agree in number and gender with the nouns they modify (Table 1c).[2]

Table 1. Agreement rules in Tashlhiyt (domain): inflectional categories affected

a. Subject-verb agreement (clause): person, number, gender

i-swa	ufruχ	aman	t-lsa	t-fruχ-t	!t-azrzi-t
3m.sg-drink:pf	boy:ms.bs	water	3f.sg-wear:pf	girl:f.sg.bs	fibula
'The boy drank the water'			'The girl wore the fibula'		

b. Noun-determiner agreement (determiner phrase): number, gender

Masculine		Feminine	
yan urgaz	'one/a man'	yat t-mʁar-t	'one/a woman'
sin irgazn	'two men'	snat t-mʁarin	'two women'
yan d-mraw n irgazn	'eleven men'	yan d-mraw-t n t-mʁarin	'eleven women'
argaz wiss-sin	'second man'	t-amʁar-t tiss-snat	'second woman'

c. Noun-adjective agreement (noun phrase, clause): number, gender

Masculine		Feminine	
addal azgzaw	'blue/green shawl'	t-aggur-t t-azgzaw-t	'blue/green door'
iddula izgzaw-n	'blue/green shawls'	t-iggura t-izgzaw-in	'blue/green doors'
imqqur ufruχ-ad	'this boy is tall'	t-mqqur t-fruχ-t ad	'this girl is tall'

As explained below, masculine and feminine gender marking is highly consistent. Feminine controllers are marked with the consonant *t*, either as a prefix or the combination of a *t*-prefix and *-t* suffix, but masculine controllers in general lack this morphology. These marks result in similarly consistent marking of targets, which again involve regular use of *t* affixes (feminine) and lack them in masculine word forms. These two agreement classes support the long-held view that Tashlhiyt has two genders.

3.3 Agreement morphology

How is gender marked in controller and target wordforms? To address this question, we require some background on noun structure and certain inflectional rules that apply to nouns, which we review here before documenting gender marking in nouns, verbs, adjectives, and determiners. For more detailed accounts, see Dell and Jebbour (1991); Dell and Jebbour (1995) and Dell and Elmedlaoui (2002).

Nouns are inflected for number, gender, and state in Tashlhiyt. Noun paradigms that exhibit the full range of contrasts have eight wordforms by permuting two numbers (singular, plural), two genders (masculine, feminine), and two states (free and bound), as shown below in Table 2.

[2] Tashlhiyt has but a small number of true adjectives, such as the color terms shown here, and many words that act as adjectives in languages like French are derived from verbs (Aspinion 1953). However, the fact that these "adjectives" participate in agreement is not in doubt, as even some verbal expressions agree in number and gender with their noun controllers.

Table 2. Neutral noun paradigm, frux 'boy/girl'

	Masculine		Feminine	
	free	bound	free	bound
sg	a-frux	u-frux	t-a-frux-t	t-frux-t
pl	i-frxa-n	i-frxa-n	t-i-frx-in	t-frx-in

Bound state forms (*état d'annexion*), referred to as construct state forms in many Afro-Asiatic languages, occur after verbs, most prepositions, conjunctions, determinative complements, and cardinal numbers (Aspinion 1953), while free state (*état libre*) forms are the elsewhere class. Plural formation in Tashlhiyt is not fully understood, but the available accounts document a complex system similar to Arabic, with so-called sound plurals that have external affixation, 'broken plurals' with internal stem changes, mixed plurals that combine both external and internal operations, and a special class of *id*-plurals (Bensoukas 2018; Bensoukas 2019; Dell & Jebbour 1995; Idrissi 2000; Jebbour 1988; Saib 1986).

There is a long and intriguing debate about the nature of morphological bases in Amazigh languages. This debate addresses questions like if Tashlhiyt has consonantal roots such as some have argued for Arabic, or if vowels are part of roots or stems (Chaker 1990; El Hamdi 2018; Galand 1988; Lahrouchi 2008; Lahrouchi 2010); see Boumalk and Bensoukas (2018) review and contemporary perspective. Since we do not need to weigh into this debate to analyze gender, we instead use the concept of a noun theme (Dell & Jebbour 1991) to define bases. Similar to the notion of stem in inflectional morphology, a theme is a word minus the inflectional affixes. Thus, *frux* is the singular theme in Table 2 for both genders, and *frxa* is the masculine plural theme, exhibiting internal changes.

The phonological structure of nouns is important in signalling gender. A broad generalization, with specific exceptions, is that free state masculine wordforms begin with a vowel, while corresponding feminine forms begin with a consonant, usually *t*. It is generally assumed that this generalization is the result of grammatical rules, and not emergent from the phonological structure of themes (see Guerssel (1983), Dell and Jebbour (1991), and Dell and Jebbour (1995), based on original ideas from Basset (1932), Basset (1952), and Basset and Picard (1948)). Thus, Dell and Jebbour (1995) formalize this assumption with a set of rules that enforce the gender-based phonological generalization (Table 3).[3] In particular, nouns with C-initial themes receive an initial vowel by the R-AUG rule (unless exceptionally marked in the lexicon), which then feeds *t-* prefixation in feminine nouns. Thus, masculine nouns are generally vowel-initial because of the broad application of R-AUG and the assumption that PRX-T does not apply to them.

[3] To avoid confusion, we use the same rules and rule labels as those developed in Dell and Jebbour (1995), and then extend them to known rules with similar functions in the inflectional morphology. Thus, their R-AUG rule is the rule inserting a vowel augment in nouns; PRF-T is the rule that works on the output of R-AUG and inserts a *t-* prefix in feminine nouns; SFX-T inserts a *-t* suffix in feminine plurals. PRF-I formalizes *i-* prefixation in most plurals, and PRF-U for bound state masculines. V-DEL accounts for the default condition in which bound state feminine nouns delete the vowel after the *t-* prefix and produce a consonant cluster, as shown for *dʁar* in Table 4.

Table 3. Gender, number, and state morphology (after Dell & Jebbour 1995)

	Rule	Conditions
R-AUG	∅ → V / [Noun ___ C	1) V filled in as *a* by default, but also *i* with lexical specification 2) -AG nouns block R-AUG
PRX-T	∅ → t / [Noun[+fem] ___ V	Ordered after R-AUG
SFX-T	∅ → t / ___]Noun[+fem, -pl]	
PFX-I	∅ → i / [Noun[+pl] ___ C	Applies to most native themes, but not lexically marked C-initial themes, like Arabic loans
PFX-U	∅ → U / [Noun[+bs, +ms] ___	Prefixes high vocoid to all bound state masculine nouns, realized as *u/w* before *a* and *u* and *i/j* before *i*; its syllabic role is determined by its syllable position
V-DEL	V → ∅ / [Noun[+bs, +fem] ___ C	Deletes vowel inserted by R-AUG or PFX-I in all bound state feminine forms

The impact of these rules, and other inflectional rules, can be illustrated by comparing nouns with consonant-initial vs. vowel-initial themes (Table 4). Nouns formed with a consonant-initial theme, such as *dʁar* (‘flap, *lambeau*’) in (a), receive the augment *a-* in ms.sg.fr, which in turn feeds PRX-T in the feminine forms. However, nouns with vowel-initial themes, like *adgal* (‘widower, *veuf’)* in (b)*,* are consistent with the vowel-initial generalization without R-AUG, but they nonetheless trigger in PRX-T in feminine nouns. The effect of the theme-initial segment is also observed in the larger paradigm: consonant-initial *dʁar* receives the plural prefix *i-* (which supplants the *a-* augment) in plurals, but vowel-initial *adgal* does not. Likewise, the bound state prefix *u-* inserted by PFX-U is glided in syllable initial position before vowel-initial *adgal* in (b), whereas *u-* causes elision of the vowel augment in (a) and so is realized as a full vowel. A final overt difference is that consonant-initial stems trigger V-DEL in bound state forms, which eliminates either the *a-* augment or the *i-* plural prefix, but vowel-initial themes are not affected by V-DEL because this rule targets consonant-initial forms. In sum, the larger rule system for inflectional morphology enforces the broad pattern that free state masculine forms are vowel-initial, while all feminine forms are consonant-initial and generally start with *t*.

