UNIVERSITY of York

This is a repository copy of *Testing semantic dominance in Mian gender : Three machine learning models*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/171940/

Version: Accepted Version

Article:
Allassonnière-Tang, Marc, Brown, Dunstan orcid.org/0000-0002-8428-7592 and Fedden, Sebastian (Accepted: 2021) Testing semantic dominance in Mian gender : Three machine learning models. Oceanic Linguistics. (In Press)

Reuse
["licenses_typename_other" not defined]

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Testing semantic dominance in Mian gender:
Three machine learning models

Author1, Author2, Author3
[1]Affiliation 1
[2]Affiliation 2
[3]Affiliation 3

Abstract
The Trans New Guinea language Mian has a four-valued gender system that has been analyzed in detail as semantic. This means that the principles of gender assignment are based on the meaning of the noun. Languages with purely semantic systems are at one end of a spectrum of possible assignment types, while others are assumed to have both semantic and formal (i.e. phonology- or morphology-based) assignment. Given the possibility of gender assignment by both semantic and formal principles, it is worthwhile testing the empirical validity of the categorization of the Mian system as predominantly semantic. Here we apply three machine-learning models to determine independently what role semantics and phonology play in predicting Mian gender. Information about the formal and semantic features of nouns is extracted automatically from a dictionary. Different types of computational classifiers are trained to predict the grammatical gender of nouns, and the performance of the computational classifiers is used to assess the relevance of form and semantics in relation to gender prediction. The results show that semantics is dominant in predicting the gender of nouns in Mian. While it validates the original analysis of the Mian system, it also provides confirmation that form-based and semantic features do not contribute equally in all languages with gender. More generally our work also demonstrates the value of computational methods to validate analyses of gender systems.

1. Introduction

Grammatical gender such as the masculine/feminine distinction in French and Spanish is a system that classifies nouns according to patterns of agreement. Grammatical gender is found in about half of the world's natural languages (Corbett 2013a). According to Contini-Morava and Kilarski (2013), one of the main functions of grammatical gender is referent identification and tracking. To be precise, the grammatical gender markers can provide informative cues in discourse to track the referents involved in the discussion. However, the principles that govern the assignment of nouns to genders differ significantly across languages (Comrie 1999; Kemmerer 2014; 2017). For instance, in French, *voiture* 'car' is feminine and *vélo* 'bicycle' is masculine.

The Trans New Guinea language Mian has a four-valued gender system (masculine, feminine and two neuters) that has been analyzed in detail as semantic (Fedden 2011), that is, gender assignment is based on the meaning of a noun. Languages with purely semantic systems are at one end of a spectrum of possible assignment types, while others are assumed to have both semantic and formal (i.e. phonology- or morphology-based) assignment. Given the possibility of gender assignment by both semantic and formal principles, we want to test the empirical validity of the categorization of the Mian system as predominantly semantic. To this end we apply three machine-learning models to determine independently what role semantics and phonology play in predicting Mian gender. We also consider two opposing hypotheses about gender assignment, one that suggests that semantic features dominate formed-based ones (the 'semantic dominance hypothesis'), and one that suggests that form-based and semantic features contribute equally (the 'equality hypothesis'). The first hypothesis ('semantic dominance hypothesis') by Corbett and Fraser (2000: 321) proposes that semantic gender assignment principles have priority over formal gender assignment principles. For example, in

Spanish, most nouns ending in /a/ are feminine. However, nouns referring to males are masculine even if they end in /a/, e.g. *el guardia* 'the policeman' is masculine despite the a-ending. The second hypothesis ('equality hypothesis'), articulated by Rice (2006), argues that forms and semantics contribute equally to gender assignment.

Previous research provided qualitative and small-scale quantitative evidence for both hypotheses, cf., Corbett and Fraser (2000) and Nesset (2006) for the semantic-dominance hypothesis and Corteen (2019) for the equality hypothesis. In terms of data size, the largest samples found in previous studies come from Corteen (2019), who used a sample of 592 nouns extracted from the DUDEN German dictionary and from Corbett and Fraser (2000), who analyzed a sample of 1,500 nouns extracted from Zasorina's (1977) Russian dictionary. Nevertheless, there is a dearth of large-scale quantitative evidence on the contributions of form and semantics to gender assignment. Moreover, no metrics have actually been used to quantify and rank the contribution of different formal and semantic features.

The current study proposes the use of computational classifiers[1] as a systematized method to evaluate the contribution of formal and semantic features to gender assignment in Mian. Mian has been chosen as it is analysed in Fedden (2011) as a language in which semantics dominates gender assignment. Given that they are not so dependent on the judgment of the researcher, machine learning methods are a good means of avoiding confirmation bias. Information about the formal and semantic features of nouns is extracted automatically from a dictionary. Then, computational classifiers are trained with the extracted information. The performance of the computational classifiers trained on the respective information source serves as the measure of the contribution of each information source to gender assignment. The importance of each variable is also extracted to measure the relevance of forms and semantics in predicting the gender of Mian nouns. The results show that semantics is dominant in predicting the gender of nouns in Mian. While it is in line with the semantic dominance hypothesis, independent validation of the original analysis of Mian system provides evidence that the equality hypothesis cannot be true for all languages with gender.

2 Background to the study

In this section, we ground the study in the current debate about factors which impact gender assignment. A brief definition of grammatical gender is provided. The two gender assignment hypotheses are also explained.

2.1 Defining gender

The experience and objects encountered by humans are stored in the mind. To ease the process of information storage and retrieval in the brain, these entries need to be categorized (Lakoff and Johnson 2003, 162–163). This need is reflected in language via mechanisms of nominal classification, i.e., classification of nouns. One of the most common of these nominal classification systems is grammatical gender, which is found in approximately half the world's languages (Corbett 2013b). As an example, nouns in French are assigned to either masculine (e.g., *verre* 'glass') or feminine (e.g., *bouteille* 'bottle'), which triggers grammatical agreement on various associated words, the agreement targets, e.g. *un grand verre* [one.M big.M glass(M)] vs. *une grande bouteille* [one.F big.F bottle(F)].

Grammatical gender, as defined in this study, is a feature of languages that distinguish between (sub-)classes of words (typically nouns) through some morphological means (Seifart 2010). The primary criteria for distinguishing a grammatical gender is the feature of 'agreement' (Hockett 1958; Corbett 1991). If a noun triggers agreement on at least one other word class, the language has grammatical gender. The two clauses in (1) display the same number, case and syntactic structure, yet the different grammatical genders (masculine/feminine) of the nouns

[1]In order to distinguish 'classifier' in the sense used in machine learning from the term 'classifier' as used for a type of nominal classification system, we employ the term 'computational classifier' for the former.

are reflected on the agreement targets, here the numeral, adjective and verb. Features not directly relevant to the analysis (e.g., number) are ignored in the example.

