Predicting grammatical gender in Nakh languages: Three methods compared

JESSE WICHERS SCHREUR[1], MARC ALLASSONNIÈRE-TANG[2], KATE BELLAMY[3], NEIGE ROCHANT[4]

[1]LEIDEN UNIVERSITY/GOETHE UNIVERSITY FRANKFURT/EPHE PARIS, [2]CNRS/MNHN/UNIVERSITY PARIS CITY (EA UMR 7206), [3]LEIDEN UNIVERSITY, [4]SORBONNE NOUVELLE/LACITO UMR 7107/LLACAN UMR 8135

Submitted: 11/03/2022 Revised version: 12/10/2022
Accepted: 20/11/2022 Published: 22/12/2022

Abstract

The Nakh languages Chechen and Tsova-Tush each have a five-valued gender system: masculine, feminine, and three "neuter" genders named for their singular agreement forms: B, D and J. Gender assignment in languages is generally analysed as being dependent on both form and semantics (e.g. Corbett 1991), with semantics typically prevailing over form (e.g. Bellamy & Wichers Schreur 2022, Allassonnière-Tang et al. 2021). Most previous studies have considered only binary or tripartite gender systems possessing one masculine, one feminine, and one neuter value. The five-valued system of Nakh thus represents a more complex and insightful case study for analysing gender assignment. In this paper we build on the existing qualitative linguistic analyses of gender assignment in Tsova-Tush (Wichers Schreur 2021) and apply three machine-learning methods to investigate the weight of form and semantics in predicting grammatical gender in Chechen and Tsova-Tush. Our main aim is thus to show how three different computational classifier methods perform on a novel set of non-Indo-European data. The results show that while both form and semantics are helpful for predicting grammatical gender in Nakh, semantics is dominant, which supports findings from existing literature (Allassonnière-Tang et al. 2021), as well as confirming the utility of these computational methods. However, the results also show that the coded semantic information could be further fine-grained to improve the accuracy of the predictions (see also Plaster et al. 2013). In addition, we discuss the implications of the output for our understanding of language-internal and family-internal processes of language change, including how loanwords are integrated from Russian, a three-gender language.

Keywords: grammatical gender; Nakh languages; computational classifiers; gender assignment.

https://doi.org/10.6092/issn.2785-0943/14545

1. Introduction

Humans intuitively categorise the world around them (Senft 2000). This categorisation emerges in language, such as through the presence of grammatical gender as a means of identifying and tracking referents in discourse (Contini-Morava & Kilarski 2013). Perhaps the most familiar gender systems are the binary masculine-feminine ones found in, for example, French, Spanish and Italian, or the tripartite masculine-feminine-neuter system of German. Yet gender is a pervasive grammatical category, being present in around half of the world's sampled languages (Corbett 2013). In a grammatical gender system, all nouns are assigned a gender, which is then formally reflected in agreement markers (their *exponence*) on other associated elements in the clause (Hockett 1958; Corbett 1991). These elements can include, but are not limited to, articles, adjectives or verbs, depending on the language in question.

Gender is assigned to nouns on the basis of semantic and/or formal (phonological and morphological) properties, with semantic features usually taking precedence over form (Corbett & Fraser 2000; Bellamy & Wichers Schreur 2022, Allassonnière-Tang et al. 2021). In some systems this assignment is completely transparent, whereas in others it is more opaque and, particularly for the L2 learner, can be difficult to systematise (e.g. Sokolik & Smith 1992 regarding French). Yet children acquire gendered languages of varying levels of transparency with equal ease, as they do languages possessing differing levels of complexity in other domains (cf. Karmiloff-Smith 1979). This begs the question, therefore, as to what the specific principles are that underpin these complex assignment systems. In this paper, we will use three computational methods to test which principles, or factors, most adequately predict the gender of nouns in Chechen and Tsova-Tush, two Nakh languages of the East Caucasian family that each possess a five-value gender system.

The present paper builds on a relatively small but expanding set of quantitative and qualitative studies concerning gender prediction in East Caucasian and other languages. Using Russian nominals as a case study, Corbett and Fraser (1993) introduced Network Morphology, a framework for analysing inflectional morphology. They followed this up with a more detailed analysis of the interrelation of meaning, gender, declension class and phonology in Russian, augmenting the Network Morphology approach with a lexical knowledge representation language known as DATR (Fraser & Corbett 1995, building on Corbett 1982; see also Evans & Gazdar 1989a 1989b 1996). The same approach has also been used to model the gender systems of Arapesh (Northern Papua; Fraser & Corbett 1997), Polish (West Slavic;

Brown 1998), and Mayali (northern Australia; Evans et al. 2002). While these studies provide intricate and informative representations of the gender systems they aim to model, they are not predictive in nature since the gender value constitutes one of the noun's attributes in their notation. As such, they are not testing previous assumptions in the same way that, later, predominantly tree-based approaches do.

Computational modelling of gender systems dates back to the early 1990s, notably Sokolik and Smith (1992) on French. In this study, the authors used a connectionist or 'parallel distributed processing' model to test "whether the information inherent within the structure of individual French nouns is sufficient to allow gender to be correctly assigned without reliance upon other types of information" (Sokolik & Smith 1992: 41). They selected 600 nouns (300 masculine and 300 feminine) from introductory French language textbooks, with 450 serving as the training set and the remaining 150 as the test set, to identify whether the model could assign the correct gender to nouns it had not yet encountered. In short, the model was able to do so with a high level of accuracy (over 76%), having 'learnt' that certain orthographic features correlate with either masculine or feminine gender. This suggests that L2 learners could also learn gender in a similar way, without having recourse to the often ambiguous cues appearing alongside a noun whose gender is not yet known, such as the singular prevocalic definite article *l'* and the plural definite article *les*.

Also in relation to French, and also using a connectionist model, Polinsky and van Everbroeck (2003) investigated the reanalysis of the Latin gender system as it transitioned to Old French. The simulations of the frequency-based neural network model aimed to uncover which factors are sufficient to lead to language change, in this case the evolution of the gender system in Old French. The authors built a training corpus from the 500 most commonly occurring nouns (excluding proper nouns and clear Greek borrowings) in the fifth-century Vulgate, in which the token frequency of each of the six possible case and number forms for each noun was calculated (for more details on building the corpus, see Polinsky & van Everbroeck 2003: 369-370). Using a feed-forward network with one hidden layer (see Section 3 for details of the neural network architecture used in the present study), the authors found that the model could adequately - around 60% at the end of the ninth generation - learn the gender of nouns in Late Latin, with a strong reliance of formal cues (i.e. case endings). Its decreasing performance over generations mirrored changes to the gender system largely instigated by phonological changes from Late Latin to Old French. Moreover, the model was also able to reflect the proposed influence of Gaulish (three genders) on gender assignment in Latin.

Bateman and Polinsky (2010) used decision trees to reconsider the gender system of Romanian, allowing them to posit a two-value rather than the traditional three-value analysis. They use the C4.5 Decision Tree Algorithm designed by Quinlan (1993 1996) to establish the rules of plural formation for Romanian nouns. Notably, the first cut relates to a semantic feature, namely whether the noun possesses masculine semantic features. The following cuts all pertain to formal features, that is, the final segment of the noun, its number of syllables, and whether the root contains a diphthong. A semantics-first assignment principle has already been observed in previous studies outlined in this section.

Indeed, Plaster, Polinsky and Harizanov (2012) applied a similar approach to what they call noun classification (and we call a gender system) in Tsez (Dido), an endangered East Caucasian language (Tsezic branch) that possesses a five-value gender system, like Tsova-Tush and various other languages of the family. Their 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" (Plaster et al. 2012: 7). Over 3,500 Tsez nouns (including loanwords), collected from dictionaries (Khalilov 1999; Rajabov, undated), were coded for formal (at least seven) and semantic (at least nine) features and then tested and run through the decision tree module of the "Orange" data mining tool (see Demsar et al. 2004). Their results demonstrate that the semantic features of a noun were most predictive of its gender, with such features overriding formal ones, as has been observed in other mixed assignment type systems (e.g. Corbett 1991). The decision tree model produced is able to predict around 70% of nouns in Tsez, with assignment to the remaining 30% possibly complicated by their status as loanwords or dialectal variants (Plaster et al. 2013: 11). Smaller semantic fields than those expressing core features such as animacy or biological sex, such as [berry] and [stone], are also highlighted as being predictive in the model. It is noteworthy, however, that the preference for semantic features in the computational model stands at odds somewhat with results from studies on Tsez language acquisition. Gagliardi and Lidz (2014) found that the Tsez-speaking children in their study had a preference for using phonological rather than semantic information for classifying nonce words, despite there being a statistical asymmetry that prefers the opposite. They suggest that this phonological bias relates to the higher value placed on phonological information in the intake: such information is available to children long before they know what a word means, and is more reliable than a semantic form (Gagliardi & Lidz 2014: 81).