Table 4. Marking gender, three basic noun classes (free state)

	Masculine				Feminine			
	free		bound		free		bound	
theme	sg	pl	sg	pl	sg	pl	sg	pl
a. dʁar	a-dʁar	i-dʁar-n	u-dʁar	i-dʁar-n	t-a-dʁar-t	t-i-dʁar-in	t-dʁar-t	t-dʁar-in
b. adgal	adgal	adgal-n	w-adgal	w-adgal-n	t-adgal-t	t-adgal-in	t-adgal-t	t-adgal-in

Exceptions to the generalization ‘vowel-initial = masculine’, which must therefore be specified [-AG] in the lexicon, include many Arabic loans which either begin with *l* or a coronal geminate, as in *lbrrad* ‘tea pot, *théière*’. So-called *bu*-nouns, which mark plurals with the proclitic *id*, are another systematic exception, as in *buttgra* ‘tortoise, *tortue*’ (Bensoukas 2015a; Bensoukas 2015b), and also an exception to the generalization that feminine nouns start with *t* (the feminine variants of this class start with *mm(u)-*). Finally, a handful of irregular feminine nouns, like *immi* ‘my mother’, are exceptional in that they are vowel-initial and they do not have the *t-* prefix.

As for agreement targets, verbs mark gender with a remarkably consistent paradigm rule, contrasting masculine and feminine forms in 3.sg, 2.pl, and 3.pl, as shown below for two verbs in Table 5. An analysis of the feminine-agreeing verbs cannot simply re-use Dell and Jebbour's (1995) rules for nouns (Table 3) because the form and grammatical conditions are not the same. However, the 3.sg.fem and 2.pl.fem forms have a *t-* prefix, and 2.pl and 3.pl feminine forms have a *-t* suffix, suggesting that the same basic operations are recycled somehow in the grammar (see Zwicky (1988)).

Table 5. Gender marking in verbs

	Paradigm rule	'remember (perfect)'	'wear (perfect)'
1.s	X-χ	kʷti-χ	lsi-χ
2.s	t-X-t	t-kʷti-t	t-lsi-t
3.m.s	i-X	i-kʷti	i-lsa
3.f.s	t-X	t-kʷti	t-lsa
1.p	n-X	n-kʷti	n-lsa
2.m.p	t-X-m	t-kʷti-m	t-lsa-m
2.f.p	t-X-mt	t-kʷti-mt	t-lsa-mt
3.m.p	X-n	kʷti-n	lsa-n
3.f.p	X-nt	kʷti-nt	lsa-nt

"Adjectives" and determiners also re-use affixes with *t*, as shown in Table 1a/b, though the existence of *t* in closed class items such as *yat* is a pattern of suppletive allomorphy consistent with *t* = feminine, rather than an outcome of a suffixation rule.

3.4 Gender in noun paradigms

Gender is realized in both full and partial paradigms, and this fact is relevant to gender assignment because paradigm structure can help predict gender. Following Payne (1998), we employ a three-way contrast for how paradigm structure relates to gender: fixed, variable, and neutral gender. Fixed genders only have wordforms inflected for one gender category, masculine or feminine. As shown in Table 6, because fixed gender nouns do not contrast in gender, they have just four or two wordforms. There is always a contrast in free/bound state, but it is sometimes the case that fixed gender forms only have singular or plural wordforms, as *χllu* 'demolition' below, which results in just two wordforms.

Table 6. Fixed gender in partial paradigms

		Masculine		Feminine	
		Free state	Bound State	Free State	Bound State
a. Masc-Fixed	sg	abʒwaj	ubʒwaj		
bʒwaj, 'delirium'	pl	ibʒwajn	ibʒwajn		
b. Fem-Fixed	sg			tamilla	tmilla
milla, 'dove'	pl			timalliwin	tmalliwin
c. Masc.sg-Fixed	sg	aχllu	uχllu		
χllu, 'demolition'					

In fixed gender nouns, gender assignment is clear because there is only one possibility. This is not the case in variable and neutral gender nouns. Variable gender nouns, as illustrated in Table 7, have both masculine and feminine wordforms, but one of them is predictable from secondary morphology. For example, the masculine free-state singular form *aʁnzur* expresses the basic level concept of 'beak', and the corresponding feminine form is a diminutive of this concept: 'small beak'. Interestingly, secondary morphology can also predict masculine gender, because basic level concepts expressed in the feminine have a corresponding masculine augmentative marked by the absence of feminine morphology, as in Table 7b, meaning 'large plug'. Though there are some special case exceptions with loans, variable gender nouns may have either diminutive or augmentative forms, but not both, because the gender markers have been used up. For example, expression of the concept 'big beak' is periphrastic: *aʁnzur mqurn* 'beak (that is) large', as opposed to the diminutive using regular feminine morphology: *taʁnzurt*. As a result, variable gender nouns can be assigned an intrinsic gender for the core concept, as in the Masculine-Variable example *aʁnzur* in Table 7a, because the other forms with the opposite gender are predictable from secondary morphology. As discussed below in section 4.1.2., secondary morphology may also contribute a unitizing meaning and polysemy (see Aspinion (1953) for more detailed description and illustration).

There are also full noun paradigms that are not predictable in this way. So-called neutral-gender nouns have both masculine and feminine morphology. These gender marks are interpretable in the sense that they reflect the intended sex/gender of the referent expressed by the noun. Thus, *amksa* in Table 7c is used to refer to a male shepherd, *tamksat* for a shepherdess. While there are a handful of 11 neutral-gender inanimate nouns that have a somewhat opaque notion of gender, the vast majority (97%) are animate and gender marking has a transparent interpretation in nouns describing humans and animals.

Table 7. Variable and neutral gender in full paradigms

		Masculine		Feminine	
		Free state	Bound State	Free State	Bound State
a. Masc-Variable	sg	aʁnzur	uʁnzur	taʁnzurt	tʁnzurt
ʁnzur, 'beak/DIM'	pl	iʁʷnzar	iʁʷnzar	tiʁʷnzar	tʁʷnzar
b. Fem-Variable	sg	asaqqun	usaqqun	tasaqqunt	tsaqqunt
saqqun, 'plug/AUG'	pl	isuqqan	isuqqan	tisuqqan	tsuqqan
c. Neutral	sg	amksa	umksa	tamksat	tmksat
mksa, 'shepherd(ess)'	pl	imksawn	imksawn	timksawin	tmksawin

The use of gender secondary morphology, like diminutives and augmentatives, can be considered inflectional in nature because the use of one value precludes expression of the opposite value, a common characteristic of inflectional morphology (Aronoff & Fuhrhop 2002). Gender is also reflected in word derivation that uses affixes and process morphology distinct from gender morphology. Though the correlations between word derivation and gender are not categorical, as they are with diminutives and augmentatives, there are some statistical trends that are striking and suggest that at least some derivational categories are predictive of gender. For example, so-called *Tirrugza* nouns are formed with a CCuCCa CV-skeletal template, as in *tinffulsa* 'guarding', and they have a strong tendency to be feminine (29 of 31 paradigms).

Agentive nouns likewise have a greater-than-chance tendency to have masculine gender (98% occurrence compared to a chance rate of 75%). Because of these trends, we have examined all noun paradigms for the existence of 12 derivational categories and used them as potential predictors in our quantitative analysis below.

The relative frequencies of fixed, variable, and neutral gender noun paradigms are shown in Table 8, cross-classified by gender. Neutral gender is not a gender itself, but a pattern within a paradigm, so we must assign it a gender to analyze neutral nouns. We can assign masculine gender to neutral gender nouns because the masculine is the 'default' gender (see evidence below), and feminine gender is expressed by adding feminine morphology. This basic assumption is widespread in prior studies (see section 3.1), and it is consistent with the native speaker intuitions of the fourth author. As shown in the column totals in Table 8, masculine nouns are statistically more prevalent than feminine nouns.

Table 8. Fixed-variable-neutral by gender

	Masculine	Feminine	Total (row %)
Fixed	762	423	1185 (61.91%)
Variable	354	60	414 (21.63%)
Neutral	315	NA	315 (16.45%))
Total (column %)	1431 (74.76%)	483 (25.24%)	1914

Fixed gender nouns are by far the most common, followed by variable gender nouns, and then neutral gender nouns. This two-category gender system is used in our quantitative analysis below. That is, it is the dependent variable we try to predict from other linguistic attributes.

3.5 Masculine as a gender default

As mentioned above, masculine gender is the default gender in a number of senses. It is far more frequent than feminine gender, constituting approximately 75% of all nouns and 70% of fixed and variable gender nouns (excluding neutral gender). It is also the default in the sense that masculine gender is assigned to words with no overt gender morphology, and feminine gender arises from the application of PXF-T and SXF-T (Table 3). Finally, the vast majority of loans into Tashlhiyt come from Arabic, and there is a strong tendency for them to be adapted first as masculine, as shown in Table 9. Native words closely match the statistical trends of the whole lexicon, but unassimilated loans (i.e., loans that have not been fully adapted to Tashlhiyt morphology), such as *lbni* 'construction', have a much higher occurrence of masculine nouns relative to the other two categories (including neutral nouns). If masculine is the default gender, assigned when a noun lacks other information, then this trend is expected.