(1) Gender agreement in French (Indo-European)

a.	[œ̃	gʁɑ̃	livʁ	ɛ	ekʁi]
	one.M	big[M]	book(M)	is	write.past[M]
	'A great book has been written'				
b.	[yn	gʁɑ̃-d	lɛtʁ	ɛ	ekʁi-t]
	one.F	big-F	letter(F)	is	write.past-F
	'A long letter has been written'				

There is no grammatical gender without agreement. Mandarin Chinese (Sinitic), for example, has sex-differentiable nouns, like *ge1ge1* 'older brother' and *jie3jie0* 'older sister', but the absence of agreement in the language clearly shows that there is no gender system. Thus, each gender system consists of two essential parts (Fedden & Corbett 2018:636): the distinctions made in the system and the outcome in form. The speaker must select a value within the classification system (in many systems this selection may be completely determined), and this selection must be realized in linguistic form. Assignment systems are models of this selection. In French, *homme* 'man' is masculine and *femme* 'woman' is feminine. These values are realized on the different agreement targets, i.e., through the system of exponence of gender. Assignment and exponence work together. Exponence is the morphological means by which gender is expressed: the evidence for gender is the systems of exponence.[2]

The underlying motivation for different types of gender assignment is still under investigation (and debate) within the field of linguistics. However, important generalisations can be made. According to Corbett (1991), with semantic and formal assignment, there are three logically possible gender assignment systems (i.e. exclusively semantic, exclusively formal, and a combination of formal and semantic). Only two of these combinations can be found, namely exclusively semantic and a combination of formal and semantic. This typological gap suggests that semantics takes precedence in gender assignment.

Assignment systems can be fully semantic, that is for all (or almost all) nouns we can predict the gender from the meaning of the noun. A classic example is the Nakh-Daghestanian language Bagvalal (Kibrik 2001), spoken in southwestern Daghestan by approximately 1,500 speakers. Bagvalal has three gender values, as evidenced by the agreement. Nouns denoting male humans are masculine; nouns denoting female humans are feminine; all remaining nouns are neuter. The neuter gender comprises all non-humans (whether animate or inanimate):

(1) Bagvalal (Kibrik 2001: 64-5)

waša	w-iRi
boy(M)	M.SG-stop
'the boy stopped'	

(2)

jaš	j-iRi
girl(F)	F.SG-stop
'the girl stopped'	

(3)

ʕama	b-iRi	
donkey(N)		N.SG-stop
'the donkey stopped'		

[2]Fedden & Corbett (2018:636) make this claim about nominal classification systems in general, but since gender is a type of nominal classification system it applies to gender in particular.

Similar systems can be found in Dravidian languages, for example Tamil (Asher 1985: 36-37; Corbett 1991: 8-9). Semantic assignment is typically along the lines of sex (masculine vs. feminine) or animacy (animate vs. inanimate, or human vs. non-human). All gender languages have such semantic core assignment principles. In some languages, like Bagvalal, they are sufficient, in other languages additional principles are required to assign a gender to the remaining nouns. These principles can be semantic as well, as for example in Bininj Gun-Wok (Evans, Brown & Corbett 2002: 125), a language of the Gunwinyguan family spoken in central Arnhem Land in Australia's Northern Territory, which has masculine, feminine, vegetable and neuter gender values. The usual semantic assignment principles are based on sex, but there are additional semantic principles, for example vegetable gender is assigned to nouns denoting plants and their products or foods and vegetables.

Often languages make use of additional formal assignment principles. These can be either morphological or phonological. Morphological principles are about inflection class or derivation (for complex words). Phonological principles are about the phonological shape of the noun stem. In Russian inflection class is an important morphological predictor. Once standard semantic principles that make nouns denoting males masculine and nouns denoting females feminine have applied, for the residue, one can predict the gender of a noun from the way it inflects, e.g. *zakon* 'law' belongs to inflection class I and is masculine, and *kniga* 'book' belongs to inflection class II and is feminine. Semantic principles take precedence, so a noun like *djadja* 'uncle', which belongs to inflection class II and should therefore be feminine, is in fact masculine by virtue of its meaning. For a more detailed, computationally implemented account of gender assignment in Russian, see Fraser & Corbett (1995).

In the East Cushitic language Qafar (Corbett 1991: 51-2) we find additional phonological principles. Generally, standard semantic principles apply, i.e. males are masculine and females are feminine. For the residue, one can predict the gender of a noun from the phonology of the stem. Nouns whose citation form ends in an accented vowel are feminine, e.g. *karmà* 'autumn', the rest is masculine, e.g. *gilàl* 'hiver', *tàmu* 'taste'. However, semantic principles take precedence, so a noun like *abbà* 'father', which ends in an accented vowel and should therefore be feminine, is in fact masculine. Again, this is a clear example where the semantic assignment principle takes precedence.

The typology therefore provides us with two basic systems. For combined systems the challenge is to model how the formal and the semantic principles interact. For purely semantic systems there is the risk that form-based generalisations might be overlooked. This is where the application of machine learning methods over a sufficiently large sample of the lexicon can help us confirm, or potentially reject, analyses. The accepted analysis of the Mian system is one where semantics dominate (Fedden 2011) and we should check how this stands up to scrutiny. Before we do this, as a point of orientation, we consider two hypotheses about gender assignment before going on to look at the Mian system and our modelling of it.

2.2 Two hypotheses on gender assignment

Previous studies suggest that the contribution of semantic features generally outranks the contribution of formal ones. Corbett (1991: 68) proposes that ''[i]f there are conflicting factors at work, semantic factors usually take precedence''. Corbett and Fraser (2000: 321) suggest that "[a]s is universally the case, the formal gender assignment rules [. . .] are dominated by the semantic gender assignment rules". As an example, in Arapesh, the noun *nakor* 'husband's father' ends in /r/. Based on the word form, it is expected to belong to gender X. However, nouns referring to male beings belong to gender VII in Arapesh. In this conflict of gender assignment rules, the semantic-based rule prevails in Arapesh and *nakor* is assigned gender VII (Dobrin 1999; Corbett and Fraser 2000; Nesset 2006: 1385). In contrast, a recent approach that uses the framework of Optimality Theory (Prince and Smolensky 1993) to evaluate the hierarchical interaction of formal and semantic features in gender assignment (Rice 2006), looking at German, Russian, French, Norwegian, and Dutch, suggests that

information about form and semantics (defined as constraints in the framework of the optimal gender assignment hypothesis) operate together as a block that cannot be ranked, which is equivalent to saying that their contributions are equal. Instead of ruling out potential candidates for gender assignment by running through the constraints (i.e., features defined over semantics and word forms) one at a time, all the constraints about semantic and formal features are considered together. Each candidate is marked with the number of constraints violated. The candidate with the fewest violations is then selected as the optimal candidate. In other words, formal and semantic features are hypothesized to have a complementary effect on gender assignment. Given the languages analyzed the coverage of the hypothesis in relation to all languages with gender is unclear, although Rice (2006: 1395) appears to argue for its general applicability.

Corbett's (1991) work on gender, in which the hypothesis proposing semantic dominance is presented, looked at a wide sample of languages, including beyond Indo-European, whereas Rice (2006), while mentioning isolated examples outside of the family, is restricted to Indo-European in its analytical focus. Corbett's (1991) original analysis of the Russian gender system, which assumes that semantics take precedence, is refined in Corbett and Fraser (2000), for which an implementation covering 1,500 high frequency nouns exist. In general, however, support for these two different views relies mostly on qualitative analyses. This means that only a small sample of nouns was selected in different languages and the process of their gender assignment was simulated, as for example in Rice (2006). As we argue in this study, progress in understanding gender systems can only be made by using computational experiments to assess how well formal and semantic features can predict the grammatical gender of nouns, with it being particularly important that we move beyond the Indo-European language family. The Mian language from the Trans-New-Guinea language family, analyzed as having a semantic system, provides a good basis to determine how well we can test such a characterization. We therefore present the standard analysis of the Mian system before discussing the experimental work.