The most recent and the most methodologically relevant previous study is Allassonnière-Tang, Brown and Fedden (2021), which applies three different computational classifier methods - simple decision trees, random forest trees and neural networks - to test the claim that Mian, an Ok language of Trans New Guinea, assigns gender predominantly on the basis of a noun's semantics. The accuracy in predicting the gender of the test nouns (30% of the nouns in the Mian dictionary used) was barely better than the majority baseline when only formal (i.e. phonological) features were fed to the computational classifiers. However, this improved considerably when only semantic features were fed in, and a little better again when both sets of features were used. The tree-based methods were the most accurate, with the random forest model only improving accuracy over a single tree by 1%. Moreover, the random forest method supports previous descriptive linguistic research on Mian (Fedden 2011), in that the top-ranked variables for predicting gender in the language are all semantic. Indeed, formal features only play a fine-grained discriminatory role in one semantic class, namely birds. These results support existing qualitative analyses of the Mian gender system as being semantics-dominant, as well as providing support for the methods used. The authors also highlight the importance of investigating and testing gender systems beyond those found in Indo-European languages, in order to contribute to the investigation of nominal classification more broadly. The present paper also represents a response to this call.

As such, the main aim of this paper is to show how three different computational classifier methods will perform on a novel set of non-Indo-European data. Some qualitative work on gender assignment in Nakh languages has been undertaken with limited results (see Section 2), therefore more fine-grained levels of description are needed. Although this paper contributes to finding semantic correlates to gender classes, our principal goal is a methodological assessment of the computational classifiers.

2. Background

2.1. Introduction to Chechen and Tsova-Tush

The Nakh languages form an outlying branch of the East Caucasian family (also known as Northeast Caucasian or Nakh-Daghestanian; Nichols 2003), and includes only three languages: Chechen (ISO 639-3 che), Ingush (ISO 639-3 inh) and Tsova-Tush (ISO 639-3 bbl). Chechen and Ingush are more closely related to each other than to Tsova-Tush, so in this paper we focus on one language from each sub-group within Nakh: Chechen and Tsova-Tush.

Chechen speakers mostly live in their ancestral homeland, Chechnya, located on the northern slopes of the Greater Caucasus Mountains and now a semi-autonomous republic of the Russian Federation. As of the 2010 Census, Chechens constitute 95.3% of the Chechen republic's population, and Standard Chechen is one of the two official languages of the republic alongside Russian. Some Chechen speakers can also be found in the Pankisi Gorge of neighbouring Georgia and in several villages in the Daghestanian lowlands.

Tsova-Tush (also known as Bats or Batsbi) is spoken in the village of Zemo Alvani in eastern Georgia by approximately 500 people who are all fluent in Georgian. These speakers have stopped transmitting Tsova-Tush to the next generation, which is why the language is considered severely endangered (Wurm et al. 2001). The Tsova-Tush ethnically self-identify as Georgian, and their language has been under the influence of Georgian for over four centuries (Desheriev 1953). Where Tsova-Tush shows little to no dialectal variation, Chechen has several distinct regional dialects. During the Soviet period, a written standard language was created for Chechen, whereas Tsova-Tush remains largely unwritten to this day.

Both Chechen and Tsova-Tush possess five gender classes, and agreement is marked by the same four consonantal prefixes. Agreement targets include a third of all underived verbs and a small number of adverbs, which agree with the nominative argument of the clause, as in Examples (1) and (2), as well as approximately ten underived adjectives and the numeral ‘four’, which agree with the head noun they are modifying (see Example (1)).

(1) Tsova-Tush (Kadagidze 2009: 44)
d-aqqo^n *xi* *d-ujt’-ŭ*
D-big water D-go-PRS
‘The big river is flowing.’

(2) Chechen (Nichols 1994: 37)
cħa *jiett* *āra* *b-ēl-ir*
one cow(B) out B-go-AOR
‘One cow went out.’

The way in which the agreement markers are distributed through the gender classes varies between Chechen and Tsova-Tush. Firstly, as shown in Table 1, Chechen has a unified human plural marker *b*-, whereas Tsova-Tush differentiates between masculine plural *b*- and feminine plural *d*-.

Gender class	**Semantics**	**Markers (sg/pl)**	
		Tsova-Tush	**Chechen**
M	male rationals	v- / b-	v- / b-
F	female rationals	j- / **d-**	j- / **b-**
B	animals, inanimates	b- / d-	b- / d-
J	animals, inanimates	j- / j-	j- / j-
D	animals, inanimates	d- / d-	d- / d-
Bb	animals, inanimates	b- / b- (6 items)	b- / b- (22 items)
Bj	body parts, ‘step’, ‘kick’	b- / j- (17 items)	-
Dj	body parts	d- / j- (6 items)	-

Table 1: Gender classes and their agreement markers in Tsova-Tush and Chechen.

Secondly, both Tsova-Tush and Chechen have several lexical items that show an agreement pattern that is different to the five main genders. In addition to several abstract nouns and nouns denoting materials that do not form a plural, both languages have words that require the agreement marker *b*- in both singular and plural contexts. Additionally, most Tsova-Tush nouns denoting body parts take *j*- in the plural and *d*- or *b*- in the singular. It has been argued that these groupings constitute additional gender classes, such that Tsova-Tush and Chechen have eight and six genders, respectively. Alternatively, these items can be viewed as exceptions or anomalies, and belong to so-called ‘inquorate’ genders, in that the number of items in these gender classes is too low to form a quorum (Corbett 1991: 170–175). Regardless of the analysis, these items will not be included in the present study.

2.2. Gender assignment in Chechen and Tsova-Tush

While the two human classes M and F mostly contain male and female rationals respectively, assignment to the three non-human classes, which can be called ‘neuter’, is far from fully predictable. On the one hand, it seems that semantics, morphology and phonology play a role in the assignment of the gender of non-rationals to one of the different neuter classes in both Tsova-Tush and Chechen. On the other hand, this role is very marginal, as the tendencies described below can only predict the gender of a small portion of non-rationals:

- Semantics: both Chechen and Tsova-Tush feature some semantically-based clusterings of genders, which allow the gender of only 15% of Tsova-Tush non-rationals to be predicted (Nichols 2007; Wichers Schreur 2021).
- Morphology: in Tsova-Tush, all verbal nouns (e.g. *st'exar* 'waiting' from *st'ex-* 'to wait') are assigned to the D class, and most de-adjectival abstract nouns (e.g. *must'ol* 'acidity' from *must'in* 'sour') are assigned the J class (Wichers Schreur 2021). Nichols (2007) also showed that a number of derived abstract nouns in Chechen are assigned D or J depending on the degree of abstractness (e.g. *goomalla* (D) 'crookedness, bend', (J) 'enmity, hostility' from *gooma* 'bent, crooked'). Inflectional morphology is not related to any gender assignment tendency.
- Phonology: Wichers Schreur (2021) showed that 60-65% of all Tsova-Tush nouns starting in *b-*, *d-* and *j-* belong to B, D and J gender classes respectively. This phenomenon is assumed to be a consequence of two processes: autogender and alliterative concord. The former refers to the phenomenon of fossilised gender markers on nouns themselves (cf. Nichols 2011: 147), while the latter refers to a phenomenon where the gender assignment of a noun is in fact influenced by its phonology (Corbett 1991: 117).

In cases where "neutral gender agreement" is required (see Corbett 1991: 205), Tsova-Tush agreement targets default to the D gender (Bellamy & Wichers Schreur 2022). Such an agreement pattern occurs in cases where the speaker leaves the gender of a human referent unspecified, when a word agrees with two or more human nouns that have different genders, or when agreement is with a clause. Additionally, a set of nouns denoting humans whose social gender is unspecified (e.g. 'friend', 'godparent', 'Christian') also triggers D agreement. However, these nouns are annotated as MF in the source dictionaries, a label which we have taken over. Neutral gender agreement is assumed to be highly similar in Chechen, as it is also found in Ingush (Nichols 2011: 435).

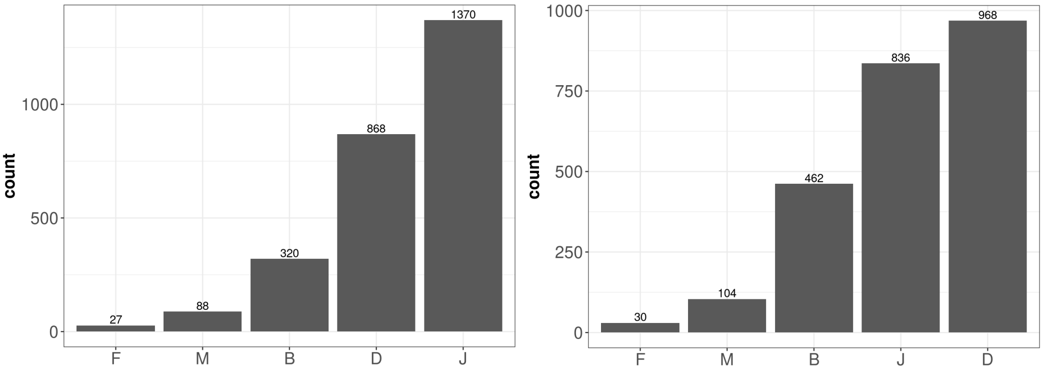
Gender assignment of loanwords has not been described in detail for Chechen, but Nichols (2007) notes that new loanwords are often given the gender of a near synonym or an immediate generic. However, a cursory query of the Chechen data (see below), shows us that all recent Russian loans have gender J. As regards Tsova-Tush loanwords, Wichers Schreur (2021) shows that they follow the same set of semantic and phonological tendencies as native nouns.