Table 9. Three-way gender classes by loanword status

	Masculine	Feminine	Neutral (Masculine)
Arabic-unassimilated	141 (77.90%)	36 (19.89%)	4 (2.21%)
Arabic-assimilated	109 (38.52%)	80 (28.27%)	94 (33.22%)
Native	825 (59.78%)	342 (24.78%)	213 (15.43%)
Total	1075 (58.30%)	458 (24.84%)	311 (16.87%)

4. Quantitative analysis of gender assignment

We start by exploring the data organized by subdomain to look for correlated variables and get a general sense of the predictive value of individual variables (or features). Non-correlated variables are then used as predictors in our classification models, with the overall aim of classifying Tashlhiyt gender within traditional typology and assessing its overall predictability.

4.1 Descriptive statistics

4.1.1 Semantic factors

Semantic factors include humanness (i.e., whether a noun refers to a human), animacy, sex/gender distinctions if relevant (unspecified for inanimate nouns), and semantic field. Humanness and animacy show the same basic pattern (Table 10 and Table 11): a bias for masculine gender in human and animate nouns. Among animate nouns, only one in twenty nouns is feminine, and that disparity is even stronger in nouns referring to humans.

Table 10. Cross-tabulation by Humanness and gender

	Feminine	Masculine
Not Human	474 (24.76%)	1118 (58.41%)
Human	9 (0.47%)	313 (16.35%)

Table 11. Cross-tabulation by Animacy and gender

	Feminine	Masculine
Inanimate	459 (23.98%)	964 (50.37%)
Animate	24 (1.25%)	467 (24.40%)

Another salient fact in our corpus is that sex/gender distinctions are rarely included in a noun's meaning. In animate nouns, only a small handful can be distinguished as male or female, and the rest are unspecified (Table 12). Thus, while the values 'female' and 'male' perfectly predict gender in nouns with sex differentiation, sex/gender is unlikely to be predictive of gender in the larger data set because so few nouns have this distinction.

Table 12. Cross-tabulation by Sex/gender and gender in animate nouns

	Feminine	Masculine
Female	10 (2.04%)	0
Male	0	27 (5.50%)
Unspecified	14 (2.85%)	440 (89.61%)

Finally, we have investigated the impact of a noun's semantic field on gender, but the cross-classification of fields by gender does not reveal any unusual distributions. The percentage occurrence of masculine nouns ranges between 87% and 61% in the 22 fields, which is centered on the 75% chance occurrence in the larger corpus and thus not skewed towards masculine or feminine in any field (see the GitHub repository for more details). To summarize, animacy and humanness appear to be predictive of gender, but sex/gender is rarely specified in nouns, and semantic field does not seem to be predictive of gender.

4.1.2 Morphological factors

At first blush, gender in Tashlhiyt seems inherently morphological. Gender is highly correlated with the C/V status of the initial segment of noun wordforms (vowel-initial = masculine,

consonant-initial = feminine), and this CV structure is the outcome of a system of inflectional rules (Table 3). As fleshed out in 3.3, the themes of masculine free state forms typically begin with a vowel, or they are provided one by R-AUG, which then forms the masculine singular free state form. PFX-T then posits a *t* in feminine nouns, producing an opposition between vowel-initial masculine and *t*-initial feminine inflected forms. True, there are many C-initial words, both masculine and feminine, that run counter to these trends. But even when these are included in the larger system, this basic C/V generalization (resulting from morphological rules) accounts for a surprising amount of the data at the level of wordforms.

However, the fact that this generalization arises from morphological processes does not mean that the morphology predicts grammatical gender. There is a strong correlation between gender and the phonological form produced by these morphological rules, but the rules themselves are not triggered by morphological attributes at the level of the theme (or the lemma level). In fixed gender nouns, for example, the CV-structure of the theme of a feminine noun can be identical to the theme of a masculine noun, but they differ in gender. The same problem arises in variable gender nouns because the basic gender (i.e., the opposite gender of that reflected in secondary morphology) is not predicted by the form of the word or the theme.

We can contrast this scenario with the role of morphology in other languages. In Russian, for example, an inherently morphological attribute, declension class, together with a semantic core, predicts gender (Corbett 1982). Tashlhiyt does not have declension classes, but it does have other idiosyncratic classes and morphological structure that could play a role in predicting gender, which we explore in the tables below. In particular, the R-Aug vowel, or lack of a vowel augment, is morphological because it represents idiosyncratic classes associated with themes, somewhat like theme vowels in Indo-European languages. While not categorical in nature, as shown in Table 13, the distribution of vowel augments does seem to be predictive of gender, with greater than chance (n > 75%) occurrence of *a* in masculines and the opposite pattern for *i* and zero R-Aug vowel.

Table 13. Cross-tabulation by R-Aug vowel and gender

	Feminine	Masculine
A	148 (14.18%)	896 (85.82%)
I	73 (42.44%)	99 (57.56%)
Zero	262 (37.54%)	436 (62.46%)
Totals	483 (25.24%)	1431 (74.76%)

The plural pattern is also an inherently morphological attribute and lexically idiosyncratic, and in Table 14 we document some trends favoring masculine gender: external and mixed plurals.

Table 14. Cross-tabulation by plural pattern and gender

	Feminine	Masculine
External	66 (13.17%)	435 (86.83%)
Internal	96 (25.33%)	283 (74.67%)
Mixed	44 (14.38%)	262 (85.62%)
Totals	206 (17.37%)	980 (82.63%)

As alluded to above, derivational morphology can be predictive, with some categories like agentives and *Tirrugza* nouns showing near categorical patterns. Table 15 shows the

principal derivational categories, including underived nouns, excluding some categories with counts less than 10 (though these can be explored in the data on the GitHub page).

Table 15. Cross-tabulation by derivational category and gender

	Feminine	Masculine
Action nouns	115 (33.63%)	227 (66.37%)
Agentives	3 (1.69%)	174 (98.31%)
Instrumental nouns	15 (33.33%)	30 (66.67%)
Stative nouns	5 (45.45%)	6 (54.55%)
Tirrugza nouns	29 (93.55%)	2 (6.45%)
Underived	302 (23.97%)	958 (76.03%)
Totals	469 (25.13%)	1397 (74.87%)

Secondary morphology is marked on 415 forms in our corpus (as in the examples in Table 7a and b), and values for the categories shown in Table 16 are categorical predictors of gender under the assumptions discussed in section 3.4.

Table 16. Cross-tabulation by secondary morphology and gender

	Feminine	Masculine
Diminutives	0	272
Polysemous	0	50
Unitizing	0	33
Augmentative	60	0
No secondary morphology	423 (28.22%)	1076 (71.78%)
Totals	483 (25.24%)	1431 (74.76%)

Finally, we include loanword status and source as a single factor and treat it as morphological because it is a gage of how well a word is incorporated into the morphological system. As shown above in Table 9, gender seems to be affected by the degree of nativization, with nativized nouns having a higher percentage of feminine gender.

To summarize, we document here four morphological attributes that have not yet been investigated in relation to gender assignment and found that some (R-Aug vowel and plural pattern) appear to be predictive of gender with above-chance patterns, while others (derivation and secondary morphology) exhibit (near) categorical predictors of gender when they are specified. Morphology, therefore, seems to have potential for predicting gender in ways that are completely separate from the inflection rules that link gender to wordforms in past accounts.

