Noun classes grow on trees

Noun classification in the North-East Caucasus*

Keith Plaster[1], Maria Polinsky[1] & Boris Harizanov[2]
[1]Harvard University and [2]University of California, Santa Cruz

Noun classes (genders) have long played an important role in the understanding of language structure and human categorization. This study presents and analyzes the division of nouns into classes in Tsez (Dido), an endangered Nakh-Dagestanian language of the North-East Caucasus. Computational modeling of the Tsez system shows that noun classification in Tsez is highly predictable, with a simple semantic core and a set of highly salient formal features that can be ranked with respect to one another. Such a system would be easily accessible to children acquiring the language, and the proposed analysis does not require additional semantic or categorical assumptions. The study serves as a proof of principle for the computational approach to the analysis of noun classification.

1. Introduction

Noun classes (genders) have long played an important role in the understanding of language structure and human categorization. The attraction of noun classes lies in their connection to diverse aspects of language. Noun categorization is relevant for understanding lexical access (see, e.g. Levelt 1989, 1993; Vigliocco et al. 2002), agreement, and conceptualization in language. On the one hand, noun classes are omnipresent – think of the pervasive patterns of agreement in such languages as Spanish or Swahili, which should probably make them stable; on the other hand, noun classes are fluid,

* We dedicate this article to Johanna Nichols, whose pioneering work in the languages of the Caucasus has long served as an inspiration to the field.

We thank our consultants – Arsen Abdulaev, Madzhid Khalilov, and Paxrutdin Magomedinov – for their assistance with the data. We are grateful to Bernard Comrie, Greville Corbett, Annie Gagliardi, the participants at the Symposium on the Languages of the Caucasus and Linguistic Theory at the 2009 Annual Meeting of the Linguistics Society of America, and two anonymous reviewers for their helpful comments and discussions about this article. All errors, of course, remain our own.

subject to restructuring and change (Nichols 1989, 2003; Corbett 1991; Polinsky & van Everbroeck 2003; Plaster & Polinsky 2007). Many subfields of linguistics claim noun categorization as their own: lexicology, grammar, historical linguistics, language acquisition, and typology. Indeed, they all have intriguing insights on noun classes to offer, and it is through a combination of these insights that new progress can be made in understanding this seemingly simple phenomenon.

This paper is an attempt to bring several subfields interested in noun categorization together by offering them a tool that they can all use: computational modeling of noun classification at the synchronic level. Although we are interested in presenting the proof-of-concept evidence for this tool, we will also use it to further the knowledge of noun classification in the North-East Caucasian (Nakh-Dagestanian) languages, most of which have multiple noun classes (from three to eight; see Kibrik & Kodzasov 1988, 1990). Researchers have just started to scratch the surface of the historical development (e.g. Nichols 1989, 1995, 2003; Polinsky & Jackson 1999; Comrie & Polinsky 1999) and synchronic organization of noun classification in these languages, and we hope that this paper will help us make another modest step forward.

2. Background on noun classification in Tsez

Tsez (Dido) is a Nakh-Dagestanian language spoken in the highland area of the Tsunta district of Dagestan and in the lowlands in the vicinity of Maxachkala, where many speakers have been migrating lately. The language has approximately six or seven thousand speakers; the accuracy of this estimate may be undermined by the fact that Tsez speakers are registered as Avars (the largest ethnic group in the area) and most adults are fluent in Avar (see van den Berg 1992 on the recent population movements among the Tsez). For an overview of the Tsez phonological system, see Comrie (2007: 1194–1195).

With regard to noun classification, Tsez possesses four noun classes (genders), which are indicated through the use of class agreement prefixes on most vowel-initial adjectives and verbs,[1] as well as on certain adverbs, postpositions, and particles. The agreement prefixes and the general content of each class are shown in Table 1.

Tsez is morphologically ergative, distinguishing between the absolutive argument (intransitive subject or direct object) and ergative (transitive subject); following

1. Agreement on consonant-initial verbs is blocked by a phonotactic restriction against initial consonant clusters. Those verbs that appear to be vowel initial but do not register agreement have a hypothetical underlying initial laryngeal (cf. Comrie & Polinsky 1999; also Nikolaev & Starostin 1994).

Table 1. Tsez noun class agreement prefixes

Class	Content	Singular	Plural
I	Males (human and divine)	*Ø-*	*b-*
II	Females (human and divine) and various inanimates	*y-*	
III	Animals and various inanimates	*b-*	*r-*
IV	Other inanimates	*r-*	

the notation used in ergative studies, the absolutive represents S and O (P), and the ergative, A (Dixon 1994). Within a clause, verbs agree with the absolutive argument (cf. (1), in which the verb agrees with the noun *baru* in the object position):

(1) *xediy-ā* *baru* *y-egir-si*
husband-ERG wife.ABS II-send-PAST.EVIDENTIAL
'The husband sent his wife.'