3. An overview of Mian

Mian belongs to the Ok family of languages (Healey 1964). The Ok family is part of the larger Trans New Guinea (TNG) family (Wurm 1982; Ross 2005; Pawley 2005). The eastern dialect of Mian, described in Fedden (2011) is spoken by approximately 1,400 people in the Telefomin District of Sandaun Province in Papua New Guinea. All Mian data presented in this paper are extracted from Fedden's fieldwork data.

Mian is a word tone language. The domain in which five lexically specified tonal melodies contrast is the entire phonological word (Donohue 1997). In the orthography, the five tonal melodies are written as follows: *mēn* 'child' (H), *mén* 'string bag' (LH), *klâ* 'properly' (LHL), *fè* 'carrion' (HL). Low tone is unmarked, e.g. *am* 'house' (L). Mian is head-marking (Nichols 1996). The unmarked constituent order is SOV. The language is strongly zero-anaphoric, i.e., all argument noun phrases are typically elided if referent identity is retrievable from context or world knowledge. Serial verb constructions and clause chaining are very frequent construction types. Arguments are marked by means of verbal affixes (mainly) following a nominative-accusative pattern. The subject is obligatorily indexed by a suffix in all finite verb forms, the object is indexed by a prefix in finite and non-finite verb forms for seven verbs only: *-têm'* 'see (PFV)', *-temê'* 'see (IPFV)', *-lò* 'hit, kill (PFV)', *-nâ'* 'hit, kill (PFV)', *-e* 'hit, kill (IPFV)', *-ntamâ'* 'bite (PFV)', and *-fû'* 'grab (PFV)'.

Trans New Guinea languages typically do not have gender systems. If they do they are usually restricted to a masculine-feminine distinction in the third person singular pronouns (Wurm 1982, 80), as for example in Oksapmin (Loughnane 2009). Mian has – like the closely related Ok languages Telefol and Tifal – gender that is not restricted to pronouns. The Mian gender system is interesting as the answer to the question how many genders the language has is not straightforward, because of a mismatch between the number of genders into which the

nouns are divided on the basis of distinct agreement forms (controller genders) and the number of genders which are marked on the agreement targets (target genders).

3.1 Forms

Mian does not mark gender overtly on the noun. Agreement targets are the enclitic articles (e.g., *naka=e* [man(M)=SG.M] 'the man') and other determiners within the noun phrase, e.g., adnominal demonstratives (e.g. *naka ēle* [man(M) DEM.SG.M] 'this man'). Outside the noun phrase, only verbs agree. Verbal affixes show gender, person and number in a portmanteau fashion. The account of the Mian gender system is based on Fedden (2011) and Corbett, Fedden and Finkel (2017). Mian gender agreement is illustrated in examples (3a) and (3b):

(3) Examples of gender agreement in Mian

a.	*ō*	*naka=e*	*a-têm'-Ø-o=be*
	3SG.F	man(M)=SG.M	3SG.M.OBJ-see.PFV-REAL-3SG.F.SBJ=DECL
	'She saw the man.'		
b.	*ē*	*unáng=o*	*wa-têm'-Ø-e=be*
	3SG.M	woman=SG.F	3SG.F.OBJ-see.PFV-REAL-3SG.M.SBJ=DECL
	'He saw the woman.'		

We restrict our description of the agreement patterns to the article since all agreement targets show exactly the same agreement pattern. The forms of all other agreement targets can be found in the appendix. Referentially used nouns are followed by an enclitic article.[3] The forms are given in (4).

(4) Examples of gender agreement on the articles in Mian

a.	*naka=e*	'a/the man'	*naka=i*	'(the) men'
b.	*unáng=o*	'a/the woman'	*unáng=i*	'(the) women'
c.	*tóm=e*	'a/the stone'	*tóm=o*	'(the) stones'
d.	*am=o*	'a/the house'	*am=o*	'(the) houses'

On the basis of distinct agreements, four controller genders can be identified in Mian. For many languages, controller genders have to be distinguished from target genders, which are the number of genders marked on the agreement targets (Corbett 1991, 151). Mian is such a language. The controller genders in Mian are given in (5):

(5) The agreement markers in Mian

a. Masculine (*=e*, *=i*), e.g. *naka* 'man'
b. Feminine (*=o*, *=i*), e.g. *unáng* 'woman'
c. Neuter 1 (*=e*, *=o*), e.g. *tóm* 'stone'
d. Neuter 2 (*=o, =o*), e.g. *am* 'house'

The paradigm of the article shows striking patterns of syncretism with the result that Mian genders have no agreement forms that are unique to them and are therefore 'non-autonomous' values (Zaliznjak 1973, 69–74; Baerman, Brown, and Corbett 2005, 15; Corbett 2012, 156). Table 1 sets the syncretism patterns using letters and different colours to indicate syncretic cells.

[3]These are articles rather than overt markers of number and gender, which a noun either invariably has or lacks. Articles are left out when a noun is used non-referentially, for example, in first elements in noun-noun compounds, e.g., *míl-blong* [bean-pod] 'bean pod', or under negation, e.g., *imen blim* [taro not_exist] 'there's no taro', *yāi=ba=be* [wound=NEG=DECL] 'it's not a wound'.

Table 1. Syncretisms in the Mian gender system (based on Corbett, Fedden and Finkel (2017: 10))

	Singular	Plural
Masculine	A	C
Feminine	B	C
Neuter 1	A	B
Neuter 2	B	B

The neuter genders are not exceptions that could be specified in the lexicon. They are not inquorate, i.e., genders with just a few nouns in them (Corbett 1991, 170). Neuter 1 contains many hundred nouns and neuter 2 contains many dozens nouns, and inanimate loan words are readily assigned to one of the neuter genders. The relation between controller genders and target genders is illustrated in Figure 2 using the agreement forms of the article.

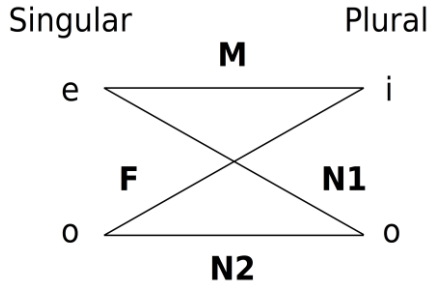

Fig2. Controller and target genders in Mian (Corbett and Fedden 2016: 517)

Mian has four controller genders independent of number, but there are two target genders in both the singular and the plural. Having established the difference between controller and target genders in Mian we now turn to the semantics of gender assignment.

Mian also has a second nominal classification system with six distinctions (e.g. long object, covering object, bundle), called verbal classifiers in Fedden (2011). The system is only marked on about 40 verbs with object handling and movement semantics, e.g. 'give', 'take', 'lift', 'turn', 'fall'. For a detailed description and analysis of this system, see Fedden (2011, ch. 5) and Corbett, Fedden and Finkel (2017).