4.1.3 Phonological factors

Phonological factors analyze the phonological structure of the theme, including initial/final segments, CV-structure, and theme length. Table 17 below gives the distribution of initial and final segments cross-classified by gender. These counts come from singular themes, and so do not include the 50 nouns that only have plural words. The consonants listed here do not distinguish consonant length, so these segment types only express generalizations for place, manner, and voicing of segments. Consonants are also not distinguished for emphasis, which is a morpheme level contrast and so not appropriate at the segment level. Examination of the % Masc column reveals some strong deviations from the 25/75% chance rates, including high rates of initial vowels in feminine nouns, as well as exceedingly high rates of final *a#, t#.* On the other

hand, initials strongly associated with masculine gender include: *#b, #k, #q, #s, #ʃ, #ʒ, #χ, #ħ, #h, #ʕ*, as are the following finals: *b#, d#, k#, g#, z#, l#.*

Table 17. Distribution of initial and final segments in theme

	Initial				Final		
Segment	Feminine	Masculine	% Masc		Feminine	Masculine	% Masc
a	73	94	56.29		139	57	29.08
i	35	29	45.31		97	156	61.66
u	31	34	52.31		20	104	83.87
b	8	90	91.84		2	36	94.74
t	14	25	64.10		14	17	54.84
d	14	47	77.05		6	112	94.92
k	5	44	89.80		3	19	86.36
g	21	73	77.66		4	36	90.00
q	4	26	86.67		4	12	75.00
f	17	54	76.06		11	50	81.97
s	26	175	87.06		14	83	85.57
z	16	63	79.75		6	53	89.83
ʃ	8	48	85.71		7	31	81.58
ʒ	2	20	90.91		2	10	83.33
χ	2	21	91.30		1	5	83.33
ʁ	12	33	73.33		7	25	78.13
ħ	2	37	94.87		3	16	84.21
h	1	8	88.89		0	2	100.00
ʕ	1	15	93.75		5	18	78.26
m	48	156	76.47		16	87	84.47
n	21	74	77.89		17	74	81.32
l	79	175	68.90		15	122	89.05
r	13	39	75.00		42	178	80.91
w	12	18	60.00		5	21	80.77
j	1	0	0.00		26	74	74.00

The claim that the themes of feminine nouns seem to be associated with initial or final vowels is explored in Table 18, which abstracts over these segments and groups them by C or V class. We see here that the percentage of masculine nouns with V-initial or V-final approaches 50%, which is far below the expected 75%, suggesting that these vowels are associated with feminine nouns. Likewise, the percentages of themes with initial or final vowels (bottom row) are twice as large for feminine as opposed to masculine nouns.

Table 18. Distribution of initial C and V structure in theme

	Initial				Final		
	Feminine	Masculine	% Masc		Feminine	Masculine	% Masc
C	327	1241	79.15		210	1081	83.73
V	139	157	53.04		256	317	55.32
% V	29.83	11.23			54.94	22.68	

The finding that feminine nouns are associated with theme initial and final vowels is interesting in two respects. First, this pattern is somewhat unexpected, given the way the morphology of gender enforces the opposite generalization in wordforms – masculine singular free state nouns are generally vowel-initial (see section 3). Second, a great deal of research into the stem phonology has shown that the CV structure of theme-initial positions (El Hamdi 2018), as well as theme-final positions (Bensoukas 2001), are important structures for predicting regular phonological processes. The fact that they appear to be important in the gender system as well is one more piece of evidence that themes can be specified for vowel structure and not simply discontinuous consonant structure.

Finally, we report on the overall length of themes, cross-classified by gender, in Table 19. The mean segment length is 4.7 segments in feminine nouns and 4.65 in masculine nouns, though the median length for masculine themes (length=5) has a much higher count relative to similar-sized themes, so it is an open question whether segment length is predictive of gender.

Table 19. Cross-tabulation of theme length by gender

Length	Feminine	Masculine
2	3	8
3	40	103
4	153	433
5	152	598
6	66	196
7	44	52
8	5	6
9	2	1
10	1	
11		1

To summarize, we have again ventured into uncharted territory by correlating theme phonological structure with gender and found important roles for initial and final segment and CV structure, though it is unclear whether segment length is predictive.

4.1.4 Summary and predictability with categorical rules

We have examined several features in three different domains and, by cross-tabulating them with gender, found some suggestive features. Semantic features do not seem to be very predictive, with the possible exception of animacy. This is due to the fact that some aspects of meaning are not specified for many nouns (sex/gender), and semantic fields, while fully specified, do not seem to have asymmetric patterns favoring a particular gender category. We also looked at a number of form-based features and we have uncovered some suggestive trends involving

lexicalized attributes like the R-Aug vowel and the plural pattern, as well as some highly predictive patterns with specified secondary and derivational morphology. Phonological form also provided some useful insights, with feminine nouns associated with vowel structures and disassociated with some consonants. We use all of these factors in our computational analysis below and employ machine learning techniques to flesh out their relative importance.

Before turning to the computational modeling, however, it is useful to motivate the research by considering how well the trends above predict gender as categorical rules. That is, given that some of the patterns are categorical patterns (e.g., secondary morphology), how well do these rules predict gender? Many past analyses use this approach to assess predictability, either explicitly or implicitly. For example, Tucker et al. (1977) propose a set of form-based conditions and assess them against a large data set of French nouns. Many of the conditions are later refined, and rules of a near categorical nature can be used to tabulate the percentage of the data accurately accounted for. Kramer (2014) uses a similar method of tabulating results for gender assignment in Amharic, given the limited lexical resources available, and uses these counts to support the contention that feminine nouns are extremely under-represented in the language.

We can take a similar approach to assess gender in Tashlhiyt, but such an analysis does not meet our expectations of the degree of predictability (i.e., 85% or higher). As shown in Table 20, specified sex/gender alone accounts for less than 2% of the data. By subtracting these data and using the residue to create a new baseline, we can also see that variable gender nouns (predictable from secondary morphology) account for an additional 22%, and, using the same procedure, derivational categories account for another 12%. Neutral nouns are not obviously predictable from surface attributes, but if we say that their masculine status is due to the existence of a gender contrast without secondary morphology, then these nouns account for another 24% of the data, coming to a cumulative total of 49% of the corpus.

Table 20. Data explained using (near) categorical rules

Residue	Data explained by factor	Cumulative total % explained
	Sex/gender: 37/1914 (1.93%)	1.93% (37/1914)
1877	Variable: 413/1877 (22%)	23.51% (450/1914)
1464	Derivation: 175/1464 (11.95%)	32.65% (625/1914)
1289	Neutral: 313/1289 (24.28%)	49% (938/1914)

In sum, if we try to tabulate all of the data points accounted for with categorical rules, using rather generous inclusion criteria (e.g., neutral nouns), we come up far short of our expectations for gender predictability, with more than half of the data unaccounted for. We need, it seems, mechanisms that can make predictions based on statistical tendencies, like vowel occurrence in themes, rather than categorical or near categorical rules. In the next section, we develop a set of computational classifiers to address this problem.

4.2 Predicting gender with supervised learning

We flesh out in more detail now the specific assumptions, methods, and results of our classifications. While space limitations prevent us from delving into all of the formal details, we document every decision about our models and conclusions, including some specific reports only available from the GitHub page.

4.2.1 Assumptions and methods

We frame the classification problem as a challenge of predicting intrinsic noun gender from a set of 15 features derived from the linguistic factors discussed in section 4.1. The features shown in Table 21 list the 13 factors discussed above and include two additional features, countC and countV, which give counts of consonants and vowels, respectively, to gage the length of theme with segment counts by C/V class.

Table 21. Features available for predicting gender, showing subclass and number of categories for each feature.

Domain	Feature	# Categories
Semantic	s_humanYN	2
	s_animateYN	2
	(s_sexGender)	4
	s_semanticField	22
Morphological	m_rAugVowel	3
	m_pluralPattern	6
	m_derivationalCat	14
	m_secondaryMorph	5
	m_loanwordSource	6
Phonological	p_start	25
	p_end	25
	p_startCorV	2
	p_endCorV	2
	p_lengthC	8
	p_lengthV	5

As discussed above, the data set is moderately imbalanced because roughly 75% of nouns are masculine. With almost 500 feminine nouns, we are not concerned that our models can learn to classify them, but it is prudent to employ methods that can handle the data imbalance. Our approach to this problem is to use classifiers and measures of success that are known to be robust to data imbalances. Thus, we included Random Forest, Gradient Boosting, and Vector Support Machine classifiers, which, because of the way they sample or represent the data and their parameters for fine-tuning, reduce the impact of the majority gender class (He & Garcia 2009). Our principal metric for evaluating the models, AUC-ROC (discussed below), is also insensitive to data imbalances (Hastie et al. 2009).

The above features were encoded for data analysis as follows. Aside from the two features, countC and countV, the features had categorical variables, and the values of these features were converted to integers using the LabelEncoder function from the sklearn.preprocessing module. We prefer this approach to one-hot encoding (which creates a feature for every value of every factor) because it drastically reduces the number of dimensions in the model (114 features for one-hot, compared to 14 for the feature set used below) and avoids problems in assessing feature correlations. However, we have tested the models discussed below using one-hot encoding and found that the results are essentially the same, except for the performance of the Support Vector Machine models, which does improve considerably with this encoding scheme. Given these findings, we believe the simpler model with fewer features is

sufficient for larger aim of comparing predictability across classes (i.e., semantic, morphological, and phonological).