Table 2 shows the agreement in noun class between the head noun and the modifying adjective:

Table 2. Paradigm of *-igu* 'good'

Class	Singular		Plural	
I	***Ø-igu***	*aħo*	***b-igu***	*aħo-bi*
	I.AGRsg-good	shepherd	I.AGRpl-good	shepherd-PL
	'good shepherd'		'good shepherds'	
II	***y-igu***	*baru*		*baru-bi*
	II.AGRsg-good	wife		wife-PL
	'good wife'			'good wives'
III	***b-igu***	*ʿomoy*	***r-igu***	*ʿomoy-bi*
	III.AGRsg-good	donkey	II-IV.AGRpl-good	donkey-PL
	'good donkey'			'good donkeys'
IV	***r-igu***	*ʿoƛ'*		*ʿoƛ'-mabi*
	IV.AGRsg-good	spindle		spindle-PL
	'good spindle'			'good spindles'

Although the existence of each of the four noun classes is readily apparent, the scope of each class is not. The classification of nouns referring to animate beings is clear; human and divine males (*uži* 'boy', *allah* 'Allah', *žek'u* 'man') fall within class I, human and divine females (*baru* 'wife', *ečju* 'grandmother', *ħurulʕin* 'fairy') fall within

class II, and all other animates (ʿ*omoy* 'donkey', *ayi* 'bird', *aw* 'mouse') fall within class III. However, nouns with inanimate referents fall within classes II, III, and IV; only class I contains no inanimate nouns.

In addition, while examination of the composition of the inanimate nouns in classes II, III, and IV reveals certain regularities (for example, the assignment of the names of berries primarily to class II and the assignment of derived abstract nouns in *-ƛi* or *-ni* to class IV), a number of other apparent categories of nouns are distributed across the three classes (for example, clothing, body parts, tools, and time terms). A sample of the inanimate nouns contained in each class is given in (2):

(2) Inanimate nouns in classes II, III, and IV

Inanimates in II: berries; paper items (letter, dictionary, newspaper); some clothing (not exclusively female); some body parts (knee, chin, shoulder blade, leg); some tools (hammer, plough, shovel); mountains, stones and rocks; some time terms (year, seasons); and various other inanimates (cage, drinking glass, salt, motor, dust, mill, science, etc.)

Inanimates in III: some clothing; some body parts (finger, calf, arm, heel, rib); some tools (hoe, chisel, sickle, tool); some time terms (month names, minute); vehicles; many Arabic loanwords;[2] and various other inanimates (alphabet, field, call, proverb, gasoline, sun, moon, etc.)

Inanimates in IV: derived abstract nouns in *-ƛi* or *-ni*; some clothing; some body parts (wrist, knuckles, belly, shoulder); some time terms (day names, day); and various other inanimates (wine glass, crib, university, navy beans, rye, stick, milk, etc.)

Thus, the basic distribution of the Tsez noun classes is as shown in (3):

(3) Basics of Tsez noun classes
I: males
II: females + [class II inanimates]
III: other animates + [class III inanimates]
IV: [class IV inanimates]

The question, then, is what comprises the [class II inanimates], [class III inanimates], and [class IV inanimates], and how do speakers know the classification of inanimate nouns? We turn to this question and various answers that have been proposed in Section 3.

2. A disproportionate number of Arabic loanwords referring to inanimates appear in class III, compared with, for example, the classification of Russian loans, which are much more evenly distributed between classes III and IV in our sample.

3. Approaches to noun classification in Tsez

Three potential approaches that could account for the distribution described above readily come to mind.

First, one may suggest that the classification of each inanimate noun is simply memorized and contained in the features for the noun stored in the speaker's lexicon, along with other information, such as declension type. Thus, upon hearing a new noun used, the learner not only stores the form, meaning, and declension type but also the noun class.

Although views on the cost of storage versus the cost of processing have shifted as advances in computer technology have dramatically reduced the cost of storage, positing memorization of the classification of each inanimate noun is nonetheless unattractive. First, memorization of the classification of each noun is a large task, particularly for a child learning the language and especially when information about the classification of each noun may not be robust, since class information surfaces only on vowel-initial agreeing elements. Second, the absence of a readily available, transparent explanation for class assignment does not indicate that no explanation exists; other "arbitrary" gender systems have been shown to be, in fact, predictable (e.g. Tucker et al. 1977; Lyster 2006 for French, Harris 1991a for Spanish, Tanenbaum 2003 for German). Finally, memorization of noun classes would fail to explain the cross-speaker consistency in the assignment of nonce forms to noun classes found by Polinsky & Jackson (1999) and Gagliardi et al. (2009).

Second, it has been suggested by Rajabov (1997) that the Tsez noun classes contain an internal logic similar to that proposed by Lakoff (1987) for Dyirbal (the source of "women, fire, and dangerous things"). Lakoff's account of Dyirbal, and Rajabov's account of Tsez, rely on the notion of a radial category, which is centered around a "prototype," or the member that possesses most of the defining characteristics of that category. Other nouns are then included within the category on the basis of their perceived resemblance to the prototype, but they do not have to actually share the criterial features of the prototype. For example, if a human female is the prototype for a class, a garment worn only by women may be placed in the same class because of the perceived resemblance between the garment and women, although the garment and a human female share none of the same characteristics. The more peripheral members are linked to the prototype through other members, and these links can be motivated by certain language-specific principles. Taken together, the members of a category thus form a radial structure, with the most representative, or prototypical, members located at the center and the less representative outliers clustered around this hub. Thus, under this account, Tsez speakers would learn the "core" or prototypical members of each class and assign the class membership of other nouns on the basis of how well they can be connected to these prototypical members under the principles applicable in Tsez.