3.2 Gender assignment

Gender assignment in Mian has been analyzed as predominantly semantic. The semantic criteria on which the assignment of the noun vocabulary is based are given in Table 2.

Table 2. Semantic criteria for gender assignment (Fedden 2011: 172)

<table>
<tr><th colspan="3">Assignment criteria</th><th>Gender</th></tr>
<tr><td rowspan="2">Animate</td><td rowspan="2">Human, animal (where sex readily discernible or relevant)</td><td rowspan="2">Sex</td><td>Masculine, e.g. naka 'man'</td></tr>
<tr><td>Feminine, e.g. unáng 'woman'</td></tr>
<tr><td rowspan="6">Inanimate</td><td colspan="2">Count nouns, e.g. mén 'string bag', imen 'taro'</td><td rowspan="2">Neuter 1</td></tr>
<tr><td colspan="2">Liquids, body fluids/wastes, substances, e.g. aai 'water', ilem 'blood', fút 'tobacco'</td></tr>
<tr><td colspan="2">Places, e.g. am 'house', dafáb 'summit'</td><td rowspan="4">Neuter 2</td></tr>
<tr><td colspan="2">Masses, e.g. afobèing 'goods, property'</td></tr>
<tr><td colspan="2">Body decoration, e.g. eit 'decoration', baasi 'pig's tusk'</td></tr>
<tr><td colspan="2">Weather phenomena, e.g. sók 'rain', ayung 'mist'</td></tr>
</table>

	Illnesses, e.g. *kweim* 'fever'	
	Intangibles/abstracts, e.g. *āns* 'song'	
	Verbal nouns, e.g. *fumin* 'activity of cooking (IPFV VN)'	
	Some tools and weapons, e.g. *káawa* 'steel axe'	

Nouns with animate referents are either masculine or feminine. The gender contrast is neutralized in the plural. The masculine and feminine genders contain only humans and animals. Nouns which refer to humans have masculine or feminine gender on the basis of the biological sex of their referents. 'Common gender' is a label for items which are arguably single lexical items, yet which have more than one gender value (Corbett 1991: 67). For example, Mian *aban* 'orphan' (Fedden 2011: 170), is masculine or feminine, a contrast that shows up only in the agreements. The same is true of some animals, e.g. *eíl* 'pig' (where the sex matters to humans) or *iwau* 'duck' (where the sexes look distinctly different).

Nouns of masculine gender referring to humans consist of male proper names, male kin relations or social relations, categories of males and jobs. Examples are: *aab* 'brother', *aaleb* 'father', *aaling* 'father's younger brother', *afín* 'friend, ally', *āi* 'dad', *awokîm* 'father's sister's husband', *ayàab* 'father's older brother', *ayàal* 'paternal grandfather', *baliam* 'male ancestor', *fanin* 'male ancestor', *hangkalebmín* '(very) old man', *hek (sūm)* 'oldest brother', *imak* 'husband', *kiab* 'kiap, patrol officer', *kimaanîn* 'boss, minder', *kingkan* 'shaman', *komók* 'leader', *máamein* 'mother's younger brother', *makáa* 'enemy', *mín* 'son', *molim* 'father-in-law', *naka* 'man', *nek* 'friend', *ning* 'younger brother', *nokai* 'maternal grandfather', *tembal* 'young man, bachelor'.

Nouns of feminine gender referring to humans consist of female proper names, female kin relations or social relations, and categories of females. Examples are: *âam* 'older sister', *afók* 'grandmother, female ancestor', *akuláb* 'parent's older sister', *alél* 'wife', *andlok* 'mother-in-law', *awók* 'mother', *báab* 'parent's younger sister', *biém* 'mum', *en (sūm)* 'oldest sister', *konokmón* '(very) old woman', *món* 'daughter', *neng* 'younger sister', *sou* 'young, unmarried woman', *unáng* 'woman'.

For many, particularly lower animals, gender is conventionalized as either masculine or feminine. Conventionalized gender is typically found with animals for which biological sex is not apparent. As an example, birds that lack sexual dimorphism, or animals for which biological sex is simply irrelevant, e.g., *tebél* 'ant' is masculine, while *slub* 'cockroach' is feminine. Table 3 gives some examples of conventionalized gender. The respective generic terms are not obligatory and appear in brackets.

Table 3. Examples of conventionalized gender for animals (Fedden 2011: 173)

	Masculine	Feminine
Animals capable of flight, i.e. birds and bats	*(wan) taimâ* 'heron' *(wan) tolim* 'eagle' *(wan) katab* 'flying fox'	*(wan) gwingwî* 'emerald dove' *(wan) alifayum* 'kingfisher'
Cassowary	n/a	*koból* 'cassowary'
Rodents, marsupials, monotreme	*(no) snuk* 'rat' *(no) kwiam* 'tree kangaroo'	*(no) befakam* 'sugar glider' *(no) yakéil* 'long-beaked echidna'
Reptiles	*(tím) ali* 'python' *(tím) heye* 'lizard sp.'	*(tím) biman* 'snake sp.'
Fish	all long fish, e.g. *(aning) finí* 'eel'	all short fish, e.g. *(aning) guguma* 'fish sp.'
Spiders	n/a (all spiders are feminine)	*(gwán) hōndou* 'spider sp.'
Insects, etc.	*tebél* 'ant' *fobiâ* 'leech'	*slub* 'cockroach' *kweng* 'grasshopper'

Neuter 1 contains mainly inanimate count nouns. The singular form denotes exactly one real world entity, the plural form denotes more than one, e.g., *tóm=e* 'a/the stone' vs. *tóm=o* '(the) stones'. Such nouns can be counted by means of a numeral, e.g. *tóm=o asumâtna* [stone(N1)=PL.N1 three] 'three stones'.

Neuter 1 also contains liquids, body fluids/wastes and substances, e.g. *aai* 'water', *al* 'faeces' and *fút* 'tobacco'. For these, the number contrast encodes a difference between small and large quantities, e.g. *aai=e* 'some water' vs. *aai=o* 'much water'. While such nouns have a distinct singular and plural, they cannot be counted by means of a numeral, e.g. **ilem asú* [blood two].

Semantically neuter 1 nouns can be subdivided into body parts, liquids and substances, other natural entities, and cultural artifacts.