With these features and this encoding scheme, we did a correlation analysis to look for highly correlated features that can be excluded from the models. In particular, we calculated the Pearson correlation of each pair of features and set a threshold of .8 for features to exclude. We found that the semantic feature of s_sexGender was negatively correlated with both s_humanYN (-0.73) and s_animateYN (-0.93), surpassing the threshold with the latter. Since s_sexGender was meaningfully specified for only 37 nouns, it was excluded. This leaves 14 features for predicting gender.

To investigate gender predictability, we created several models using five different classifiers that can handle high dimensional data sets and extract complex relationships between target values (gender) and features (linguistic factors). These models (see Table 22) cover a range of complexity, with Support Vector Machines being the simplest, and the Ensemble Voting Classifier being most comprehensive. The Ensemble Voting Classifier consisted of Random Forest, Gradient Boosting, and Adaboosting classifiers with soft aggregation.

Table 22. Classification models used (with label used in later graphics)

rfc = Random Forest Classifier (Ho 1995) Ensemble method that builds and combines multiple decision trees to improve accuracy.
svc = Support Vector Machine Classifier (Cortes & Vapnik 1995) Model that finds the optimal hyperplane to separate data into different classes.
gbd = Gradient Boosting Classifier (Mason et al. 1999) Ensemble method that builds decision trees sequentially, each improving on the last, to improve accuracy
abc = Adaboost Classifier (Freund & Schapire 1995) Ensemble method that builds strong classifiers from weak ones that previously produced errors.
evc = Ensemble Voting Classifier Ensemble method that combines predictions of multiple classifiers by voting.

We have avoided using deep learning neural network classifiers because the relatively small data set makes it hard to train these models sufficiently. In addition, we have explored two additional models, Logistic Regression and Categorical Naïve Bayes classifiers, for data analysis purposes (e.g., probe effect of feature encoding) and found that they make predictions very similar to the ones in Table 22. An important point is that our goal is not to find and then fine-tune the best possible classifier for predicting gender. Rather, we aim to create a profile of predictions from multiple sensible models, and then use this profile to tease apart the role of different linguistic factors. We believe that this collection of models is sufficient for this task.

We report all standard metrics in evaluating models below (i.e., accuracy, precision, recall, F1) in the appendix and the complete reports on the GitHub page. However, in the discussion below we focus on one metric, AUC-ROC (Area Under the Receiving Operating Characteristic Curve), which is a commonly used measure in binary classification tasks (Spackman 1989). In a nutshell, the AUC-ROC score is a summarization of confusion matrices

(for positive/negative outcomes) at different threshold values. It gives a single scalar that quantifies the overall performance of the model. A higher AUC-ROC score indicates better performance, where a score of 0.5. indicates random guessing and 1.0 is perfect classification. As discussed above, the AUC-ROC score is insensitive to data imbalance, making it suitable for our data. We also report in the appendix the Brier Loss score for each model, which covers the main limitation of the AUC-ROC score, namely that it is independent of the probability magnitude values.

We used a standard procedure of 5-fold cross-validation training repeated 20 times. Thus, the data were split into five parts, and the model was trained on four parts and tested each time on the remaining fifth part. All metrics were computed on the test data, then stored for individual tests, and the reported metrics were the mean of 20 such trials. To probe different classes of features, we tested our models with seven different feature sets (see Table 21): phonological, morphological, semantic, phonological and morphological, phonological and semantic, morphological and semantic, and all 14 features. By distinguishing models by linguistic factors, we can assess the predictive power of these distinct feature sets.

4.2.2 Results

The results depicted below investigate the predictability of noun gender relative to these linguistic factors. The relative AUC-ROC scores for each classifier are shown in Figure 1, given the feature subsets associated with the three linguistic domains. This first assessment uses the default parameters for all classifiers provided by the sklearn module, but we also examine hyper-parameterized models (i.e., models with optimized parameters) below.

If we evaluate these models relative to the 85% predictability threshold from literature on gender assignment (Corbett 1991), morphology is the only single linguistic domain to reach this threshold. Ignoring the Support Vector Machine results, which consistently under-perform in all models, the AUC-ROC score for the morphology subset is .85, which out-performs both the phonology (.76) and semantic (.69) subsets by wide margins.[4] The weak predictive power of semantic features is underscored by the fact that adding semantic features to either morphological or phonological features only marginally improves performance, whereas the combined morphology + phonology feature model (i.e., both 'form' domains) makes a significant gain of the basic single-domain models, with a mean AUC-ROC score of .89.

[4] The use of the AUC-ROC metric here does not directly relate to the 85% predictability benchmark in the literature, which corresponds to Accuracy. However, it is standard to use AUC-ROC scores in classification tasks such as ours to address data imbalances and other issues, so it is the superior measure of performance as explained in 4.2.1. Additionally, the AUC-ROC scores are in the same range as Accuracy, so they can be used to assess general predictability, especially when making comparisons across feature subsets.

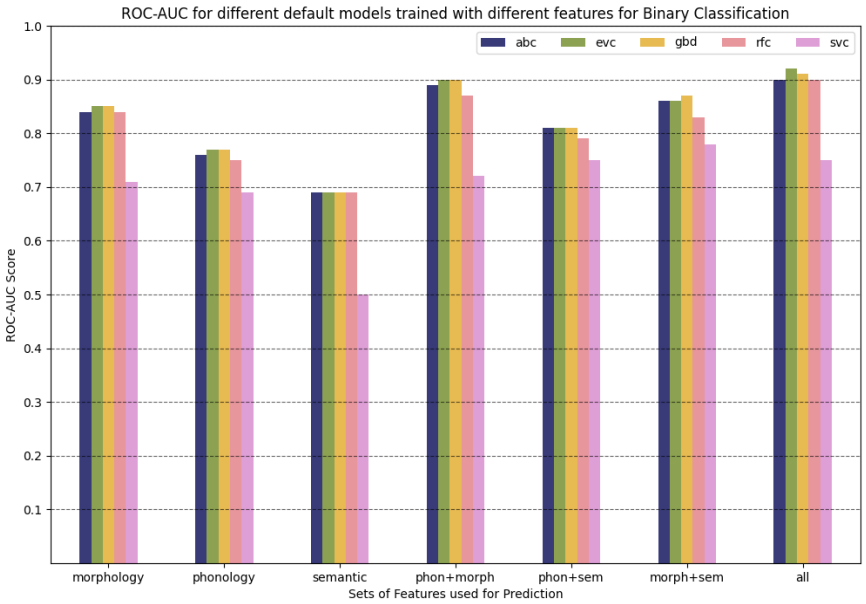

Figure 1. Results with default parameter values

A standard technique in machine learning is to improve on the performance of classifiers with hyper-parameterization tuning, that is, by searching the parameter space to find the best-performing parameters. Using standard hyper-parameterization tools from sklearn, we optimized the parameters of the models using ensemble methods (i.e., Random Forest, Adaboosting, Gradient Boost classifiers), and again investigated any differences among the seven models. The optimized parameters are reported in the appendix. As shown in Figure 2, hyper-parameterization did little to improve these models. For example, the mean AUC-ROC for the morphology model is essentially unchanged, and the combined morphology + phonology only improves from a mean of .89 (default) to .90 (hyper-parameterized). While our goal is not to find the best-performing model, since we are interested in relative performance across models, hyper-parameterization of these models validates our findings with default parameters because the same relative relationships are maintained in these slightly improved models: morphology is the best performing single-domain model, followed by phonology and then semantics. Furthermore, the combined morphology + phonology competes with the all-features model with a mean AUC-ROC value of .90.