Rajabov (1997) proposes a variety of principles for the assignment of nouns to classes in Tsez. These principles are set forth in (4):

(4) Principles for assignment of noun classes in Tsez (Rajabov 1997)
   a. Material: if X is the material out of which Y is made, Y may be assigned to the same class as X (e.g. 'wood' and 'chair')
   b. Shape: flat items tend to go into class II; round, nonflat things tend to go into class III; long, thin items tend to go into class IV
   c. Internal feature: liquidness and density sometimes are relevant to class assignment ('ice' is in class III because of its association with 'rock', but it could be expected to be in class IV under the 'material' principle)
   d. Function: if Y is used for or resembles X functionally, Y may be assigned to the same class as X (e.g. 'fortress' is in class III because 'fight' is in class III)
   e. Semantic domain association: the assignment of nouns may create semantic domains (e.g. 'sock' is assigned to class IV on the basis of 'wool', and a semantic category of 'footwear' is subsequently created in class IV on the basis of the assignment of 'sock')
   f. Species to genus association: nouns referring to specific instances of more general nouns will be assigned to the class of the more general noun (e.g. the words for different fingers are assigned to the same class as 'finger')
   g. Concept association (analogy): loanwords that duplicate existing words may be assigned the class of the duplicated words
   h. Opposites: words expressing opposite concepts are placed in the same class (e.g. 'fire' and 'water' are in class IV, 'medicine' and 'poison' are in class III)[3]

This explanation is also unappealing for several reasons. First, the links between members and the principles purportedly causing such links seem to act more as after-the-fact generalizations than operational principles. In addition, the vast majority of the principles proposed by Rajabov (1997) are identified as tendencies rather than systematic rules for the assignment of nouns. Thus the account does not motivate either the links between members or the overall class assignments in an unambiguous and predictive manner.

A third possibility is that Tsez speakers rely on a combination of semantic and formal features to classify nouns. Under this approach, noun classification should

3. Rajabov (1997) also proposed a principle calling for the assignment of homonymous nouns to the same class, but his examples appear to be the result of the semantic extension of a single noun to a second meaning rather than two separate homonymous nouns (e.g. 'moon' and 'month', 'tongue' and 'language', 'year' and 'leaf', and 'nose' and 'hill').

be able to be explained by appealing to simple semantic and formal features, in both cases of the sort young children are sensitive to (cf. Jusczyk et al. 1994; Saffran et al. 1996, Karmiloff-Smith 1979; Levy 1983; Berman 1985; Smoczyńska 1985; Slobin 1973; Gerken et al. 2005). Formal features have been shown to be relevant in the determination of gender or noun class in a variety of languages, including Russian (Corbett 1991), French (Tucker et al. 1977), and Romanian (Bateman & Polinsky 2010) and even in such complex systems as German (Tanenbaum 2003) or Dyirbal (Plaster & Polinsky 2007, 2010).

The groundwork for this approach to noun classification in Tsez has already been laid by Comrie & Polinsky (1999) and Polinsky & Jackson (1999). Comrie & Polinsky (1999) identified a synchronic connection between both *i*/*y* and *u*/*w* with class II, with these segments appearing in initial or final position, serving as predictors of assignment to class II. The connection between *i*/*y* and class II was unsurprising, as it is seen in a number of other Nakh-Dagestanian languages (see, e.g. Nichols 1989 for the Nakh languages, Kibrik 1977 for Archi). The connection between *u*/*w* and class II is less clear; Comrie & Polinsky proposed that the feature may have been identified in high-frequency class II words like *baru* 'woman'.

Polinsky & Jackson (1999) also examined class II and identified class II as resulting from the merger of two earlier classes, both of which exist in other Nakh-Dagestanian languages.[4] In addition, they performed nonce-word testing to identify whether the presence of initial or final segments that are identical to the current class prefixes was used by speakers as a cue to class assignment, which resulted in several relevant findings. First, the presence of an initial or final segment identical to a class prefix was predictive of class assignment, with 92 percent of nonce forms beginning with *i* or *y* and 78.5 percent of nonce forms ending in *i* or *y* being assigned to class II. The presence of an initial or final bilabial or *r* was also associated with assignment to class III or IV, respectively, but with a lower level of predictiveness. Although both initial and final segments appeared to be predictive, Polinsky & Jackson found initial segments to be more predictive, and in the event of a conflict between the initial and final cue in a nonce form, the form was assigned to the class cued by the initial segment. This finding indicates that the similarity with the prefixal exponents of gender agreement may provide a powerful formal cue for noun classification in Tsez.[5]

---

4. Polinsky & Jackson (1999) proposed that the merger was enabled by the phonological development of the singular class agreement prefix of the former class V from **d*- > *y*-, causing it to be identical to the singular class agreement prefix of the much smaller class II.