- Examples of body parts are: *aal* 'skin' (and the compound *sitaal* 'lip' [lit. tooth-skin]), *abín* 'navel', *anang* 'mouth', *ban* 'arm', *bān* 'palm, sole', *báan* 'jaw', *bēl* 'wing', *dáang* 'back, spine', *debelón* 'forehead', *dlong* 'knee', *éit* 'penis', *fiaam* 'tail fin', *gabáam* 'head', *háang* 'tongue', *ikam* 'leg', *ín* 'liver', *kin* 'eye', *klōn* 'ear', *kwéil* 'hand', *kwel* 'neck', *kwīng* 'shoulder', *mokók* 'ankle', *mukùng* 'nose', *mutum* 'heel', *nái* 'vagina', *ōn* 'bone', *sít* 'tooth', *skíl* 'foot', *tub* 'breast'.
- Examples of liquids and substances are: *aai* 'water', *al* 'faeces', *as* 'wood', *atol* 'flame', *déib* 'moss', *dēn* 'tree sap', *fút* 'tobacco', *gáam* 'juice, grease', *gabangnak* 'male or female genital fluids, cassowary faeces', *ibal* 'dust', *ifá* 'sweat', *ilem* 'blood', *imán* 'urine', *isá* 'pus', *māt* 'bile'.
- Examples of natural entities (excluding body parts and liquids and substances) are: *áam* 'pandanus (P. antaresensis; Tok Pisin *karuka*)', *aket* 'flower', *amún* 'lake', *as* 'tree', *deit* 'nest', *dingding* 'taro rhizome', *éim* 'pandanus (P. conoideus; Tok Pisin *marita*)', *imen* 'taro', *kimit* 'cucumber', *kimkim* 'root', *mifím* 'sago palm', *níng* 'thorn', *som* 'banana', *tek* 'vine', *un* 'egg', *wán* 'sweet potato'.
- Examples of cultural artifacts are: *afong* 'walking stick', *aful* 'ball', *ān* 'arrow', *anòk* 'bow', *atit* 'wooden stick used for eating', *ayal* 'light(source)', *báangkli* 'stone adze', *fabí* 'stone adze', *geim* 'pronged arrow', *mén* 'string bag', *tlúm* 'brace, bridge', *was* 'drum', *yóum* 'piece of clothing'.

For the neuter 2 gender the agreement forms do not allow the encoding of a number contrast. Neuter 2 contains nouns of the following semantic subclasses: masses; places, locations, and types of terrain; traditional body decoration; weather phenomena; abstract notions and intangibles which include the subclasses of illnesses and verbal nouns.

- Examples of masses are: *afobèing* 'goods', *atum* 'smoke', *awitnîn* 'star(s)', *difib* 'rubbish' (e.g. torn paper, small bits of wood), *dím* 'flesh', *fub* 'rubbish bits', *kibi* 'face' (consisting of eyes, nose, mouth, etc.), *kutab* 'white ash', *unín* 'food'.
- Examples of places, locations, and types of terrain (especially places with certain functions, for example the abode of humans or animals) are: *am* 'house' (and all its compounds, such as *gilam* 'house without kitchen' [lit. 'cold-house'], *itam* 'dance house', *kwoisâm* 'spirit house', *katabam* 'cave' [lit. 'flying_fox-house']), *basal* 'veranda', *betan* 'area, place', *bib* 'village, place', *damìb* 'garden', *dāng* 'garden', *deib* 'path', *mon* 'old garden', *sesá* 'bush', *smē* 'cave'.
- Examples of traditional body decoration are: *amún* 'hole in nosetip', *baasi* 'pig tusk (put through the septum)', *eit* 'decoration', *mitakla* 'hole through septum (receiving the *baasi* tusk)', *klōn maalu* 'pig tusk (put through the ear)'.
- Examples of weather phenomena are: *ayung* 'mist', *ēimawe* 'haze', *íb* 'cloud(s)', *sók* 'rain'.
- Examples of intangibles and abstract notions are: *am* 'day', *angkusil* 'war magic', *āns* 'song', *awá* 'fight', *dam* 'dream', *fotom* 'shame', *hōb* 'breath, spirit', *kél* 'black magic' (rough

equivalent of Tok Pisin *poisin*), *kukúb* 'way, custom', *ninín* 'name', *titil* 'strength, power', *wasi* 'war(fare)', *wéng* 'talk, language, voice' [and all its compounds, such as *glolwêng* 'rumour' (lit. 'wind-talk'), *kwelwêng* 'whisper' (lit. throat-talk)], *taan* 'sunlight', *tang* 'smell', *tēing* 'generosity', *usem* 'sorcery' (rough equivalent of Tok Pisin *sanguma*); furthermore all illnesses, e.g. *klō* 'ringworm (a fungal skin infection)', *genin* 'illness (general state of being unwell)', *kweim* 'fever'; and all verbal nouns, e.g., *fumin* 'activity of cooking'.

The neuter 2 gender contains a few nouns that refer to discrete and countable entities, e.g., *am* 'house' and some tools and weapons, such as *káawa* 'steel axe', *mók* 'stone adze', *skemdâng* 'knife'. The agreements show no number contrast, e.g., *káawa=o* 'a/the steel axe, (the) axes'. Example (6a) has two readings, depending on the context; example (6b) illustrates how neuter 2 nouns referring to a discrete real-world entities can be counted by means of a lexical numeral.

(6) Examples of neuter 2 gender in Mian

a. *am=o* *yē* *bi-Ø-o=be*
house(N2)=N2 there exist.IPFV-IPFV-N2.SBJ=DECL
'There is a house.' OR 'There are houses.'

b. *am=o* *asú* *yē* *bi-Ø-o=be*
house(N2)=N2 two there exist.IPFV-IPFV-N2.SBJ=DECL
'There are two steel axes.'

Loan words from Tok Pisin, most of which come ultimately from English, are assigned to the four genders largely on the basis of the semantic properties of their referents. Animates are assigned on the basis of the sex of the referent, e.g., masculine: *kiab* 'kiap, patrol officer', *kounsol* 'councillor', *bolis* 'policeman', *bailot* 'pilot', *emeief* 'someone who works for the Mission Aviation Fellowship (M.A.F.)', *soldia* 'soldier'. All of these have exclusively male referents. Inanimates referring to discrete countable objects are usually assigned to neuter 1, as one would expect, e.g., *senso* 'chainsaw', *hàs* 'hat', *balu* 'plane' (< Tok Pisin *balus* 'pigeon, aeroplane'), *tòs* 'torch', *siòt* 'shirt', *sù* 'shoe', *ben* 'pen', *bòks* 'box', *kàb* 'cup'. Nouns referring to locations and institutions are neuter 2, e.g., *klabus* 'prison' (< Tok Pisin *kalabus* 'prison', *kot* 'court, trial', *skùl* 'school', and *lotu* 'church, worship' (< Tok Pisin *lotu* 'church, worship'). Mass nouns like *monî* 'money' are also assigned to neuter 2.

4. Materials and method

This section describes how the materials are gathered and how the information on formal and semantic features in Mian are extracted. An overview of the computational classifiers used in this study is also provided. The following R (R-Core-Team 2019) packages are used to perform the quantitative analyses: data.table (Dowle and Srinivasan 2019), parsnip (Matt Kuhn and Vaughan 2019), randomForest (Liaw and Wiener 2002), randomForestExplainer (Paluszynska and Biecek 2017), recipes (Max Kuhn and Wickham 2019), reprtree (Banerjee, Ding, and Noone 2012), rpart (Therneau and Atkinson 2019), rpart.plot (Milborrow 2019), rsample (Max Kuhn, Chow, and Wickham 2019), tidyverse (Wickham 2017).

4.1 Materials

The information on grammatical gender is directly extracted from the Mian dictionary (Fedden n.d.), which contains 917 nouns. Four gender annotations are found in the dictionary: 'masculine', 'feminine', 'neuter 1', and 'neuter 2'. The value 'masc_fem' refers to nouns for which gender can be assigned to either masculine or feminine genders, as the noun denotes a higher animate but is not specific as to sex. The distribution of the five categories is shown in Figure 3.