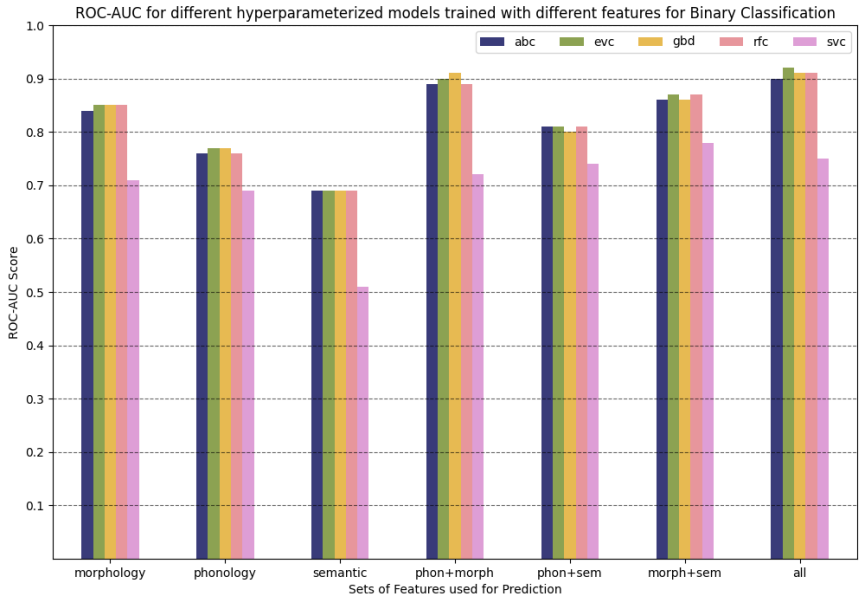

Figure 2. Results with hyperparameter tuning

We have seen above how machine learning techniques can be used to discover the linguistic domains most important to predicting gender. These models can also be used to drill down into specific features within these domains using feature importance. To get a better sense of what specific factors are important in gender assignment, we used the permutation_importance function from the sklearn.inspection module for each of the features in Table 21. This function probes feature importance by shuffling the values for one feature and assessing the impact of this data shuffling on the larger model. Thus, a greater decrease in performance means that the shuffled feature has greater importance in the model. The results of this analysis are shown in Figure 3 for each of the 14 features, aggregated by specific feature subsets (i.e., the seven models).

These results can give us additional insights into the underlying factors predicting noun gender. Starting with the features within a purely morphological model (top left panel in Figure 3), the single most predictive feature is m_secondaryMorph, followed by m_rAugVowel and m_pluralPattern. The role of secondary morphology is not a surprise, since it is categorically predictive of large numbers of nouns, especially with diminutive marking (see Table 16). The fact that the R-Aug vowel and plural pattern are more predictive than derivational morphology, however, is interesting because it tells us that, though derivational morphology has some near-categorical predictors (e.g., agentives), in the broader analysis these categories are less important than the R-Aug vowel and plural pattern of a noun. As for phonology, the size of a theme seems to be relatively unimportant to gender, but the phonological structure of initials and finals are quite important, especially p_endPhoneme. Finally, though semantics features are the least important as a class, animacy and semantic field make non-trivial contributions to predicting gender. As for combined models, most of these feature importances hold: secondary morphology, end phoneme, and animacy have the greatest importance in the all-features model (right panel), with the interesting emergence of the loanword source as a semi-important predictor.

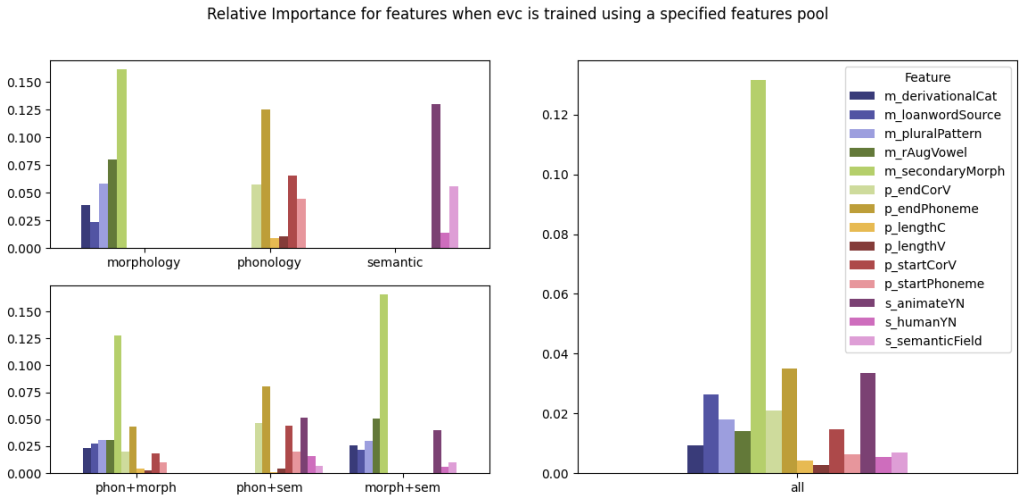

Figure 3. Feature importance by model for ensemble voting classifier.

In sum, these facts further support our contention that form features, i.e., morphology and phonology, are primary, though the combined models also have contributions from other domains, like the role of animacy in the all-features model.

5. Conclusion

This article has given a rule-based analysis of gender by establishing the number of genders, summarizing the inflectional rules for gender in Tashlhiyt, documenting default masculine gender, and investigating the paradigm structure of gender marking (i.e., variable vs. fixed vs. neutral). Using this linguistic system for gender, we constructed a database of 1914 Tashlhiyt noun paradigms with information about many attributes relevant to predicting gender. These included semantic attributes like animacy and semantic field, phonological attributes like the initial and final segment of noun themes, and morphological attributes like the use of secondary morphology and certain vowel augments (i.e., the R-Aug vowel) that are part of the morphological system. This qualitative analysis of gender assignment establishes the key factors for predicting gender.

We have also constructed a quantitative analysis of Tashlhiyt gender assignment. We have given descriptive statistics in several cross tables that show how gender categories are correlated with the specific values of semantic, phonological, and morphological features. Using these tables, we tried to enumerate gender predictability with categorical or near categorical features (e.g., secondary morphology or sex-gender) and found that such an account does not meet the expected levels of predictability from prior research. This negative result motivated the use of supervised learning to develop computational classifiers for predicting gender. Using the same features and the labelled data in the database, the classification models reached the expected levels of predictability. These models also enabled us to compare different linguistic domains and features, and we found that morphology was especially important in predicting gender, with phonological structure a close second. The classification models also supported an analysis of feature importance, which was consistent with our broader analysis, because specific morphological (secondary morphology) and phonological features (final phoneme of theme) were very important, but also this analysis revealed that animacy plays an important role. The

machine learning methods have been broadly successful in that they allowed us to establish an agreed upon standard of predictability and make new discoveries in the factors predicting gender assignment.

What do these findings tell us about the typology of gender assignment? First, they tell us that gender assignment in Tashlhiyt is form-based, with morphology playing an important role, followed by phonology. Interestingly, while morphology alone can predict gender well, phonology alone is not far behind, and when they are combined, they approach 90% predictability. This differs from semantic features, which are not sufficiently predictive on their own and, despite the feature importance of animacy, semantic features do not significantly improve performance when they are added to other feature subsets. In sum, we can say that gender assignment is due to a combination of morphology and phonology.

Turning to methodology, we believe that our analysis constitutes a strong case for using machine learning tools to analyze gender assignment. Building on the insights of Allassonnière-Tang et al. (2021), we have shown that relatively simple classifiers produce results superior to an enumerative approach using categorical features. These computational models have off the shelf tools that address many of the problems one must face in analyzing gender, including data imbalances, missing data, feature engineering, and feature importance. The success of our models in identifying the correct predictor domains (morphology and phonology) and pin-pointing the most important features shows that for at least one gender system, these methods can lead to new discoveries.

Finally, we can ask how to extend these methods to new languages, including languages with less linguistic documentation. We can think of the overall problem as one of predicting *y*, the gender of a noun, from *X,* a set of the noun's features. In our case, we constructed a database containing values for both *y* and many sensible features for *X*. Since one of the co-authors is a native speaker of Tashlhiyt, and the size of the database is relatively small, the task of filling in these values was a tractable one. While we believe that there is no substitute for documenting and verifying noun information with native speakers, this kind of documentation may not always be possible. In addition, the mental lexicons of humans are likely to have far more than 1900 nouns, so it may be useful to accelerate the process of data acquisition for larger lexicons, whether a native speaker is available or not.

There are a host of mechanisms for acquiring these attributes programmatically. Most lexicons have word lemmas as a headword, either in orthographic or phonetic form, so this is a good place to start building a database. If only an orthographic representation exists, it can typically be converted to a phonetic representation using one of many scripts in the NLP community or one created for this purpose. From there, feature engineering can extract a whole host of phonological attributes, including segments in special positions (initial, final, etc.), stress or tone patterns, syllable structures by position, and even higher-level prosodic structure if the rules for forming metrical stress feet and other units are well-known. Likewise, morphological attributes can also be developed using special purpose scripts designed to extract patterns. For example, declension class can be extracted from either dictionaries or texts by searching for unique markers of declension class. In Tashlhiyt, for example, R-Aug can be calculated by searching for glides versus vowels in masculine bound patterns using regular expressions or consonant clusters in feminine bound forms and then extracting the relevant vowel if the theme is consonant-initial (i.e., has a glide in masculine or a cluster in feminine bound forms).