5. We hypothesize that the strength of word-initial predictors is specific to Tsez and is due to the fact that agreement exponents are also word-initial. For Dyirbal, Plaster & Polinsky (2007, 2010) found that gender-predictive segments coincided with the stressed syllable, which is the first syllable of the word. The same correlation is found in French (final segments are strongly predictive; cf. Tucker et al. 1977). Stress in Tsez is relatively weak and, except in particular

Thus Comrie & Polinsky (1999) and Polinsky & Jackson (1999) identified the beginnings of an alternate analysis of noun classification in Tsez that appeals to the use of semantic and formal features to determine the classification of nouns. We have sought to expand this approach to noun classification in Tsez through a larger-scale, systematic analysis of a lexicon of Tsez nouns.

4. The current project

To establish whether noun classification in Tsez can be explained by appealing to a small number of simple semantic and formal features, we set out to computationally model the Tsez gender system. Computational modeling is not new to the analysis of gender systems (see Sokolik & Smith 1992, Polinsky & van Everbroeck 2003 (neural networks) and Tanenbaum 2003, Bateman & Polinsky 2010 (decision trees)) and has the benefit of enabling the testing of a variety of user-specified features in an efficient manner. Our goal was to identify a set of formal and semantic features of the sort to which young children acquiring language are known to be sensitive and to produce a decision tree containing these features that will predict the classification of nouns in Tsez.

4.1 Decision-tree modeling

Before turning to our results, an aside on decision-tree modeling will be beneficial. Decision trees are likely familiar to most readers; they are tools designed to assist with the making of a particular decision. Decision trees present a series of connected questions to be answered, beginning with a single question (or node) and branching from there, with the answer to each question leading further down the tree and eventually ending in a decision based on the answers given. They are commonly used in a wide range of fields, such as education, operations research and management, data mining, and machine learning. Decision trees are induced from a set of examples (a training set), each of which consists of a set of input attributes and a single output value (called a goal). Finding the smallest decision tree consistent with the examples requires identifying which decisions should be made before others: the idea is to test the most important attributes first – namely, those that make the most difference to the classification of a noun.

---

grammatical forms, falls on a word's final syllable if it is closed, or on the penult if the final syllable is open (Alekseev & Radžabov 1989:118). It may be the case that stressed syllables serve as a cue for gender only in languages in which stress is fixed at a word edge.

As an example, assume that we are examining the data set and attributes shown in Table 3. The training set consists of four Spanish nouns; for each noun, two attributes are identified ('female?' and 'final a?'), and the goal is the noun's gender (M or F).

Table 3. Spanish example

Noun	Female?	Final *a*?	Gender?
actriz 'actress'	Y	N	F
chica 'girl'	Y	Y	F
jardín 'garden'	N	N	M
pijama 'pajamas'	N	Y	M

If we examine the values of the two attributes, we find that the values of each attribute split the examples into the two subsets shown in (5) and (6):

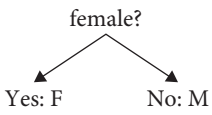
(5) 'female?'
subset 1 (=Y): *actriz* (F), *chica* (F)
setset 2 (=N): *jardín* (M), *pijama* (M)

(6) 'final *a*?':
subset 1 (=Y): *chica* (F), *pijama* (M)
setset 2 (=N): *actriz* (F), *jardín* (M)

As shown in (5), both nouns in the sample that denote females are feminine, and both nouns that do not denote females are not feminine. However, as shown in (6), the presence of a word-final /a/ does not separate the feminine nouns from the masculine nouns in our sample; two of the nouns in the sample end in /a/, one of which is masculine and one of which is feminine.

Thus we see that, based on our sample, it is more informative for the purposes of determining gender assignment to ask whether a noun refers to a female than whether it ends in an /a/. The presence of the feature [female] indicates assignment to F, and it turns out that no other feature is needed to explain the assignment of the non-[female] nouns in the data set. As a result, based on this training set, a simple decision tree (7) emerges:

(7)



We can run each of the four nouns in the data set through the decision tree; those nouns with a [female] feature will be assigned feminine gender, and those without the [female] feature will be assigned masculine gender, explaining the genders of the nouns in our data set perfectly.

However, as many readers will have already remarked, gender classification in Spanish is far from that simple (see Harris 1991a, b for a detailed discussion of the Spanish gender system), and the example serves to illustrate some of the potential pitfalls of decision-tree modeling. A vast number of feminine nouns, including the nouns shown in Table 4, do not have a [female] feature; these nouns would be predicted to be masculine under the decision tree in (7). The gender of forms like *leche* and *mano* would not be predicted based on the features examined in the data sets; like the masculine noun *jardín*, these forms do not have a female referent or end in /a/.