In sum, although a number of semantic, morphological and phonological tendencies have been observed, gender assignment for the vast majority of nouns in Chechen and Tsova-Tush remains poorly understood. This gap in our understanding holds for both qualitative and quantitative approaches, a void we aim to begin filling with the present study.

3. Data and Methods

3.1. Data

Two lexical databases were used in the present study. For Chechen we used the database developed by Chechen language specialist Erwin Komen, consisting of the combined dictionaries of Matsiev (1961) and Jamalkhanov and Aliroev (1991). The database contains 4339 nouns with gender annotation and Russian translation. After removing nouns that only occur in singular or plural, proper nouns, and nouns with ambiguous or inquorate genders, we arrive at 2673 items. We used these items as our dataset and annotated each item with a value "yes/no borrowed".[1] 298 items were classified as borrowings. The distribution of the gender categories found in the Chechen data is visualised in Figure 1a.



(a) Chechen

(b) Tsova-Tush

Figure 1: The distribution of gender categories in different languages of the data.

[1] Where "yes" means 'obvious borrowing' (i.e. the form differing in one or two segments from its Russian translation), and "no" 'not known to be a borrowing'.

For Tsova-Tush, we used the Kadagidze and Kadagidze (1984) dictionary, containing 2775 nouns with gender annotation and Georgian and Russian translation. After removing nouns that only occur in singular or plural, proper nouns, and nouns with ambiguous or inquorate genders, we arrive at 2400 items. The distribution of the gender categories found in the Tsova-Tush data is visualised in Figure 1b. We translated all nouns to English and annotated each item with a value "yes/no borrowed".[2] Within all the items, 1365 are marked as borrowed from Georgian, whereas 72 additional items are marked as borrowed from other languages (mostly Russian).

For both datasets, several annotations were added that classify items in semantic categories. Firstly, the following broad semantic categories were added: Male, Female, Human (i.e. gender-unspecified human), Animal, Inanimate. For example, in the Tsova-Tush dataset, we marked 2081 items as Inanimate, 175 items as Animal, 104 items as Male, 13 as Human, and 27 as Female. Secondly, English translations were matched against their corresponding Concept Sets in the online Concepticon (List et al. 2016). This database was developed to serve as a reference concept list to aid in studies on semantic change, cross-linguistic polysemies, and semantic associations. Both the abstract concept (Concept Set in Concepticon's terminology), as well as the semantic category of this concept (e.g. 'agriculture and vegetation' or 'the body') were added as annotations to items in the Tsova-Tush and Chechen databases where the English translation matched a Concepticon Concept Set. This was possible for 57% (1363/2400) of Tsova-Tush items and 44% (1164/2673) of Chechen items. Thirdly, a more fine-grained semantic annotation was added for 69% (1655/2400) of Tsova-Tush items and 1490/2673 (56%) for Chechen. These annotations consisted of semantic domains we chose, to represent the common practice of linguists inventing semantic domains based on intuitive and language-internal principles. As an example, within the Tsova-Tush data, 228 items are annotated as 'Abstract', 87 items are marked as 'Person', 137 as 'Implement', and 96 as 'Food'. Annotating our data with three independent semantic layers allows us to test which is preferred by the different computational methods described below. It is necessary to distinguish between synonymous labels of different annotation layers, i.e. between the label [person] in the fine-grained semantic domain layer (indicating nouns that refer to professions and other identities that humans have, other than kinship relations and ethnicities),

[2] Where "yes" means 'obvious borrowing' (i.e. the form differing in one or two segments from its Georgian or Russian translation), and "no" 'not known to be a borrowing'.

Concepticon's label [person/thing] (as opposed to [action/process]), and the label [human] in the layer in the broad semantics layer (indicating any human that is not explicitly female or male).

In terms of form, we included the information of the first three and the last three phonemes of each noun. This choice was motivated by the fact that nominal features such as gender are generally found at the start and/or the end of nouns (Dryer 2013; Basirat et al. 2021). Information regarding word length was also included. As an example, the first three phonemes of Tsova-Tush *haer* 'air' are /h/, /a/, and /e/, and its word length is counted as 4. An example of the raw data used for Tsova-Tush is provided in Table 2. The full raw data is provided in the supplementary materials.

Noun	haer	mar
Gloss	air	husband
Gender	J	M
Concepticon_category	Person/thing	Person/thing
Concepticon_field	Physical world	Kinship
Semantic_broad	Inanimate	Male
Semantic_domain	Natural	Kinship
Borrowed_Arab	0	0
Borrowed_GE	1	0
Borrowed_Turk	0	0
Borrowed_Russian	0	0
Word length	4	3
Last first phoneme	r	r
Last second phoneme	e	a
Last third phoneme	a	m
First phoneme	h	m
Second phoneme	a	a
Third phoneme	**e**	r

Table 2: A sample of the raw Tsova-Tush data used in this paper.

3.1. Method

We used three computational classifiers to evaluate the predictive power of form and semantics on the grammatical gender of nouns in Chechen and Tsova-Tush. The first two classifiers apply the method of binary recursive partitioning (Breiman et al. 1984). First, one of the classifiers generates a decision tree by recursively partitioning the data in a binary way to create homogeneous groups. The output is represented as a decision tree that allows us to visualise the hierarchical interaction of the semantic and formal variables when it comes to predicting the grammatical gender of nouns. As an example, if both a formal and a semantic feature are helpful for predicting the grammatical gender of nouns, the decision tree will display which variable should be considered first during the decision-making process. Such a tree could be subject to overfit when it comes to multiple replications, therefore the second classifier generates a forest of trees instead of a single tree, hence its name: *random forests*.

The random forests computational classifier creates a forest of 300 decision trees that are considered as a whole to evaluate the relevance of formal and semantic variables for predicting the grammatical gender of nouns. This classifier uses the same algorithm as a single decision tree, but it creates a forest of trees instead of only one tree. For each tree in the sample of 300 trees, the classifier takes a bootstrap sample of the data along with a subset of the variables of the data. In other words, the classifier extracts a random subset of rows and columns in the data to generate each of the 300 trees. A statistical test carried out for each sample shows if an interaction between some variables is consistently observed across all the samples. This process of random sampling is one of the main strengths of random forests, as it allows for the analysis of different data sizes (ranging from small to large), as well as the consideration of potential auto-correlation between variables (Tagliamonte & Baayen 2012). Another main strength of decision tree-based classifiers such as random forests, is that they allow for a relatively transparent understanding of the interaction between the variables. For example, such classifiers have been used to investigate linguistic universals (One-Soon & Tang 2020), automatic identification of sounds based on phonetic features (Ulrich et al. 2021), tone paradigms (Lemus-Serrano et al. 2021), and also the gender affiliation of nouns (Allassonnière-Tang et al. 2021). More specifically, random forests provide information on the relative importance of the predictor variables. The larger the importance of a variable, the more predictive it is. For instance, if the accuracy of the classifier drops the most when it does not take into account a specific feature, this feature can be considered to have the highest ranking within all the variables.

However, the transparency of decision-tree based classifiers can also be their weakness, as they might fail to capture extremely complex interactions between the variables. Therefore, the third classifier we consider in our study has a neural network architecture (Haykin 1998; Parks et al. 1998). *Neural network* is a non-linear discriminative classifier that identifies boundaries between the data points with regard to their predicted variable. While it is much more complex to extract the interactions between the variables using a neural network, we can use such a classifier to assess whether decision-tree based classifiers missed some non-linear information encoded in the data. For instance, if decision-tree based classifiers and neural networks reach a similar level of accuracy for predicting the grammatical gender of nouns, we can assume that both classifiers are capturing most of the information encoded in the data and that additional information is needed to further improve the performance of the classifiers.

In this study, we use a simple feed-forward neural network that consists of an input layer, a hidden layer, and an output layer. Each layer has a specific number of neurons that are connected to each other. The input layer has one neuron for each predictor (i.e., each variable). The number of hidden layers and their quantity of neurons is flexible. The size of the output layer is equal to the number of categories to predict. As an example, when predicting gender in Tsova-Tush, the output layer has five neurons, which represent the gender categories M, F, B, D and J. In our experiments, we set the size of the hidden layer to ten. More experiments could be conducted to finetune the number and size of hidden layers. For example, we could have an architecture with hundreds of neurons and a dozen hidden layers. Nevertheless, since we are only interested in the relative performance of formal and semantic features, we do not perform such experiments in this study.

Some parameters are shared across all of the classifiers. First, each classifier is trained with 70% of the data. Then, the trained classifier is evaluated when predicting the other 30% of the data that it did not encounter during the training. That is to say, the training set and the test set do not overlap. If a noun is used in the training set, it will not be used in the test. To avoid coincidental biases from the random sampling between the training and test sets (which can never be precisely 70% and 30% respectively), this sampling and evaluation process is repeated ten times for each classifier. If the results are similar across the ten replications, it shows that the results are robust. We also tested the experiments based on 100 and 1000 replications, which gave similar results. For sake of simplicity and to avoid over-computation, we thus report the results based on ten replications. The code in the supplementary materials allows for additional tuning of the parameters, such as conducting more iterations and only selecting specific gender categories for making predictions.