Semantic information in *X* may be more difficult to collect if it is missing from available lexical resources, but one alternative is to substitute word embeddings for a traditional meaning

and develop features from these representations. Word embeddings are vector representations that encode word meaning as a function of distributions in texts, and they are used in a number of NLP tasks involving numerical computation of meaning (Jurafsky & Martin 2021). These embeddings may have predictive value in their own right, but they can also be used to calculate other values with feature engineering. For example, word embeddings can be used by animacy detection algorithms that combine them with other information, like part of speech tags, to predict animate versus inanimate nouns (Jahan et al. 2018). Thus, there seem to be at least some tractable ways of generating values for *X* programmatically. Many under-studied languages have repositories with word embeddings (as in the IndicNLP GitHub repository for many Indic languages), and they can be generated from texts if embeddings do not yet exist for a language. We think that with methods like these, future work can extend supervised learning approaches to a wider range of languages, providing a wider empirical basis for generalizations about gender assignment cross-linguistically.

Appendix – Modeling results

The table below lists all standard performance metrics, cross-classified by classification model and feature subset. The optimized parameters for the Random Forest, AdaBoost, and Gradient Boosting classifiers are also given.

Feature set	Model	Accuracy	Precision	Recall	f1	AUC-ROC	Brier loss
morphology	rfc	0.82	0.86	0.91	0.88	0.84	0.13
morphology	abc	0.82	0.85	0.92	0.88	0.84	0.23
morphology	gbd	0.82	0.84	0.94	0.89	0.85	0.13
morphology	svc	0.77	0.78	0.96	0.86	0.71	0.17
morphology	evc	0.82	0.85	0.92	0.89	0.85	0.14
phonology	rfc	0.78	0.83	0.89	0.86	0.75	0.17
phonology	abc	0.8	0.81	0.95	0.88	0.76	0.24
phonology	gbd	0.81	0.82	0.95	0.88	0.77	0.15
phonology	svc	0.78	0.78	0.99	0.87	0.69	0.16
phonology	evc	0.8	0.82	0.94	0.87	0.77	0.15
semantic	rfc	0.74	0.76	0.96	0.85	0.69	0.17
semantic	abc	0.75	0.75	1	0.85	0.69	0.25
semantic	gbd	0.74	0.76	0.97	0.85	0.69	0.17
semantic	svc	0.75	0.75	1	0.86	0.5	0.19
semantic	evc	0.74	0.76	0.96	0.85	0.69	0.18
phon+morph	rfc	0.85	0.88	0.94	0.91	0.87	0.11
phon+morph	abc	0.84	0.86	0.94	0.9	0.89	0.23
phon+morph	gbd	0.85	0.87	0.95	0.9	0.9	0.1
phon+morph	svc	0.79	0.78	0.99	0.87	0.72	0.16
phon+morph	evc	0.86	0.87	0.95	0.91	0.9	0.12
phon+sem	rfc	0.79	0.83	0.9	0.86	0.79	0.15
phon+sem	abc	0.79	0.82	0.94	0.87	0.81	0.24
phon+sem	gbd	0.81	0.82	0.94	0.88	0.81	0.14
phon+sem	svc	0.78	0.77	0.99	0.87	0.75	0.16
phon+sem	evc	0.8	0.83	0.93	0.87	0.81	0.15
morph+sem	rfc	0.8	0.85	0.89	0.87	0.83	0.14
morph+sem	abc	0.82	0.85	0.92	0.88	0.86	0.23
morph+sem	gbd	0.82	0.85	0.93	0.89	0.87	0.12
morph+sem	svc	0.75	0.75	1	0.86	0.78	0.17
morph+sem	evc	0.81	0.85	0.91	0.88	0.86	0.13
all	rfc	0.86	0.88	0.94	0.91	0.9	0.1
all	abc	0.85	0.87	0.93	0.9	0.9	0.22
all	gbd	0.85	0.87	0.94	0.91	0.91	0.1
all	svc	0.78	0.78	0.99	0.87	0.75	0.16
all	evc	0.86	0.88	0.95	0.91	0.92	0.11

Hyperparameterization

The optimized parameters for the Random Forest classifier are:

- n_estimators: 500
- max_features: 'auto'
- max_depth: 90
- min_samples_split: 5
- min_samples_leaf: 4
- bootstrap: False

A simple grid search for the Adaboost classifier paramters produced the following optimized parameters:

- Learning_rate: 1.0
- n_estimators: 200

A randomized search for Gradient Boosting classifier parameters produced the following optimized parameters:

- n_estimators: 100
- min_samples_split: 40
- min_samples_leaf: 5
- max_depth: 5
- learning_rate: 0.1

Acknowledgements

We are grateful to Donna Gerdts, Lana Leal, and Rachid Ridouane and two anonymous Linguistic Analysis reviewers for questions and comments on earlier drafts of this article. This research was supported in part by an Insight grant from the Social Science and Humanities Research Council of Canada (435-2020-0193).

References

Adouane, Wafia, Nasredine Semmar & Richard Johansson. 2016a. Romanized berber and romanized arabic automatic language identification using machine learning. Paper presented to the Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3), 2016a.

Adouane, Wafia, Nasredine Semmar, Richard Johansson & Victoria Bobicev. 2016b. Automatic detection of arabicized berber and arabic varieties. Paper presented to the Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3), 2016b.

Ahlberg, Malin, Markus Forsberg & Mans Hulden. 2015. Paradigm classification in supervised learning of morphology. Association for Computational Linguistics 2015.1024-29.

Alderete, John, Abdelkrim Jebbour, Bouchra Kachoub & Holly Wilbee. 2015. Tashlhiyt Berber grammar synopsis (ed.) S.F. University.

Allah, Fadoua Ataa & Siham Boulaknadel. 2012. Toward computational processing of less resourced languages: Primarily experiments for Moroccan Amazigh language. Theory and Applications for Advanced Text Mining, ed. by S. Sakurai, 197-218. Rijeka: InTech.
Allassonnière-Tang, Marc, Dunstan Brown & Sebastian Fedden. 2021. Testing semantic dominance in Mian gender: Three machine learning models. Oceanic Linguistics 60.302-34.
Amri, Samir, Lahbib Zenkouar & Mohamed Outahajala. 2017. Build a morphosyntaxically annotated amazigh corpus. Paper presented to the Proceedings of the 2nd international Conference on Big Data, Cloud and Applications, 2017.
Aronoff, Mark. 1991. Noun classes in Arapesh. Yearbook of morphology, 21-32. Dordrecht: Springer Netherlands.
Aronoff, Mark & Nanna Fuhrhop. 2002. Restricting suffix combinations in German and English: Closing suffixes and the monosuffix constraint. Natural Language and Linguistic Theory 20.451-90.
Aspinion, R. 1953. Apprenons le berbère; initiation aux dialectes chleuhs. Rabat: Editions Félix Moncho.
Barkin, Florence. 1980. The role of loanword assimilation in gender assignment. Bilingual Review/La Revista Bilingüe 7.105-12.
Basset, André. 1932. Note sur l'état d'annexion en berbère. Bulletin de la Société de Linguistique de Paris 33.173-74.
—. 1952. La langue berbère. Oxford: Oxford University Press.
Basset, André & A Picard. 1948. Éléments de grammaire berbère (Kabylie/Irjen). Alger: La Typo-Litho et Jules Carbonel.
Bensoukas, Karim. 2001. Stem forms in the nontemplatic morphology of Berber: Mohammad V University, Rabat, Morocco.
—. 2015a. bu-nouns in TashlhitAn oft-overlooked complex morphosyntactic corpus. Corpus.165-88.
—. 2015b. Expressing ownership in tashlhit: Phrasal affix (ation) vs. bound word (hood). Revue de l'Institut Royal de la Culture Amazighe 10.11-38.
—. 2018. Concurrent cognate and contact-induced plural traits in Afro-Asiatic: Amazigh id-and Arabic-at plurals. International Journal of Arabic Linguistics 4.59-102.
—. 2019. id-Pluralization in Tashlhit: Definitely not a root-and-pattern morphology! La catégorisation grammaticale en amazighe. Actes des journées d'étude 10 et 11 novembre 2016, ed. by A. Boumalk & H. Souifi, 28-46. Rabat: IRCAM publications.
Boukous, Ahmed. 1977. Langage et culture populaires au Maroc, essai de sociolinguistique. Casablanca: Dar El-kitab.
Boumalk, Abdallah & Karim Bensoukas (eds) 2018. *La racine dans les langues chamito-sémitique: nature et fonction* **13**). Rabat: L'Institute Royal de la Culture Amazighe.
Boussayer, Abdelaaziz. 2021. Gender and Number Marking in Amazigh Language. International Journal of Linguistics and Translation Studies 2.91-106.
Buck, Carl Darling. 2008. A dictionary of selected synonyms in the principal Indo-European languages. . Chicago: University of Chicago Press.
Chaker, Salem. 1990. La parente chamito-sémitique du berbère: un faisceau d'indices convergent. Etudes et Documents Berberes 7.28-57.
—. 1998. Genre (grammatical) en Berbère. Encyclopédie berbère 20.3042-45.