The gender neuter 1 accounts for the majority of the nouns (48.2%, 442/917). Categories such as 'masc_fem' are admittedly small, however, they are still kept in the current sample to provide a faithful representation of the gender system in Mian.

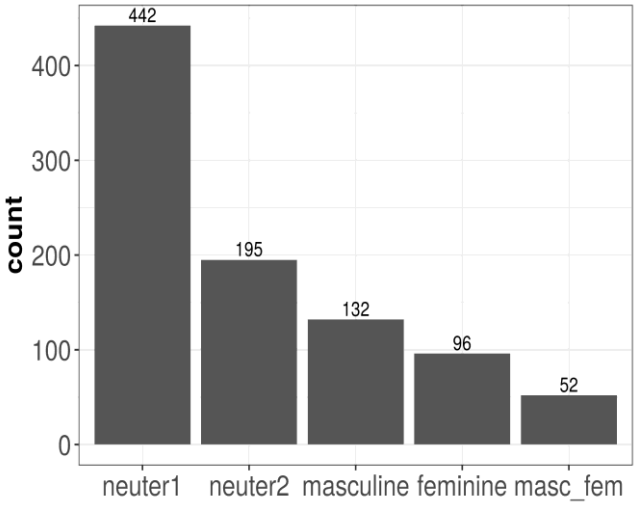

Fig3. The distribution of gender assignment in the Mian dictionary

The information on semantics is directly obtained from the existing categories in the dictionary. Salient semantic features of the nouns are annotated in the dictionary. The 15 most frequent semantic features are listed in Table 4. For instance, 122 nouns are semantically related to body parts in the corpus. A few examples of body parts are *báanyang* 'chin', *debelôn* 'forehead', *gabáam* 'head', among others.

Table 4. The most frequent 15 semantic features represented in the Mian dictionary

Feature	Count	Feature	Count	Feature	Count
body_part	122	tool	46	landscape	21
plant	119	human	44	female_human	18
entity	115	location	34	male_human	17
bird	101	insect	33	water_related	16
undefined	46	plant_part	28	animal	14

The information on form is extracted automatically from the entries of the dictionary. The first three and the last three phonemes of each entry are encoded for each noun. Since tone is a property of the whole word rather than individual phonemes, tone is extracted at the word level. As an example, for the word *báanyang* 'chin', the tone is LH, the first three phonemes are /b/, /a$^{\varsigma}$/, and /n/. The last three phonemes are '/y/, /a/, and /ŋ/.[4] The order of the phonemes is encoded in the data. This selection is based on the observation that nominal features mostly appear at the beginning and/or the end of nouns (Dryer 2013) under monosyllabic forms (Manova 2015; Sánchez-Gutiérrez, Mailhot, Deacon and Wilson 2018). An overview of the five most frequent phonemes for each of the three noun-final positions is shown in Table 5.

Table 5. The top 5 noun-final phonemes in the Mian dictionary

Last.first	Count	Last.second	Count	Last.third	Count

[4]We consistently use orthographic representations for Mian words. The mapping of letters to phonemes is one-to-one, except in the following cases, where more than one letter represents a single phoneme: <ng> = /ŋ/, <kw> = /k^{w}/, <gw> = /g^{w}/, <aa> = /a$^{\varsigma}$/, <ai> = /ai/, <au> = /au/, <aai> = /a$^{\varsigma}$i/, <aau> = /a$^{\varsigma}$u/, <ei> = /ɛi/, <ou> = /ou/.

n	168	a	187	l	98
m	152	i	163	m	96
l	127	o	113	t	78
ŋ	91	e	101	k	73
b	82	aˀ	72	b	65

The extracted formal and semantic features are then combined to create the training and testing data for the computational classifiers. A sample of the final data is shown in Table 6. The semantic features are encoded in the 'sem' variable, whereas the formal features are represented by the following variables: 'first.phoneme', 'second.phoneme', 'third.phoneme', 'last.first.phoneme', 'last.second.phoneme', 'last.third.phoneme', and 'tone'. The entries in the dictionary may not have a perfect matching with the phonemes in Mian. However, this matching is considered sufficiently high for the current study. The variables are thus named with the term 'phoneme' rather than 'character'.

Table 6. A sample of the data fed to the computational classifiers

	Example 1	Example 2
Noun	*aaleb*	*afunón*
gloss	father	shinbone
Gender	masculine	neuter 1
sem	male_kin	body_part
first.phoneme	aˀ	a
second.phoneme	l	f
third.character	ɛ	u
last.first.character	b	n
last.second.character	ɛ	o
last.third.character	l	n
tone	L	LH

As a summary, formal and semantic information is extracted automatically from the Mian dictionary. This information is then fed to different computational classifiers for training them on predicting the grammatical gender of nouns in Mian. The performance of the different computational classifiers is used as a representation of how useful formal and semantic information is in gender assignment of nouns in Mian. The following subsection provides an overview of the computational classifiers used in this study.

4.2 Method 1 – Decision trees

Three computational classifiers are used to assess the relevance of form and semantics with regard to predicting the grammatical gender of nouns in Mian. The first two computational classifiers are based on binary recursive partitioning (Breiman et al. 1984). The first computational classifier generates one decision tree based on the data and visualizes the interaction of the variables. The second computational classifier is called 'random forests'. The random forests computational classifier generates a series of 500 decision trees that are analyzed as a whole and used to assess the importance of each formal and semantic feature with regard to predicting the grammatical gender of nouns in Mian. The functioning of the tree-based

computational classifiers is summarized as follows. During the classification task, the data is recursively partitioned binarily to create homogeneous groups. The first computational classifier only generates one tree based on the provided data. A toy example of decision trees is shown in Figure 4 based on the iris dataset (Fisher 1936). The data consists of observational data on 150 flowers in total. 50 flowers from three different species: Setosa, Versicolor, and Virginica. Each flower is labeled with its sepal length, sepal width, petal length, and petal width. The data is fed to a decision-tree-based algorithm to try to predict the species of the flowers based on their sepal and petal measurements.