Table 4. Additional Spanish feminine nouns

Noun	Female?	Final *a*?	Gender?
casa 'house'	N	Y	F
leche 'milk'	N	N	F
mano 'hand'	N	N	F

As this small example shows, a comprehensive data set is critical. If the lexicon of Spanish contained only the four nouns in the data set in Table 3 or if all nouns in Spanish behaved as these four nouns do, the tree in (7) would model Spanish perfectly. Thus it is important to include as many nouns as possible in the hope of gathering representatives of all gender-assignment rules. In addition, a comprehensive sample will provide a better picture of the relative importance of attributes. For example, while the example shows that 'female?' is more informative than 'final a?' with respect to the forms in Table 3, this ceases to be true with the addition of the forms in Table 4. Even with the forms in Table 4, the predictive value of a final /a/ is not clear, when, in fact, a final /a/ is strongly predictive of feminine gender.

Second, the attributes selected for testing are critical. If relevant attributes are not identified, they cannot be tested. For example, in addition to the presence of a [female] feature or a word-final /a/, one may want to test whether the presence of a [male] feature or a word-final /o/ is predictive of Spanish gender. Unless those attributes are identified and tested, the resulting model will not be able to appeal to them, possibly leaving certain nouns without a basis for class assignment and possibly overstating the importance of other attributes.

Finally, it is important to note that a decision tree will not account for lexical exceptions, which must be memorized under any model. For example, while a final

/a/ is in fact a cue to feminine gender in Spanish, certain exceptions exist (nouns such as *día* 'day' or *pijama* 'pajamas'), whose masculine gender must be memorized by speakers.

4.2 Testing

Having provided an overview of decision-tree modeling, we now turn to our project and results. Our data set consisted of over 3,500 nouns culled from Khalilov 1999 and Rajabov (undated). To ensure the accuracy of noun-class information and dialect consistency, the classification of each noun was confirmed with native Tsez speakers of the Kidiro and Mokok dialects. Our lexicon contains a large number of loanwords, some of which are older (e.g., loanwords of Arabic origin, such as *amru* 'order', *din* 'religion', *sual* 'question', *alim* 'scholar, teacher') and some of which are more recent (e.g. the many Russian loans currently in use, from the earliest – such as *konka* 'public transportation' or *istoli* 'table' to *pilet/billet* 'ticket', *tilipon* 'telephone', *učitel* 'teacher' – to such recent words as *ewro* 'euro' or *nowutbuk* 'laptop').

The distribution of the nouns in our data set among the four classes is shown in Table 5:

Table 5. Distribution of nouns in data set

Class	% of total
I	12.6%
II	12.4%
III	41.4%
IV	33.6%

Class III is the largest class in our sample, with class IV coming in a relatively close second. We included all *nomina agentis* and other nouns that can be I or II depending on the referent (e.g. teacher, student, boss, American (person)) as class I nouns in our sample, causing the number of class I nouns to be slightly inflated, but since the class of these nouns is determined by the gender of the referent, our results would not be affected if these nouns were instead included as class II nouns.

As we saw in the Spanish example above, attribute selection is very important to successful modeling. Therefore, we tested a broad set of semantic and formal features to which children may be sensitive. The formal features that we tested included the identity of the noun's first segment, the identity of noun's last segment, the identity of the noun's first two segments, the identity of the noun's last two segments, the noun's declension class, and the number of syllables in the noun. The semantic features

we tested included [male], [female], [animate], [berry], [paper], [edible], [vehicle], [container], and [stone].

We coded each noun in our sample for each of the features we tested and ran the resulting data set through the decision-tree module of the Orange data-mining tool (Demsar & Zupan 2004), based on Quinlan's C4.5 algorithm (Quinlan 1993), to identify which attributes were predictive of noun classification and the relative rankings among the attributes. Decision-tree algorithms, like Quinlan's C4.5 algorithm, are simple but powerful learning algorithms widely used in data mining and the field of machine learning. We then used these results to produce the decision tree shown below.

4.3 Results

The results of the decision-tree modeling demonstrate that even a complicated system such as that of Tsez may be explained through the use of simple formal and semantic features of the sort that children are known to pay close attention to.

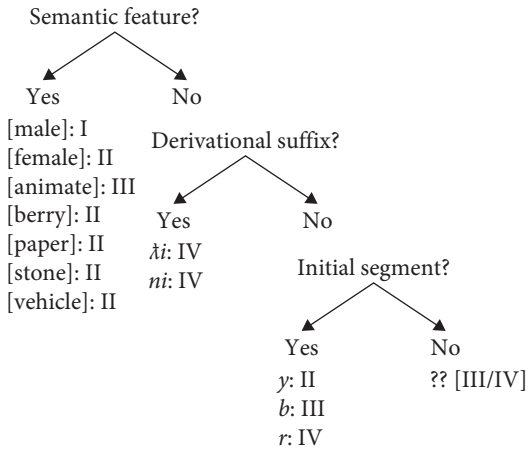
The presence of a semantic feature was found to be most predictive of noun classification. In addition, semantic features override conflicting formal features, as is typical in gender systems with both types of gender cues (e.g. Corbett 1991, Gentner & Namy 1999). As noted earlier, all males appear in class I, all females appear in class II, and all other animates appear in class III. Several smaller semantic classes also appear to be predictive, including [berry], [paper], and [stone], all of which are strongly correlated with assignment to class II. The feature [vehicle] is strongly predictive of class III and may reveal that [vehicle] and [animate] may both be better identified as a single feature [mobile].