The performance of the classifiers is evaluated using accuracy, precision, and recall. First, the accuracy shows how good the classifier is at predicting grammatical gender on the data. The accuracy is equal to the ratio of all the correctly retrieved tokens within the entire data. For example, if the classifier correctly predicted the gender of 70 out of 100 nouns, the accuracy of the classifier is 70%. This accuracy value needs to be compared to a baseline in order to be interpreted as good or bad. In this study, we compare the accuracy to a random baseline and a majority baseline. The random baseline constitutes what the model would obtain by making totally random guesses. Taking a binary classification as an example, we could imagine a model that has to guess if a coin toss will be heads or tails. The random baseline is calculated by adding the probability of heads in the data multiplied by the probability that you guess heads, with the probability of tails in the data multiplied by the probability that you guess tails. Assuming that the coin is fair and that there is a 50-50 chance to get heads or tails, the random baseline equals to 0.5*0.5 + 0.5*0.5 = 0.5. Taking Chechen as an example, the random baseline is equal to the sum of the square of the proportion of each gender category in the data, i.e., (320/2673) * (320/2673) + (868/2673) * 868/2673) + (27/2673) * (27/2673) + (1370/2673) * (1370/2673) + (88/2673) * (88/2673) = 38.3%. This accuracy is what the model would guess by making a random guess for each of the nouns in the data. That is to say, for each noun, the model has a probability to guess that the noun belongs to a gender category based on the proportion of nouns belonging to that gender category. Conducting a random shuffling of gender labels of the nouns would also get this level of accuracy. If the accuracy of a classifier is above this baseline, it performs better than chance.

We also consider the majority baseline, which is generally higher (and thus harder) than the random baseline. The majority baseline represents what the model would obtain by making an informed guess and affiliating each noun in the test set to the largest gender category in the data. For example, in Chechen, the J gender category is equal to 51.3% (1370/2673) of the data. The classifier could thus reach an accuracy of 51.3% just by guessing that all nouns belong to the J gender category. Therefore, the accuracy of the classifiers trained with the information of forms and/or semantics should at least exceed the accuracy of 51.3% to be considered as having good discriminatory power.

Second, precision and recall are used to assess the performance of the classifier on each of the gender categories in the target language. Precision assesses how many tokens are correct of all the tokens assigned to a gender category by the classifier, while recall evaluates how many tokens belonging to a gender category are correctly retrieved by the classifier. These measures are used in a similar way as Suppliance in Obligatory Context (SOC) and Target-Like Usage (TLU) in studies of language acquisition (Ting 2010). Using values such as precision and recall allows us to have a more precise understanding of which gender categories are more difficult to predict for the classifiers, as opposed to the accuracy, which only provides an overview of the performance of the classifiers.

4. Results

In the following subsections, we present the quantitative analyses for both Chechen and Tsova-Tush. For each language, three computational classifiers are trained and tested ten times: single decision tree, random forests, and neural networks. Each classifier is fed with three types of input data. First, the classifier is trained with information on form to predict gender in a given language. Second, the classifier is trained with semantic information to predict gender. Finally, the classifier is trained with both formal and semantic information.

4.1. Chechen

The accuracy of each combination of parameters is first compared to visualise the effect of form and semantics for predicting the gender of nouns. In Figure 2, each point represents the accuracy of a parameter across its ten iterations. First, we observe that the accuracy of a parameter does not vary much visually across the ten iterations, showing that the classifier is likely not influenced by random splits between the training and test sets. Second, we observe that the combination of formal and semantic information generally results in the highest accuracy for each classifier (random forests, single tree, neural networks). The performance of neural networks is not higher than decision trees, showing that the output of decision trees is capturing as much information as can be captured in the variables of the data. Third, the accuracy of the classifiers is far above the random baseline, demonstrating that the interaction of the variables captured by decision trees can be further analysed.

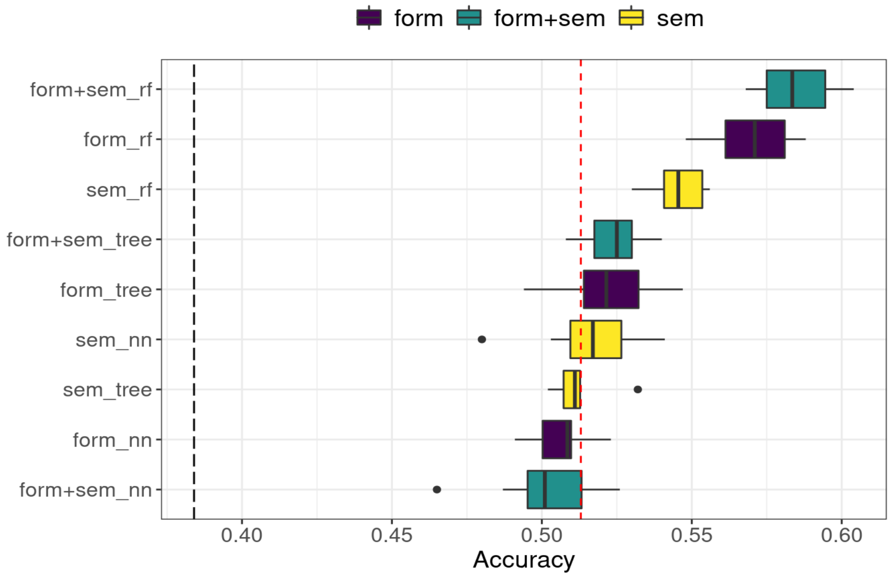


Figure 2: The accuracy of different parameters for predicting the gender of nouns in Chechen. The abbreviations are interpreted as follows: sem = semantics, rf = random forests, tree = single decision tree, nn = neural networks. The red dashed line indicates the majority baseline. The black dashed line indicates the random baseline.

Figure 2 only shows the overall accuracy of the model. However, it is also important to analyse how the classifiers perform on each gender category found in Chechen. For example, a classifier could have a high accuracy but only be good at predicting two gender categories, even when a language possesses more categories than that. Therefore, in Table 3 we visualise the precision and recall of each classifier on each gender category found in Chechen. We can observe that the classifiers have difficulties identifying items from the categories B and F. This is expected from a quantitative point of view, as these categories have a smaller number of tokens in the data. The detailed confusion matrices of the classifiers are provided in the supplementary materials.

Since the performance of combining formal and semantic information results in the highest accuracy for both single tree and random forests classifiers, we only display the output of combining formal and semantic information in this paper. However, the detailed output of each parameter with each classifier is provided in the supplementary materials.

Classifier	**Setting**	**Mean Acc**	**Pr/Rc [B]**	**Pr/Rc [D]**	**Pr/Rc [F]**	**Pr/Rc [J]**	**Pr/Rc [M]**
Tree	form	52.2 (51.2-53.3)	0/0	0/14.7	0/0	53.1/92.5	0/0
Tree	sem	51.2 (50.6-51.7)	0/0	0/5.3	0/0	52.0/96.5	0/8.6
Tree	form + sem	**52.4 (51.6-53.1)**	0/0	0/7.2	0/0	52.3/97.1	0/12.2
RF	form	56.9 (56.0-57.9)	46.5/13.6	50.5/45.4	0/0	60.5/78.7	61.7/4.7
RF	sem	54.5 (53.9-55.2)	0/1.0	46.8/22.9	0/23.8	56.4/87.8	54.7/56.7
RF	form + sem	**58.5 (57.6-59.4)**	49.8/9.3	52.8/44.9	73.8/36.2	61.0/80.8	64.0/33.9
NN	form	50.6 (49.9-51.3)	26.1/20.9	44.3/40.9	0/0	58.9/68.0	0/0
NN	sem	**51.7 (50.4-52.9)**	20.8/9.3	43.3/35.7	0/17.3	58.5/72.8	49.0/44.7
NN	form + sem	50.1 (48.9-51.4)	23.5/21.0	45.2/46.3	0/0	61.2/62.4	20.8/18.3

Table 3: The performance of the classifiers across ten replications ranked according to their mean accuracy when predicting gender in Chechen. The numbers in parentheses indicate the upper and lower confidence intervals of the accuracy. The abbreviations are interpreted as follows: Acc = accuracy; Pr = precision; Rc = recall. The values in bold indicate the parameters with the highest accuracy for each classifier.

In Figure 3, we first present a single decision tree that is generated when feeding both formal and semantic information to the classifier. Since the accuracy of the ten iterations does not vary much, we consider the last tree as an example. However, it is important to point out that this tree is displayed as a visualisation of how the classifier works rather than an absolute truth of gender assignment in the language under discussion. The tree can be read in the following way: the buckets at the bottom of the tree indicate with different colours the predicted gender. For example, the nouns predicted as M by the classifier are coloured in green. The numbers in each bucket indicate the number of predictions and how many of them are correct. As an example, in node 3 (bottom right) 206 nouns are predicted as J by the classifier, and of those 206 nouns, 173 do indeed belong to J gender. Each prediction is read starting from the top of the tree, descending until a bucket is reached. As another example, if the word is not a Russian loan (node 1 to node 2), but the semantic field is kinship (node 2 to node 5), the predicted gender is M. Following the path, 32 nouns are predicted as M, and 16 of them indeed belong to the M gender, which results in an accuracy of 16/32 = 50%. This accuracy is quite low, although it is important to emphasise that this accuracy should be compared with either the random baseline (38.4%) or the proportion of the predicted category. As an example, node 5 relates to gender M, which has a proportion of 3.3% in the entire data. The same reading can be applied for the other branches of the tree.