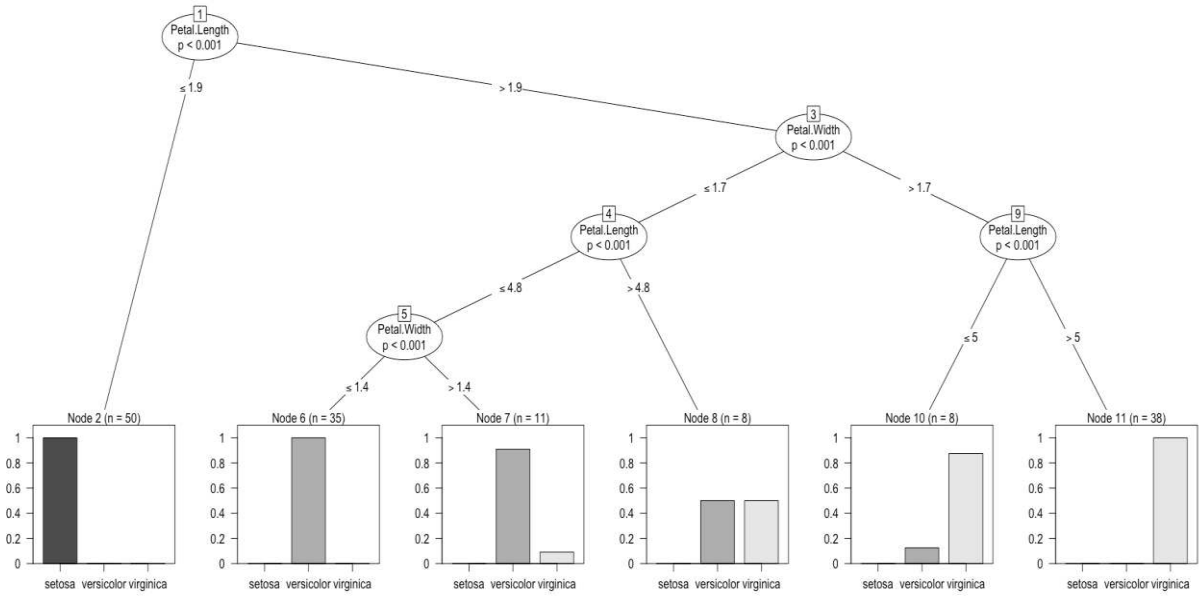

Fig4. Decision tree with flower species as the response variable and sepal length/width and petal length/width as explanatory variables (Tang, in press)

The tree can be read as follows. The bars of the buckets at the bottom of the graph indicate the ratio of the predicted flower species, i.e., Setosa, Virginica, and Versicolor. The homogeneity of the buckets is reflected by the correct predictions. As an example, Node 2 only has flowers from the Setosa species, which makes it extremely homogeneous. However, Node 8 has 50% of Versicolor flowers and 50% of Virginica flowers, which make it not homogeneous. A lack of homogeneity means that the model is not sure about the prediction and is very likely to result in wrong predictions. Thus, the higher the homogeneity of each node, the better the predictions. The decision tree can be read from the top. Depending on the answer at each numbered node, a prediction is made about the species of the flower being analyzed. For instance, if the petal length is shorter than 1.9 centimeters (Node 1 to Node 2), the flower is most likely a Setosa. If the petal length is larger than 5 centimeters (Node 9) and the petal width is above 1.7 centimeters, the flower is most likely a Virginica. The variables that are shown in the decision tree are the variables considered to have statistically significant interpretability on the data. The variables that are not included in the tree are considered not helpful to identify the species of the flowers. Information on sepal length and width is not included in our illustration.

The second computational classifier (i.e., random forests) uses the same algorithm, but generates a sample of trees. For each tree, the computational classifier uses a bootstrap sample of the entire dataset and a random subset of the variables encoded in the entire dataset. If the data is visualized as a table in which the rows represent the nouns and the columns the formal and semantic features, each partitioning selects a random sample of rows and columns. A statistical test is carried out for each random sampling. The results are considered statistically significant if the statistical test is consistently valid across most of the surveyed samples. This process of random sampling is also the main strength of random forests, as it allows the analysis

of small-scale data and consideration of the possible auto-correlation of variables (Tagliamonte and Baayen 2012).

4.2 Method 2 - Neural-network-based computational classifier

The third computational classifier uses a neural network architecture (Haykin 1998; Parks, Levine, and Long 1998). 'Neural network' is a non-linear discriminative classifier that searches for a boundary between the data points with regard to their predicted variable to minimize the classification errors. The current study uses a 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 variable (predictor) in the classification task. 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. Taking again the example from the iris data (Figure 5), the input layer on the left has four neurons that represent sepal length, sepal width, petal length, and petal width. The output layer has three neurons that represent the three flower species. In this example, there is only one hidden layer between the input layer and the output layer. The hidden layer has ten neurons.

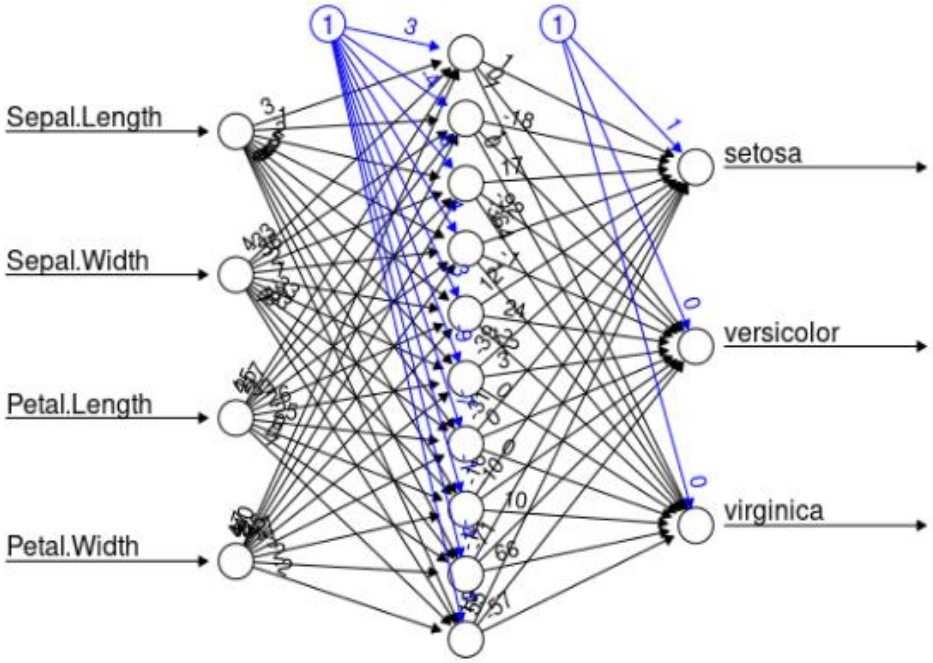

Fig5. An example of a feed-forward network based on the iris data (Fisher 1936)

The operation of the neural network computational classifier is summarized as follows. Each node in the hidden layer receives the inputs from the input layer. Each source of input are multiplied by weights (the arrows in the plot) and summed up. The sum is then transformed with a function and passed on to the nodes of the output layer as a result. The network is trained by searching for the weights that produce the desired output. The learning algorithm adjusts the connection weights between the neurons to minimize the divergence between the real values of the target variables and those predicted by the neural network computational classifier.

In the experiment on gender in Mian, the size of the input layer is identical to the number of variables. The size of the output layer is equal to the number of grammatical gender categories in Mian (four). The size of the hidden layer is set to ten. Additional experiments could be conducted to find which number and size of hidden layers results in the highest accuracy of the neural network classifier. As an example, the neural network architecture could have thousands of neurons and several hidden layers, which would increase its accuracy and the required computational power. Nevertheless, the current study is only interested in the relative

performance of formal and semantic features. Therefore, we have not conducted these additional experiments.

4.3 Comparing the output of the three computational classifiers

The first computational classifier can generate a decision tree to show the hierarchical interaction of the variables within the dataset. For instance, if both a formal and a semantic feature have a significant effect on predicting the grammatical gender of a noun, the decision tree will show which of the two variables has a stronger predictive power when they are both considered. The second computational classifier (i.e., random forests) can provide information on the relative importance of the predictors. The larger the importance of a variable, the more predictive it is. As an example, if the accuracy of the computational classifier drops the most when it does not take into account a specific feature, this feature is considered to have the highest ranking within all the variables. The third computational classifier using the architecture of neural networks cannot provide transparent information about the interaction of the variables. However, it can detect the presence of non-linear information that is not captured by tree-based computational classifiers. If the results are consistent with all three types of computational classifier, the conclusion of the hypothesis testing can be strengthened. As an example, if all three computational classifiers show a better performance based on semantic features, the results provide stronger evidence supporting the semantic dominance hypothesis.