The presence of one of the two abstract-forming derivational suffixes, *-ƛi* and *-ni,* was also strongly predictive of noun class. All nouns ending in the productive suffix *-ƛi* are assigned to class IV, and the vast majority of nouns ending in the formerly productive suffix *-ni* are also assigned to this class. It is important to note that it is the presence of these derivational suffixes rather than an [abstract] feature that is responsible for assignment to class IV because abstract nouns appear in classes II and III as well (e.g. *gaq'u* 'destruction' (class II), *kep* 'happiness' (class II), *adab* 'politeness, respect' (class III), and *bax* 'luck' (class III)).

In addition, our results confirmed that the presence of a word-initial segment identical to one of the class prefixes was predictive of assignment to the class of the class prefix. However, these segments were not found to be predictive in word-final position, against the results of Comrie & Polinsky (1999) and Polinsky & Jackson (1999), both of which found final *i*/*y* and *u*/*w* to be predictive of class II.

Our current decision tree, which explains the classification of 69 percent of the nouns in our data set, is shown in (8).

(8)[6]



Thus, with only a small number of simple semantic and formal features, we have been able to account for almost 70 percent of the noun classifications in our sample. The burning question, of course, is how to explain the remaining 30 percent of the nouns in our sample. The answer to this question likely lies in the answers to a number of questions and potential confounds that our investigation has turned up and that merit further investigation. Although we cannot currently provide answers to these questions, we share our preliminary thoughts below.

First, we suspect that the number of loanwords in the lexicon may be affecting the results, particularly where the basis for the assignment of these loanwords to classes is driven by the classification of a semantically identical or similar Tsez word rather than as the result of the semantic or formal features possessed by the loanword. Although one may attempt to remove the loanwords from the data set studied, to do so would provide a less comprehensive view of the task faced by a child learning the language or the competence of a mature speaker; many of these loanwords are part of the Tsez daily vocabulary, and children learning the language do not have the benefit of knowing which words are native and which words are not. To the extent that formal features possessed by loanwords are simply ignored in their classification, the value of these features for the native Tsez vocabulary is understated; at the same time, however, the value of these features for the entire Tsez vocabulary is not. As a result of the heavy influx of loanwords, the principles of class assignment in Tsez may be expected to change as children are required to make generalizations over a different lexicon from

6. To be clear, each node of the tree asks only about relevant values of each feature and treats any nonrelevant (i.e. nonpredictive value) as a negative response.

that know by previous generations. Thus it is possible that noun classification in Tsez currently may be in an interim stage of development. Studies of the acquisition of Tsez noun classification, such as that of Gagliardi et al. (2009), may provide useful insight on these issues.

Second, dialectal variation in noun classification may be responsible for a portion of the classifications not covered by the tree in (8). There are at least five dialects of Tsez (Bokarev 1959): Asakh, Mokok, Kidiro, Shaitli, and Sagada (the latter is likely to be a separate language). Phonological and lexical differences between the dialects are known to exist, and it is possible that the classification of some nouns may differ as well. We confirmed the classification of each noun in our sample with native speakers of the Kidiro and Mokok dialects. However, in some cases, Rajabov (undated) provides a different classification from the one identified by Khalilov (1999) and our consultants, which may be due to the fact that Rajabov is a speaker of the Asakh dialect. Dialectal variation is known to yield some differences in other languages of the family (e.g. Kibrik 1999: 48–49), and some of its effects may be expected in Tsez as well.

In addition, study of acquisition of the noun classification system may shed light on the importance of the smaller semantic fields that we have identified. As noted above, children, from an early age, are known to be sensitive to certain core semantic features (sex, animacy). As a result, we would predict that children should acquire the smaller clusters of nouns forming a semantic sphere at a later age than they acquire [male], [female], and [animate]; in other words, we would predict that a child's decision tree may develop through not only the addition of nodes but also through the addition of relevant items under each node as he or she is able to identify generalizations. In addition, it may be possible to identify at what age children identify the smaller semantic categories and which factors (e.g. frequency, number of relevant nouns) may be correlated with the earlier or later identification of a category.

It is also possible that additional or slightly different semantic categories may be relevant for the classification of Tsez nouns. While we want to explain the classification of as many nouns as possible, we also want to ensure that we do not fall into the trap of simply positing after-the-fact generalizations that fit the data solely to explain more data. For this reason we have focused on features that young children are known to or may be expected to be sensitive to, and we have avoided semantic features that require complex, abstract connections or complex cultural knowledge that children are likely not to possess. However, a continued examination of the semantic features involved in the assignment of nouns to classes may prove useful.