Semantic features are found in the tree: the tree considers the information of ‘Kinship’ (node 2) and ‘Person’ (node 8). As an example, if a noun is coded as possessing kinship semantics, it is more likely to be the M category. Formal features are found lower in the tree. For example, if a noun ends with /e/ (node 4), it is more likely to be from the J category. As another example, if a noun starts with /d/ (node 16), it is more likely to be from the D category (recall Section 2). Since different trees generated by the classifier could result in the use of different variables, we consider the forest of trees generated by the random forests classifier and extract the importance of variables based on three metrics.

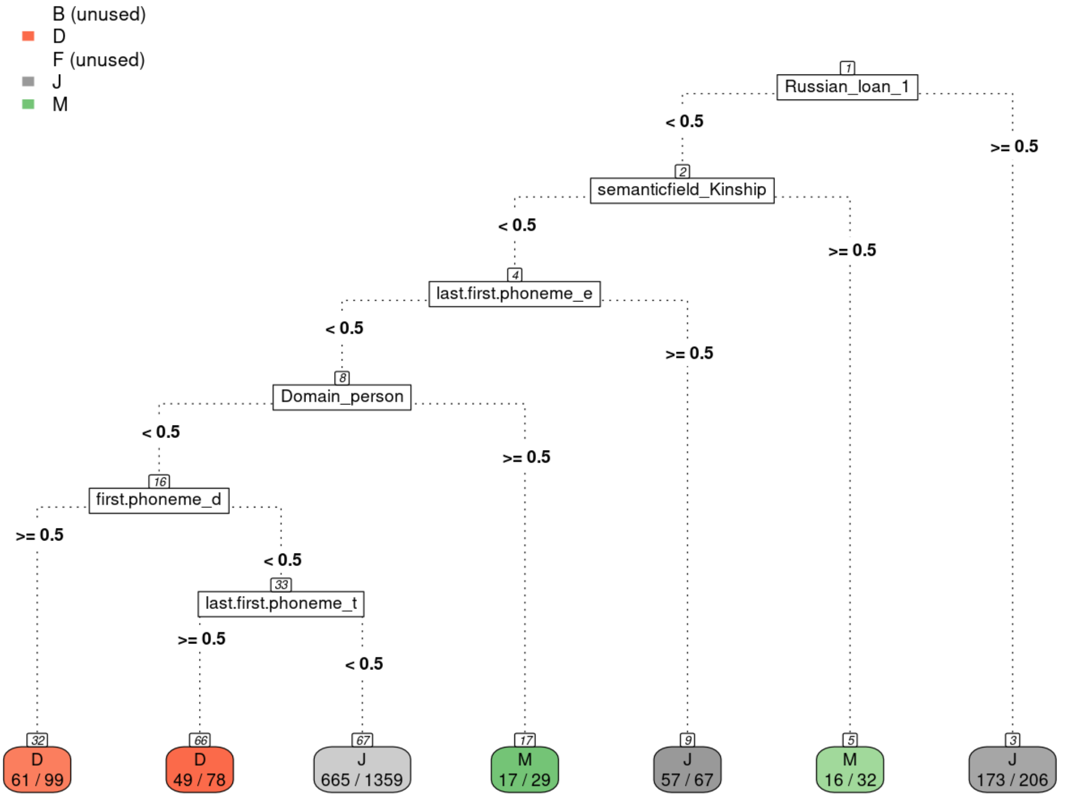


Figure 3: One of the ten decision trees generated for predicting gender in Chechen based on formal and semantic information. Labels marked as unused refer to categories that the models did not predict based on this sample tree.

First, we consider the minimal depth of a variable in a decision tree. For example, in Figure 3, the variable ‘Russian loan’ is at the top of the tree, which means that it has a depth of 0. As another example, the semantic category Kinship appears immediately after the root of the decision tree, which gives it a depth of 1. The closer a variable is found to the root, the larger group it creates within the data, which gives it higher importance. Second, we consider the decrease of accuracy for the classifier when a variable is removed. For example, if the tree has an accuracy of 60% with all variables considered, and the accuracy drops by 30% if we remove the semantic category of Person, it means that the semantic category Person is highly relevant for the classification task. Third, we consider the decrease of purity in predictions (i.e., the Gini coefficient). As an example for calculating the Gini coefficient, if the predictions of a node are all correct, it results in a high purity of the predictions and therefore a high Gini coefficient. In contrast, the predictions of node 5 have a lower purity/Gini coefficient, as they are correct at 50% (16/32). The top ten most important variables according to these three metrics are shown in Table 4. The full ranking of the variables is provided in the supplementary materials.

Ranking	Minimap depth	Mean decrease accuracy	Mean decrease Gini coefficient
1	Word length	Person	Word length
2	Person.thing	Russian.loan_1	Russian.loan_1
3	Russian.loan_0	Kinship	Person
4	Russian.loan_1	First.phoneme_d	First.phoneme_d
5	Person	Last.second.phoneme_u	Last.first.phoneme_i
6	First.phoneme_d	Last.first.phoneme_e	Person.thing
7	Last.first.phoneme_a	Ethnicity	Russian.loan_0
8	Second.phoneme_a	Word length	Last.first.phoneme_a
9	Third.phoneme_r	Last.first.phoneme_i	Third.phoneme_r
10	Inanimate	Wild.plant	Kinship

Table 4: The importance of variables in random forests according to different metrics for predicting gender in Chechen. Only the top ten variables for each metric are listed. The variables highlighted in grey are the variables that are found in the top ten of all three metrics.

Four variables are found consistently across the three rankings. First, borrowing from Russian has an effect on gender assignment. Second, the variable ‘Person’ shows that the human/inanimate distinction is relevant for grammatical gender assignment. Finally, in terms of form, we also see that word length is relevant, combined with the use of /d/ as the first phoneme. Additional linguistic interpretation of these variables is provided in Section 5.

4.2. Tsova-Tush

The same analysis was conducted for the Tsova-Tush data. First, we compare the accuracy of each combination of parameters. see Figure 4. As found with Chechen, we see that the accuracy does not vary much across the ten iterations, which indicates that the output of the models is stable. We can observe that, as in Chechen, combining formal and semantic information results in the highest accuracy for each of the models. As an example, the accuracy of random forests is at its highest when combining formal and semantic information. The accuracy of all models is also above the majority baseline.

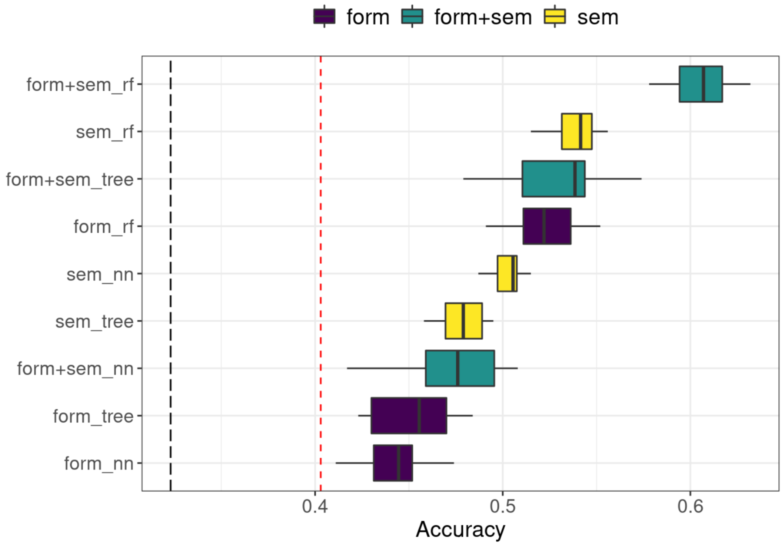


Figure 4: The accuracy of different parameters for predicting the gender of nouns in Tsova-Tush. The abbreviations are interpreted as follows: sem = semantics, rf = random forests, tree = single decision tree, nn = neural networks. The red dashed line indicates the majority baseline. The black dashed line indicates the random baseline.

The detailed performance of each classifier and each combination of parameters is shown in Table 5. In general, we see that the classifiers perform well on the M and F categories, but have more difficulties identifying the nouns from the B, D, and J categories. For example, the precision and recall of the gender category J are generally low, since the model did not correctly identify the majority of nouns belonging to this category. The detailed confusion matrices of the classifiers are provided in the supplementary materials.