Corbett, Greville. 1982. Gender in Russian: An account of gender specification and its relationship to declension. Russian Linguistics 6.197-232.
Corbett, Greville G. 1991. Gender. Cambridge: Cambridge University Press.
—. 2007. Gender and noun classes. Language typology and syntactic description (second edition). Volume III: Grammatical categories and the lexicon, ed. by T. Shopen, 241-79. Cambridge: Cambridge University Press.
Corbett, Greville G & Norman Fraser. 2000. Gender assignment: A typology and a model. Systems of nominal classification, ed. by G. Senft, 293–325. Cambridge: Cambridge University Press.
Corbett, Greville G. (ed.) 2014. *The expression of gender*. Berlin: De Gruyter Mouton.
Cortes, Corinna & Vladimir Vapnik. 1995. Support-vector networks. Machine Learning 20.273-97.
Dell, Francois & Mohamed Elmedlaoui. 1989. Clitic ordering, morphology and phonology in the verbal complex of Imdlawn Tashlhiyt Berber. Langues Orientales Anciennes: Philologie et Linguistique 2.165-94.
—. 1991. Clitic ordering, morphology and phonology in the verbal complex of Imdlawn Tashlhiyt Berber. Langues Orientales Anciennes: Philologie et Linguistique 3.77-104.
Dell, François & Mohamed Elmedlaoui. 1992. Quantitative transfer in the nonconcatenative morphology of Imdlawn Tashlhiyt Berber. Journal of Afroasiatic Languages 3.89-125.
—. 2002. Syllables in Tashlhiyt Berber and in Moroccan Arabic. Dordrecht: Kluwer.
Dell, Francois & Abdelkrim Jebbour. 1991. Phonotactique des noms à voyelle initiale en berbère (chleuh de Tiznit, Maroc). Linguistic Analysis 21.119-47.
—. 1995. Sur la morphologie des noms en berbère (chleuh de Tiznit, Maroc). Langues Orientales Anciennes: Philologie et Linguistique 5-6.211-32.
El Hamdi, Fatima. 2018. On Tashlhit root structure and its implications for the organization of the lexicon: Paris 8, Universite Vancennes/Saint-Denis Doctoral dissertation.
El Moujahid, El Houssaïn. 1997. Grammaire générative du berbère: Morphologie et Syntaxe du Nom en Tashelhit. Série Thèses et mémoires n. 38. : Université Mohamed V Doctoral.
Freund, Y. & R. E. Schapire. 1995. A decision-theoretic generalization of on-line learning and an application to boosting. Paper presented at the European conference on computational learning theory.
Galand, Lionel. 1988. Le berbère. Les langues dans le monde ancien et moderne. Troisième partie: les langues chamito-sémitiques, ed. by D. Cohen & J. Perrot, 207-42. Paris: Editions du CNRS.
Gerdts, Donna. 2011. The purview effect: Feminine gender on inanimates in Halkomelem Salish. Berkeley Linguistics Society 37.417-26.
Guerssel, Mohamed. 1983. A phonological analysis of the construct state in Berber. Linguistic Analysis 11.309-30.
Hastie, Trevor, Robert Tibshirani & Jerome Friedman. 2009. The elements of statistical learning. New York: Springer.
He, Haibo & Edwardo A Garcia. 2009. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering 21.1263-84.
Ho, Tin Kam. 1995. Random decision forests. Paper presented to the Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995.

Idrissi, Ali. 2000. On Berber plurals. Research in Afroasiatic grammar. Papers from the Third Conference on Afroasiatic Languages, ed. by J. Lecarme, 101-24. Philadelphia: John Benjamins.
Jahan, Labiba, Geeticka Chauhan & Mark A. Finlayson. 2018. A new approach to animacy detection. Paper presented to the Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, 2018.
Jebbour, Abdelkrim. 1988. Les processus de formation du pluriel nominal en Tachelhit de Tiznit - Approche non-concaténative: University of Mohammed V, Rabat, Morocco DES Thesis.
—. 1996. Contraintes prosodiques et morphologie en berbère (tachelhit de Tiznit - Maroc). Analyse linguistique et traitement automatique: Université Mohammed V, Rabat, Morocco.
Jebbour, Abdelkrim, Ahmed Boukous, Abdelkrim El Alami, Jane SY Li, Rachid Ridouane & John Alderete. 2021. Nine Tashlhiyt texts: Structured representations of 18,000 words (First Release). Burnaby, BC: Department of Linguistics, Simon Fraser University.
Jurafsky, Daniel & James H Martin. 2021. Speech and language processing: An Introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, NJ: Pearson Prentice Hall.
Kramer, Ruth. 2014. Gender in Amharic: A morphosyntactic approach to natural and grammatical gender. Language Sciences 43.102-15.
—. 2020. Grammatical gender: A close look at gender assignment across languages. Annual Review of Linguistics 6.45-66.
Kupisch, Tanja, Miriam Geiss, Natalia Mitrofanova & Marit Westergaard. 2022. Structural and phonological cues for gender assignment in monolingual and bilingual children acquiring German. Experiments with real and nonce words. Glossa: a journal of general linguistics 7.1-37.
Lafkioui, Mena B. 2008. Dialectometry analyses of Berber lexis. Folia Orientalia 44.71-88.
Lahrouchi, Mohamed. 2008. A templatic approach to gemination in the imperfective stem of Tashlhiyt Berber. Studies in African Linguistics 37.21-59.
—. 2010. On the internal structure of Tashlhiyt Berber triconsonantal roots. Linguistic Inquiry 41.255-85.
Mason, L., J. Baxter, P. Bartlett & M. Frean. 1999. Boosting algorithms as gradient descent. Paper presented to the Advances in neural information processing systems 12, 1999.
McCarthy, John J. 1979. Formal problems in Semitic phonology and morphology: MIT Doctoral dissertation.
Mettouchi, Amina. 1999. (1999). Le" t" n'est-il qu'une marque de féminin en berbère (kabyle)? Faits de langues 7.217-25.
Oussikoum, Najat. 2019. L'accord en genre dans les catégories nominale, adjectivale, verbale et pronominale amazighes. Études et Documents Berberes 41.113-28.
Payne, Doris L. 1998. Maasai gender in typological perspective. Studies in African Linguistics 27.159–75.
Quint, Nicolas & Marc Allassonnière-Tang. 2022. Inferring case paradigms in Koalib with computional classifiers. Corpus Linguistics and Linguistic Theory 19.237-69.
Rice, Curt. 2006. Optimizing gender. Lingua 116.1394–417.
Saib, Jilali. 1986. Noun pluralization in Berber: A study in internal reconstrution. Langues et Littératures 5.109-33.

Spackman, Kent A. 1989. Signal detection theory: Valuable tools for evaluating inductive learning. Paper presented at the Proceedings of the sixth international workshop on machine learning, Morgan Kaufmann.
Taghbalout, I, F Ataa Allah & M El Marraki. 2015. Amazigh noun inflection in the universal networking language. International Journal of Education and Information Technology 9.122-28.
Taine-Cheikh, Catherine. 2002. Morphologie et morphogenèse du diminutif en zénaga (berbère de Mauritanie). Articles de linguistique berbère. Mémorial Werner Vycichl, ed. by K. Naït-Zerrad, 427-54. Paris: L'Harmattan.
Tucker, G. R., W. E. Lambert & A. A. Rigault. 1977. The French speaker's skill with grammatical gender: An example of rule-governed behavior. The Hague: Mouton.
Vycichl, Werner. 1961. Diminutiv und Augmentativ im Berberischen. Zeitschrift der Deutschen Morgenländischen Gesellschaft 111.243-53.
Zaliznjak, A. A. 1964. K voprosu o grammatičeskix kategorijax roda i oduševlennosti v sovremennom russkom jazyke. Voprosy jazykoznanija 4.25-40.
Zwicky, Arnold. 1988. Morphological rules, operations and operation types. Proceedings from the Fourth Eastern States Conference on Linguistics, ed. by A. Miller & J. Powers, 318-34. Columbus: The Ohio State University.