Finally, an outstanding question remains the identity of the Tsez "default" class, or, in other words, the class containing nouns not bearing a pertinent semantic or formal feature, as noted in the decision tree in (8) above; it is unclear whether unassigned nouns are placed into class III or class IV by default. Class III is the largest class in our sample, as noted above, but size alone is not sufficient to justify its identification as the default class.

Gagliardi et al. (2009) found an apparent difference in the default class used by children and adults in the classification of nonce forms: while for children, class III was the default class, for adults the default class appeared to be class IV. Thus, the uncertainty about the identity of the default class appears to extend beyond our current analysis.

5. Conclusions

In this paper, we have shown that noun classification in Tsez is highly predictable. The Tsez system of noun classification can be reduced to a simple semantic core, as is common in all noun classification (gender) systems (Corbett 2011), and a set of highly salient formal features. The semantic core consists of the common features (natural gender, animacy) that children acquiring language are sensitive to and several smaller semantic categories that do not require abstract connections or cultural knowledge which may not be available to young language learners. The formal features similarly are of the sort that would be easily accessible to children learning the language – they are mainly initial segments and a small number of suffixes. The predictive initial segments are specifically reinforced because they coincide with the inflectional exponents of noun-class agreement in Tsez.

In addition, the current study serves as a proof of principle for the approach taken here – namely, that noun classification (gender) is reducible to a set of simple semantic and formal features that can be ranked with respect to one another. While we have adopted a decision-tree model to depict the relative rankings of these features, ordered rules or other models enabling these rankings to affect noun classification would also be consistent with our approach. It is our hope that the simple computational modeling we relied on in this project will prove useful to other studies of complex gender systems, as well as diachronic reanalyses of genders.

References

Alekseev, M.E. & R.N. Radžabov. 1989. Tsez. In M. Job (ed.), *The indigenous languages of the Caucasus,* vol. III: *The North East Caucasian Languages,* part I, 115–163. Ann Arbor, MI: Caravan Books.

Bateman, Nicoleta & Maria Polinsky. 2010. Romanian as a two-gender language. In Donna Gerdts, John Moore & Maria Polinsky (eds.), *Hypothesis A/Hypothesis B*, 41–78. Cambridge, MA: The MIT Press.

Berg, Helma van den 1992. The Tsezic peoples and the policy of resettlement (with special reference to the Hunzib). *Annual of the Society for the Study of Caucasia* 3–4. 45–53.

Berman, Ruth. 1985. The acquisition of Hebrew. In Dan I. Slobin (ed.), *The cross-linguistic study of language acquisition*, 255–371. Hillsdale: Lawrence Erlbaum Associates.

Bokarev, Eŭgeno A. 1959. *Cezskie (didojskie) jazyki Dagestana*. Moscow: Izdatel'stvo Akademii Nauk SSSR.

Comrie, Bernard. 2007. Tsez (Dido) morphology. In A. S. Kaye (ed.), *Morphologies of Asia and Africa,* vol. 2, 1193–1204, Winona Lake, IN: Eisenbrauns.

Comrie, Bernard & Maria Polinsky. 1999. Gender in a historical perspective: Radial categories meet language change. In C. Justus & E. C. Polomé (eds.), *Language Change and Typological Variation: In Honor of Winfred P. Lehmann on the Occasion of His* 83rd *Birthday* – Vol. 2*, Grammatical Universals and Typology*, 566–589. Washington, DC: Institute for the Study of Man.

Corbett, Greville. 1991. *Gender*. Cambridge: Cambridge University Press.

Corbett, Greville. 2011. Sex-based and non-sex-based gender systems. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*, Chapter 31. Munich: Max Planck Digital Library. ⟨http://wals.info/chapter/31⟩ (31 August 2013).

Demsar, J. & B. Zupan. 2004. Orange: From experimental machine learning to interactive data mining. Ljubljana: University of Ljubljana. ⟨www.ailab.si/orange⟩ (31 August 2013).

Dixon, R.M.W. 1994. *Ergativity*. Cambridge: Cambridge University Press.

Gagliardi, Ann, Jeffrey Lidz & Maria Polinsky. 2009. *The acquisition of noun classes in Tsez: Computational and experimental results*. Poster presented at the 34th Boston University Conference on Language Development (BUCLD34), November 2009.

Gentner, Dedre & Laura Namy. 1999. Comparison in the development of categories. *Cognitive Development* 14. 487–513.

Gerken, LouAnn, Rachel Wilson & William Lewis. 2005. Infants can use distributional cues to form syntactic categories. *Journal of Child Language* 32. 249–268.

Harris, James. 1991a. The exponence of gender in Spanish. *Linguistic Inquiry* 22. 27–62.

Harris, James. 1991b. The form classes of Spanish substantives. *Yearbook of Morphology* 1991. 65–88.

Jusczyk, Peter W., Paul A. Luce & Jan Charles-Luce. 1994. Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language* 33. 630–645.

Karmiloff-Smith, Annette. 1979. *A functional approach to child language: A study of determiners and reference*. Cambridge: Cambridge University Press.