As with Chechen, in the following analyses we only consider the models that combine formal and semantic features, since they have the highest accuracy. Nevertheless, the details regarding other combinations of parameters are included in the supplementary materials. In Figure 5, we visualise one of the ten decision trees generated based on formal and semantic information for predicting the gender of nouns in Tsova-Tush. The decision tree uses both formal and semantic variables. In terms of formal features, the tree considers the first phonemes /b/ and /n/, along with the penultimate phoneme /a/, and word length (the cut-off point being five phonemes). With regard to semantic features, the tree uses the broad semantic categories of Male, Female, and Inanimate. It also considers the more specific category of ‘Wild plants’ and ‘The physical word’.

Classifier	**Setting**	**Mean Acc**	**Pr/Rc [B]**	**Pr/Rc [D]**	**Pr/Rc [F]**	**Pr/Rc [J]**	**Pr/Rc [M]**
Tree	form	45.2 (43.6-46.8)	55.7/15.9	46.4/66.8	0/0	41.8/43.1	0/0
Tree	sem	47.9 (46.9-48.7)	0/0.6	50.4/59.4	0/65.5	42.0/54.1	100/100
Tree	form+sem	**52.8 (50.7-54.9)**	56.9/15.9	52.9/63.7	0/67.3	46.5/54.0	100/100
RF	form	52.2 (50.9-53.4)	52.5/29.1	52.1/70.4	0/0	52.0/50.6	85.7/13.2
RF	sem	53.9 (52.9-54.7)	44.1/11.7	54.0/64.7	90.8/84.9	49.2/57.9	98.4/100
RF	form+sem	**60.7 (59.4-61.9)**	62.6/27.1	58.5/71.8	93.0/76.7	57.6/60.6	99.8/100
NN	form	44.4 (42.9-45.8)	33.6/25.7	48.6/56.6	0/1.7	45.0/47.3	0/3.3
NN	sem	**50.3 (49.7-50.9)**	34.2/15.0	53.8/53.6	78.8/76.7	46.0/59.8	94.0/95.3
NN	form+sem	47.4 (45.2-49.5)	33.0/33.7	51.9/57.8	0/24.8	49.2/42.8	0/57.5

Table 5: The performance of the classifiers across ten replications ranked according to their mean accuracy when predicting gender in Tsova-Tush. The numbers in parentheses indicate the upper and lower confidence intervals of the accuracy. The abbreviations are interpreted as follows: Acc = accuracy; Pr = precision; Rc = recall. The values in bold indicate the parameters with the highest accuracy for each classifier.

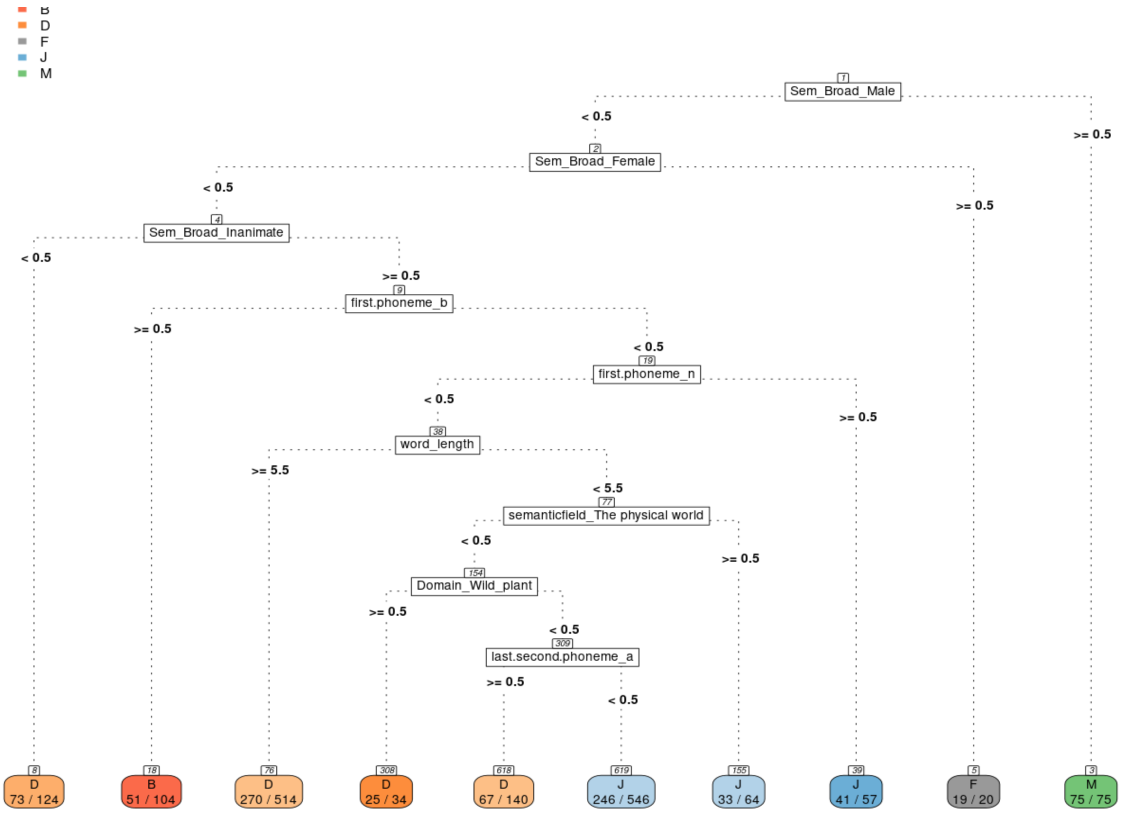


Figure 5: One of the ten decision trees generated for predicting gender in Tsova-Tush based on formal and semantic information.

We then consider the output from the random forests classifier to consider the overall importance of the variables when generating a forest of trees. We take into account the three metrics minimal depth, mean decrease of accuracy, and Gini coefficient. The top ten most important variables according to these three metrics are listed in Table 6. The full ranking of the variables is provided in the supplementary materials.

We can observe that the recurrent variables are word length, the first phoneme /b/, the penultimate phoneme /a/, along with the semantic information of male and inanimate. The information of Person also seems to be relevant, as it is encoded by different variables that are found in the rankings. For example, the metric of mean decrease of Gini coefficient has the Concepticon category ‘Person.thing’ in its top ten, while the rankings of accuracy and purity consider the dictionary information regarding person. Additional linguistic interpretation of these variables is provided in Section 5.

Minimap depth	**Mean decrease accuracy**	**Mean decrease Gini coefficient**
Word_length	Male	Male
Borrowed_GE	Inanimate	Word_length
Inanimate	First.phoneme_b	Inanimate
Person.thing	Word_length	Person
Last.second.phoneme_a	Female	Borrowed_GE
First.phoneme_b	Abstract	First.phoneme_b
Second.phoneme_a	Animal	Last.second.phoneme_a
Male	First.phoneme_d	Person.thing
Last.first.phoneme_r	Person	Female
First.phoneme_d	Last.second.phoneme_a	Animal

Table 6: The importance of variables in random forests according to different metrics for predicting gender in Tsova-Tush. Only the top ten variables for each metric are listed. The variables highlighted in grey are the variables that are found in the top ten of all three metrics.

5. Discussion

5.1. Methodological and technical aspects

In terms of accuracy, we observe that the classifiers generally perform above the random and the majority baseline, which indicates that interpreting the interaction of the variables detected by the classifiers is worthwhile. However, the accuracy stabilises at around 60%, which shows that the classifiers are still far from being able to perfectly predict the gender of nouns. While similar results are found in the literature, widening and deepening the variables is likely to improve the performance of the classifiers. On the one hand, additional variables related to form and semantics could provide additional information to the classifiers. For example, it is possible that adding frequency information could help the classifiers to find different rules for frequent and less-frequent words. On the other hand, the semantic information used in our experiments could be further refined. As an example, the semantic domains could be annotated as individual variables rather than being merged as a single variable, as a noun may belong to different semantic categories simultaneously. Furthermore, additional tuning of the classifiers is also likely to improve their accuracy. In the current experiments, we did not tune the classifiers to find the parameters that fit best the gender classification task. Running additional tests to identify those parameters will also improve the accuracy of the classifiers. As an example, preliminary tests show that allowing the tree to split the data into smaller buckets (e.g., buckets of two units) and/or allowing the tree to go much deeper (i.e., have more branches) does not alter the general results. Nevertheless, additional testing could help to identify the settings that result in the highest possible accuracy.

5.2. Language-internal and family-internal aspects

For Tsova-Tush, the results in Figure 5 and Table 6 corroborate findings by Bellamy and Wichers Schreur (2022) and Wichers Schreur (2021): Firstly, in our classifier experiments, the variable “Male” has the most predictive power when all metrics are considered (Table 5), pointing to 100% accuracy in Table 6. It is known from corpus research (Wichers Schreur 2021) that the variable “Female” holds the same predictive power. That this is not shown in our outcomes is likely due to the low number of items in this category. Semantically, we find few other variables that serve as cues for gender assignment. Only the label “wild plant” points to gender D, which was not observed in previous literature, and the broad label “Inanimate” (after subtracting phonological clues and “wild plants”) points to gender J. In terms of phonology, we see that nouns starting in *b-* correlate with gender B, as expected from previous qualitative and descriptive studies. Also as in Tsova-Tush, in terms of phonology, we find words starting in *d-* to correlate with gender D.