Khalilov, M. Sh. 1999. *Tsezsko-russkij slovar'*. Moscow: Academia.

Kibrik, A.E. 1977. *Opyt strukturnogo opisanija arčinskogo jazyka*, vols. 2 and 3. Moscow: MGU.

Kibrik, A.E. (ed.). 1999. *Elementy caxurskogo jazyka v tipologičeskom osveščenii.* Moscow: Nasledie.

Kibrik, A.E. & S. V. Kodzasov. 1988. *Sopostavitel'noe izučenie dagestanskix jazykov: Glagol.* Moscow: MGU.

Kibrik, A.E. & S. V. Kodzasov. 1990. *Sopostavitel'noe izučenie dagestanskix jazykov: Imja. Fonetika.* Moscow: MGU.

Lakoff, George. 1987. *Women, fire, and dangerous things: What categories reveal about the mind.* Chicago, IL: University of Chicago Press.

Levelt, Willem J.M. 1989. *Speaking: From intention to articulation*. Cambridge, MA: The MIT Press.

Levelt, Willem J.M. (ed.). 1993. *Lexical access in speech production*. Cambridge: Blackwell.

Levy, Yonata. 1983. It's frogs all the way down. *Cognition* 15. 75–93.

Lyster, Roy. 2006. Predictability in French gender attribution: A corpus analysis. *French Language Studies* 16(1). 69–92.

Nichols, Johanna. 1989. The Nakh evidence for the history of gender in Nakh-Daghestanian. In Howard I. Aronson (ed.), *The Non-Slavic Languages of the USSR: Linguistic Studies*, 158–175. Chicago, IL: Chicago Linguistic Society.

Nichols, Johanna. 1995. Diachronically stable structural features. In Henning Andersen (ed.), *Historical linguistics 1993. Selected papers from the* 11th *International Conference on Historical Linguistics, Los Angeles, 16–20 August 1993*, 337–355. Amsterdam: John Benjamins.

Nichols, Johanna. 2003. The Nakh-Daghestanian consonant correspondences. In Dee Ann Holisky & Kevin Tuite (eds.), *Current trends in Caucasian, East European and Inner Asian linguistics: Papers in honor of Howard I. Aronson*, 207–264. Amsterdam: John Benjamins.

Nikolaev, Sergey & Sergey Starostin. 1994. *A North Caucasian etymological dictionary.* Moscow: Asterisk.

Plaster, Keith & Maria Polinsky. 2007. Women are not dangerous things: Gender and categorization. In *Harvard working papers in linguistics,* vol. 12. 115–158. Cambridge, MA: Harvard University.

Plaster, Keith & Maria Polinsky. 2010. Features in categorization, or a new look at an old problem. In Anna Kibort & Greville Corbett (eds.), *Features: Perspectives on a key notion in linguistics*, 109–142. New York, NY: Oxford University Press.

Polinsky, Maria & Ezra van Everbroeck. 2003. Development of gender classifications: Modeling the historical change from Latin to French. *Language* 79. 356–390.

Polinsky, Maria & Dan Jackson. 1999. Noun classes: Language change and learning. In Barbara Fox, Dan Jurafsky & Laura A. Michaelis (eds.), *Cognition and function in language*, 29–58. Chicago, IL: University of Chicago Press.

Quinlan, J. Ross. 1993. *C4.5: Programs for machine learning.* San Mateo, CA: Morgan Kaufmann Publishers.

Rajabov, R. 1997. The class category in Tsez: Underlying principles. In *Dagestanskij lingvisticheskii sbornik*. Vyp. 4. Moscow: Academia.

Rajabov, R. Undated. Tsez dictionary (unpublished manuscript). Los Angeles: University of Southern California.

Saffran, Jenny, Elissa Newport & Richard Aslin. 1996. Word segmentation: The role of distributional cues. *Journal of Memory and Language* 35. 606–621.

Slobin, Dan I. 1973. Cognitive prerequisites for the development of grammar. In Charles Ferguson and Dan I. Slobin (eds.), *Studies of child language development*, 175–208. New York, NY: Holt, Rinehart & Winston.

Smoczyńska, Margaret. 1985. The acquisition of Polish. In D. I. Slobin (ed.), *The crosslinguistic study of language acquisition,* Vol. 1: *The data*, 595–685. Hillsdale: Lawrence Erlbaum Associates.

Sokolik, Margaret E. & Michael E. Smith. 1992. Assignment of gender to French nouns in primary and secondary language: A connectionist model. *Second Language Research* 8. 39–58.

Tanenbaum, Karen. E. 2003. *Modeling German gender* (unpublished manuscript). San Diego: University of California, San Diego.

Tucker, G. Richard, Wallace E. Lambert & A. Rigault. 1977. *The French speaker's skill with grammatical gender: An example of rule-governed behavior*. The Hague: Mouton.

Vigliocco, Gabriella, Marcus Lauer, Markus Damian & Willem Levelt. 2002. Semantic and syntactic forces in noun phrase production. *Journal of Experimental Psychology: Learning, Memory and Cognition* 28. 46–58.