The variable “word length” (i.e., possessing more than 5 phonemes) is an unexpected predictor, especially since it scores highly in all three metrics presented in Table 5. Word length corresponds highly with gender D, which we explain by the high percentage of abstract nouns formed by suffixes (increasing the word length), many of which trigger gender D (see Wichers Schreur 2021). Conversely, one suffix forming abstract nouns that triggers J gender is the suffix *-ol*, which was unexpectedly not used as variable by any of the decision trees. As expected from previous literature, loanword status is not used as a variable, as it does not correlate with any particular gender.

In our Chechen results, “word length” is similarly an important variable, although in this language it is correlated with gender J. It remains unclear whether there is a larger number of suffixes deriving J gender nouns in Chechen (as opposed to Tsova-Tush), or whether there are simply more attestations of such suffixes in our dataset; further research is thus necessary in this regard. Additionally, the majority of Russian loanwords (coded as such, see Section 3.1) in the dataset are assigned gender J. As indicated in Figure 1a, J is the largest gender category in Chechen, comprising just over half (51.3%) of the nouns included in the present study. It is hard to detect any formal (i.e. phonological) or semantic patterns in this group of loanwords, therefore it is tempting to suggest that they are being assigned to the unmarked gender (also referred to as a ‘default assignment strategy’; Bellamy & Parafita Couto 2022; Corbett 1991: 77; Poplack et al. 1982: 21-23). That said, in reference to German loanwords in Russian, Corbett (1991) argued that the morpho-phonological form of the former dictated their assignment to the appropriate declension classes of the latter. Most of the exceptions to the J gender assignment rule are animate nouns, which can more transparently be assigned a gender in the recipient language, especially since Chechen possesses clear M and F categories. Nonetheless, further investigation is needed to verify whether the inanimate exceptions, which represent 22% of Russian borrowings in this dataset, could be explained by semantic analogy or other factors (ibid.).

Both the Chechen and Tsova-Tush datasets contain clusters of items correlating strongly with one particular gender, which are not used by the decision tree algorithm and which score low on all importance metrics (Tables 4 and 6). Examples of this include abstract nouns in *-lo* (J) in Tsova-Tush, and nouns annotated as “Male” (M) in Chechen, and nouns annotated as “Female” (correlating with F) in both languages. The fact that these variables are not used can be explained by (i) the number of items in a category being too low; (ii) this cluster being split up into other categories that are deemed more important by the algorithm; or (iii) there was no need to explicitly separate this cluster in the decision tree, as it is subsumed under larger categories, most importantly “inanimate” and “word length”.

These results tie in well with hypotheses formulated elsewhere regarding gender assignment in East Caucasian. Particularly, they corroborate the observation that broad semantic categories such as animacy, humanness and abstractness show a high correlation with certain gender classes (Carling et al. 2021). On the other hand, whereas more fine-grained categories such as “animal”, and “metal” show a general tendency of correlating with certain gender classes in East Caucasian (B for animals, D for metals), no such tendency has been found in our experiments. Animals in Nakh are divided between the three neuter classes, and metals do correlate with gender D, but form too small a category to be picked up by the models used here.

5.3. Broader linguistic implications

Other studies, both qualitative and quantitative, have shown that semantics plays a more important role in gender assignment than (phonological-morphological) form (see overview in Section 1), despite many systems relying on both features to different degrees (e.g. Corbett 1991). Indeed, we found that the form alone of a noun is always less predictive than purely its semantics, but that form and semantics together are always more informative, in all models. The Nakh languages therefore sit towards the semantic end of the spectrum of possible gender assignment types, albeit not as categorical as, for example, Mian (Allassonnière-Tang et al. 2021). An advantage of the computational methods presented here is that they can test and validate, or not, descriptive analyses of gender assignment systems. Moreover, they can also identify the relative weights of semantics and form in assignment, as well as the relative importance of the various semantic domains involved.

However, the prevalence of semantics-based assignment in the Nakh systems, amongst many others (such as Mian), seems difficult to reconcile with the observation that children pay more attention to phonetic cues when acquiring gender systems, especially those found noun-externally, namely the agreement markers that occur on other elements in the clause (see Gagliardi & Lidz 2014 for such evidence from Tsez, another Daghestanian language). This reliance on phonetic input is to be expected, since in the earliest stages of language acquisition, children do not have access to the meanings of all the input they receive. There appears, therefore, to be a discrepancy between the earliest stages of assignment and later processes, including the integration of loanwords, nonce words or neologisms into the language. Such a discrepancy begs the question as to what happens with regard to the processing and storage of gender assignment information between (early) childhood and adulthood. How can different cues gain greater prominence as age and language develop? More comparative studies of children of different ages and adults speaking the same language are required to investigate this issue further.

That said, certain semantic cues are likely to be prominent in gender assignment from early childhood, notably animacy. We have seen how Nakh languages have clear masculine and feminine animate categories (labelled M and F respectively), and humanness is almost exceptionlessly diagnostic for gender assignment in many Indo-European languages. Moreover, there is evidence from code-switched speech that animacy is also the most important feature of a noun from one language inserted into a phrase or clause of another. Cruz (2021) highlights how English nouns representing feminine animates are assigned feminine gender when inserted into Spanish, despite there otherwise being a preference for a default masculine assignment strategy in inanimates (cf. Balam 2016 who finds feminine animates also assigned masculine gender in code-switching mode).

Finally, it should be stressed here that computational models of gender assignment do not necessarily represent mental classification, and an interdisciplinary approach to investigating gender assignment still remains to be undertaken.

6. Conclusion

The main aim of this paper was to show how three different computational classifier methods perform on a novel set of non-Indo-European data. We applied three machine-learning methods to investigate the relative weight of (phonological) form and semantics in predicting grammatical gender in the Nakh languages Chechen and Tsova-Tush. The results showed that the combination of form and semantics gives the best results for both languages, and that semantics is dominant in Tsova-Tush, which supports findings from existing literature. However, the results also suggest that making the coded semantic information more fine-grained could improve the accuracy of the gender predictions.

Our results confirm observations about the East Caucasian family, which relies heavily on humanness and abstractness as classifiers for gender assignment. Additionally, the first segment of a phonological form of a given noun being /b/ or /d/ is once again found to correlate highly with the corresponding genders B and D, respectively (see e.g. Nichols (1989; 2011: 147) for the concept of autogender, and its possible historical explanations).

As the many descriptive analyses demonstrate, gender assignment is language-specific, especially with respect to the specific semantic domains that emerge as important. Nonetheless, we have presented results here supporting the claim that certain common trends can be identified, notably the greater primacy of semantics as a categorisation principle, and within this, the importance of animacy. Whether this principle holds across all gendered languages (that is, whether it can be considered ‘universal’) requires further empirical testing, from both qualitative and quantitative perspectives.

Acknowledgements

The authors would like to thank two anonymous reviewers and Francesca Di Garbo for their comments on earlier versions of this paper. The contributions by JWS were in part thanks to funding of the Linked Open Dictionaries project, an Early Career Research Group funded by the German Ministry for Education and Research. KB acknowledges funding from a Postdoctoral Mandate at KU Leuven, and the Dutch Research Council (NWO) Veni grant number Vl.Veni.211C.077. MAT is thankful for the support from the French National Research Agency (ANR-20-CE27-0021). NR acknowledges funding from a doctoral contract granted by École Normale Supérieure - PSL, hosted at Sorbonne Nouvelle University (Paris).

Abbreviations

AOR = aorist	B = B gender	D = D gender
PRS = present		

References

Allassonnière-Tang, Marc & Dunstan Brown & Sebastian Fedden. 2021. Testing semantic dominance in Mian gender: Three machine learning models. *Oceanic Linguistics* 60(2). 302–334. https://doi.org/10.1353/ol.2021.0018

Balam, Osmer. 2016. Semantic categories and gender assignment in Contact Spanish: Type of code-switching and its relevance to linguistic outcomes. *Journal of Language Contact* 9(3). 405–435. https://doi.org/10.1163/19552629-00903001

Basirat, Ali & Marc Allassonnière-Tang & Aleksandrs Berdicevskis. 2021. An empirical study on the contribution of formal and semantic features to the grammatical gender of nouns. *Linguistics Vanguard* 7(1). 20200048. https://doi.org/10.1515/lingvan-2020-0048

Bellamy, Kate & M. Carmen Parafita Couto. 2022. Gender assignment in mixed noun phrases: State of the art. In Dalila Ayoun (ed.), *The acquisition of gender: Crosslinguistic perspectives*. 14–48. Amsterdam / Philadelphia: John Benjamins.

Bellamy, Kate & Jesse Wichers Schreur. 2022. When semantics and phonology collide: Gender assignment in mixed Tsova-Tush-Georgian nominal constructions. *The International Journal of Bilingualism* 26(3). 257–285. https://doi.org/10.1177/13670069211039559

Breiman, Leo & Jerome H. Friedman & Richard A. Olshen & Charles J. Stone. 1984. *Classification and regression trees.* Boca Raton: Routledge.

Brown, Dunstan. 1998. Defining 'sub-gender': Virile and devirilized nouns in Polish. *Lingua* 104(3-4). 187–233.

Carling, Gerd & Kate Bellamy & Jesse Wichers Schreur. 2021. *Gender stability in Nakh-Daghestanian*, paper presented at Languages, Dialects and Isoglosses of Anatolia, the Caucasus and Iran, March 2021, Paris.

Contini-Morava, Ellen & Marcin Kilarski. 2013. Functions of nominal classification. *Language Sciences* 40. 263–299. https://doi.org/10.1016/j.langsci.2013.03.002

Corbett, Greville. 1982. Gender in Russian: An account of gender specification and its relationship to declension. *Russian Linguistics* 6(2). 197–232.

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

Corbett, Greville. 2013. Systems of gender assignment. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/32, Accessed on 2022-02-02)

Corbett, Greville G. & Norman M. Fraser. 1993. Network Morphology: A DATR account of Russian nominal inflection. *Journal of Linguistics* 29(1). 113–142. https://doi.org/10.1017/S0022226700000074

Corbett, Greville & Norman M. Fraser. 2000. Default genders. In Barbara Unterbeck & Matti Rissanen (eds.), *Gender in grammar and cognition I: Approaches to gender.* 55–98. Berlin: Mouton de Gruyter.

Cruz, Abel. 2021. A syntactic approach to gender assignment in Spanish–English bilingual speech. *Glossa: a journal of general linguistics* 6(1). 1–40. https://doi.org/10.16995/glossa.5878

Desheriev, Y. D. [Дешериев]. 1953. *Bacbijskij jazyk: fonetika, morfologija, sintaksis, leksika* [The Tsova-Tush language: phonetics, morphology, syntax, lexicon]. Moscow: Izdatel'stvo AN SSSR.

Demsar, Janez & Blaz Zupan & Gregor Leban & Tomaž Curk. 2004. Orange: From experimental machine learning to interactive data mining, white paper. *European Conference of Machine Learning: 2004; Pisa, Italy* 3202. 537–539.

Dryer, Matthew S. 2013. Prefixing *vs*. suffixing in inflectional morphology. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/33, Accessed on 2022-02-02)

Evans, Nicholas & Dunstan Brown & Greville Corbett. 2002. The semantics of gender in Mayali: Partially parallel systems and formal implementation. *Language* 78(1). 109–153.

Evans, Roger & Gerald Gazdar. 1989a. Inference in DATR. *Proceedings of the fourth conference of the European Chapter of the Association for Computational Linguistics, Manchester, England*. 66–71.

Evans, Roger & Gerald Gazdar. 1989b. The semantics of DATR. In A. G. Cohn (ed.), *Proceedings of the seventh conference of the Society for the Study of Artificial Intelligence and Simulation of Behaviour,* 79–87. London: Pitman/Morgan Kaufmann.

Evans, Roger & Gerald Gazdar. 1996. DATR: A language for lexical knowledge representation. *Computational Linguistics* 22(2). 167–216.

Fedden, Sebastian. 2011. *A Grammar of Mian*. Berlin / Boston: Mouton de Gruyter.

Fraser, Norman M. & Greville G. Corbett. 1995. Gender, animacy, and declensional class assignment: A unified account for Russian. In Geert Booij & Jaap van Maarle (eds.), *Yearbook of Morphology 1994,* 123–150. Amsterdam: Kluwer Academic Publishers.

Fraser, Norman M. & Greville G. Corbett. 1997. Defaults in Arapesh. *Lingua* 103(1). 25–57.

Gagliardi, Annie & Jeffrey Lidz. 2014. Statistical insensitivity in the acquisition of Tsez noun classes. *Language* 90(1). 58–89. https://doi.org/10.1353/lan.2014.0013

Haykin, S. 1998. *Neural networks: A comprehensive foundation*. Prentice-Hall: Englewood Cliffs.

Her, One-Soon & Marc Tang. 2020. A statistical explanation of the distribution of sortal classifiers in languages of the world via computational classifiers. *Journal of Quantitative Linguistics* 27(2). 93–113.
https://doi.org/10.1080/09296174.2018.1523777

Hockett, Charles F. 1958. *A course in modern linguistics*. New York: MacMillan.

Jamalkhanov, Z. D. [Джамалханов] & Aliroev, I. Y. [Алироев]. 1991. *Slovar' pravopisanija literaturnogo čečenskogo jazyka* [Orthographical dictionary of literary Chechen]. Grozny: Kniga.

Kadagidze, E. [ქადაგიძე]. 2009. *C'ova-tušuri t'ekst'ebi* [Tsova-Tush texts]. Tbilisi: TSU gamomcemloba.

Kadagidze, D. & N. Kadagidze. [ქადაგიძე]. 1984. *C'ova-tušur-kartul-rusul leksik'oni* [Tsova-Tush-Georgian-Russian dictionary]. Tbilisi: Mecniereba.

Karmiloff-Smith, Annette. 1979. *A functional approach to child language*. New York / London: Cambridge University Press.

Khalilov, M. S. [Халилов]. 1999. *Cezsko-russkij slovar'* [Tsez-Russian dictionary]. Moscow: Academia

Lemus-Serrano, Magdalena & Marc Allassonnière-Tang & Dan Dediu. 2021. What conditions tone paradigms in Yukuna: Phonological and machine learning approaches. *Glossa: a journal of general linguistics* 6(1). 60. https://doi.org/10.5334/gjgl.1276

List, Johann-Mattis & Michael Cysouw & Robert Forkel. 2016. *Concepticon: A resource for the linking of concept lists*. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2393–2400.

Matsiev, A. G. [Мациев]. 1961. *Slovar' čečenskogo jazyka* [Chechen dictionary]. Moscow: Gosudarstvennoe izdatel'stvo inostrannyx i nacional'nyx sloverej.

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: Chicago Linguistic Society, University of Chicago.

Nichols, Johanna. 1994. Chechen. In Riks Smeets (ed.) *The North East Caucasian languages, part 2,* 1–78. Delmar: Caravan.

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.

Nichols, Johanna. 2007. Chechen morphology with notes on Ingush. In Alan S. Kaye (ed.), *Morphologies of Africa and Asia,* 1188–1207. State College: Penn State University Press. https://doi.org/10.1515/9781575065663-044

Nichols, Johanna. 2011. *Ingush grammar.* Berkeley, Los Angeles: University of California Press.

Parks, Randolph & Daniel S. Levine & Debra L. Long. (eds.). 1998. *Fundamentals of neural network modeling: Neuropsychology and cognitive neuroscience.* Boston: MIT Press.

Plaster, Keith & Maria Polinsky & Boris Harizanov. 2013. Noun classes grow on trees: Noun classification in the North-East Caucasus. In Balthazar Bickel, Lore A. Grenoble, David A. Peterson & Alan Timberlake (eds.), *Language typology and historical contingency: In honor of Johanna Nichols,* 153–170. Amsterdam: John Benjamins.

Polinsky, Maria & Ezra Van Everbroeck. 2003. Development of gender classifications: Modeling the historical change from Latin to French. *Language* 79(2). 356–390.

Quinlan, J. Ross. 1993. *C4.5: Programs for Machine Learning.* Burlington: Morgan Kaufmann Publishers.

Quinlan, J. Ross. 1996. Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research* 4. 77–90.

Rajabov, Ramazan. Undated. *Tsez Dictionary*. Unpublished MS (Los Angeles: University of Southern California).

Senft, Gunter (ed.). 2000. *Systems of nominal classification.* Cambridge: Cambridge University Press.

Sokolik, M. E. & Michael E. Smith. 1992. Assignment of gender to French nouns in primary and secondary language: A connectionist model. *Second Language Research* 8(1). 39–58.

Tagliamonte, Sali A. & R. Harald Baayen. 2012. Models, forests and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24(2). 135–178. https://doi.org/10.1017/S0954394512000129

Ting, Kai Ming. 2010. Precision and recall. In Claude Sammut & Geoffrey I. Webb (eds.), *Encyclopedia of Machine Learning,* 781–781. Boston: Springer. https://doi.org/10.1007/978-0-387-30164-8_652

Ulrich, Natalja & Marc Allassonnière-Tang & François Pellegrino & Dan Dediu. 2021. Identifying the Russian voiceless non-palatalized fricatives /f/, /s/ and /ʃ/ from acoustic cues using Machine Learning. *Journal of the Acoustical Society of America* 150(3). 1806–1820. https://doi.org/10.1121/10.0005950

Wichers Schreur, Jesse. 2021. Nominal borrowings in Tsova-Tush (Nakh-Daghestanian, Georgia) and their gender assignment. In Diana Forker & Lore A. Grenoble (eds.), *Language contact in the territory of the former Soviet Union,* 15–33. Amsterdam: John Benjamins.

Wurm, S. A. & I. Heyward & Unesco. 2001. *Atlas of the world's languages in danger of disappearing*. Paris: Unesco Pub. Website http://www.unesco.org/languages-atlas/ consulted on 7-12-2021.

CONTACT
jesse.jessews@gmail.com
marc.allassonniere-tang@mnhn.fr
k.r.bellamy@hum.leidenuniv.nl
neige.rochant@sorbonne-nouvelle.fr