All the computational classifiers are trained with 70% of the data as the training set and their performance is recorded based on the other 30% of the data. The training and test sets do not overlap. The percentages are consistently represented in each predicted gender. As an example, neuter 1 has 442 nouns. The training data thus has 311 (70.4%, 311/442) neuter 1 nouns while the test data has 131 (29.6%, 131/442) neuter 1 nouns. The percentages are not exactly 70% and 30% since the whole set of nouns does not exactly divide into 70% and 30%. The performance of the computational classifiers is assessed with two measures, accuracy and f-score. The accuracy provides an overview of the performance on the entire dataset. It is equal to the ratio of all the correctly retrieved tokens within the entire data. This value is expected to be compared with a baseline. On the one hand, it is possible to compare the accuracy of the model with the random baseline, which represents the accuracy the model would get by making totally random guesses. In our case, the random baseline would be equal to the square of the proportion of each gender category in the data, i.e., (442/917)*(442/917) + (195/917)*(195/917) + (132/917)*(132/917) + (96/917)*(96/917) + (52/917)*(52/917) = 31.2%. If the model can surpass this baseline, it is considered as performing better than chance. On the other hand, the random baseline is easily affected by the different sizes of each category in the data. Therefore, we select the majority baseline as a threshold, which refers to the biggest category in the dataset. Since most nouns in the Mian data are assigned neuter 1 gender (48.2%, 442/917), the computational classifier could reach a precision of 48.2% just by guessing that all the nouns belong to the gender neuter 1. Therefore, the performance of the computational classifiers based on the information of forms and/or semantics should at least exceed the accuracy of 48.2% to be considered as having good discriminatory power. This baseline is by default equal or higher than the random baseline, which makes it harder to beat and more reliable when evaluating the performance of computational models. The f-score is a combination of two other measures: precision and recall. Precision assesses how many tokens are correct in the output of the computational classifier, while recall evaluates how many tokens are correctly retrieved among all the expected correct output. The f-score is equal to the harmonic mean of the precision and recall, i.e. 2(recall × precision)/(recall + precision) (Ting 2010).

5. Results

The accuracy of the three computational classifiers is compared to the majority baseline in Figure 6. Accuracy refers to the ratio of correctly predicted tokens in the entire dataset, i.e. the number of nouns from the overall number of nouns that are assigned the correct gender by the computational classifier. For instance, if only 100 nouns of each of the four genders are predicted correctly, the accuracy of the computational classifier is 400/917 = 43.6%. On the one hand, the information extracted from formal features can only generate an accuracy slightly higher than the majority baseline. On the other hand, the performance of the computational classifiers is much higher when trained with semantic features. When both formal and semantic features are fed to the recursive partitioning-based computational classifiers, their accuracy slightly improves, but does not exceed by much the accuracy based on semantic features. These results suggest that semantic features provide most of the information relevant for gender assignment in Mian. In terms of types of computational classifier, using a sample of trees with random forests only improves the accuracy by 1% when both formal and semantic features are considered. This indicates that the classification task is not extremely complex and most of the information can be captured by a single tree.

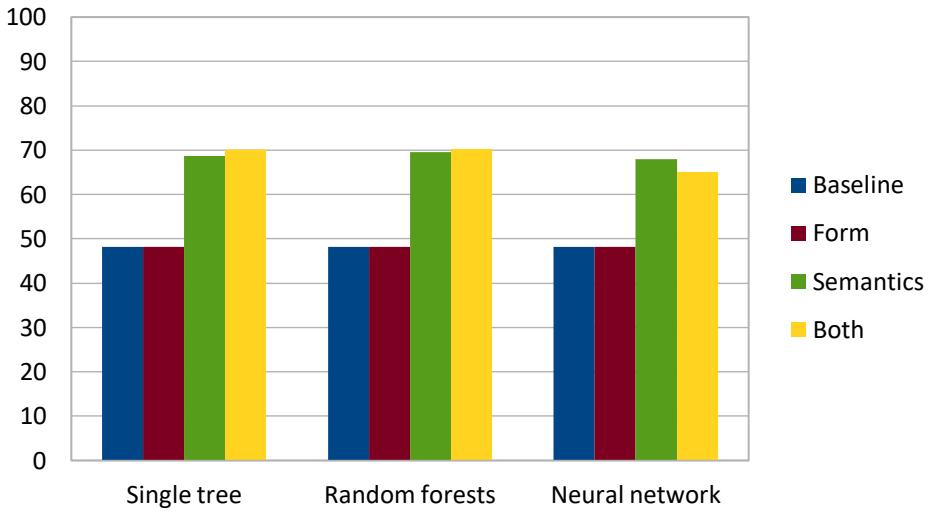

Fig6. The performance of the computational classifiers

The accuracy based on the formal and semantic features does not improve by much even when using a neural network architecture. This suggests that the current formal and semantic features can only provide so much for the task of gender assignment in Mian. For the neural network classifier, the accuracy of semantic features surpasses the accuracy of using both formal and semantic features. It is likely that the model detects patterns of forms that occur in the training data but are not found in the test data. The model thus overfits the training data and makes erroneous guesses when predicting the gender of the test data. This suggests that additional features and/or data are required to increase the accuracy of the classification. On the one hand, the semantic features represented in the dictionary are not generated automatically, which limits their width and coverage. Increasing the number of semantic features is expected to improve the accuracy of the computational classifiers. On the other hand, the form-based features are generated automatically but only for a small sample of words. The neural network computational classifier is expected to get more stable results with a larger set of nouns. As a summary, the random forests classifier and the neural network classifier have similar performance. Tree-based computational classifiers are less complex but their functioning logic is easier to understand and extract. The following analysis thus considers the tree-based computational classifiers as the main source of analysis.

Two types of output are extracted from the computational classifiers. First, a decision tree is generated from the single-tree-based computational classifier. This tree displays the interaction of the variables within the dataset. Second, the ranking of the formal and semantic

features is retrieved from the random forests computational classifier. Since the accuracy of the two methods is rather similar, the decision tree is expected to fairly accurately reflect the results based on a sample of trees. For both types of output, formal and semantic features are used so that their relevance can be assessed simultaneously.

The decision tree is shown in Figure 7. As mentioned in Section 4, this decision tree is generated based on 70% of the data and evaluated with the other 30%. The tree is read from the top node to the bottom nodes. A value smaller than 0.5 indicates 'no', while a value greater than 0.5 indicates 'yes'. As an example, starting from the top node (node 1), if the noun is annotated with the semantic feature of 'bird' (sem_bird), and if the noun has /l/ as the last but one phoneme (node 2), the noun is interpreted as 'masculine' (node 5, third from the left at the bottom) by the computational classifier. The tree indicates that form only plays a role within a semantically defined class, i.e., birds.

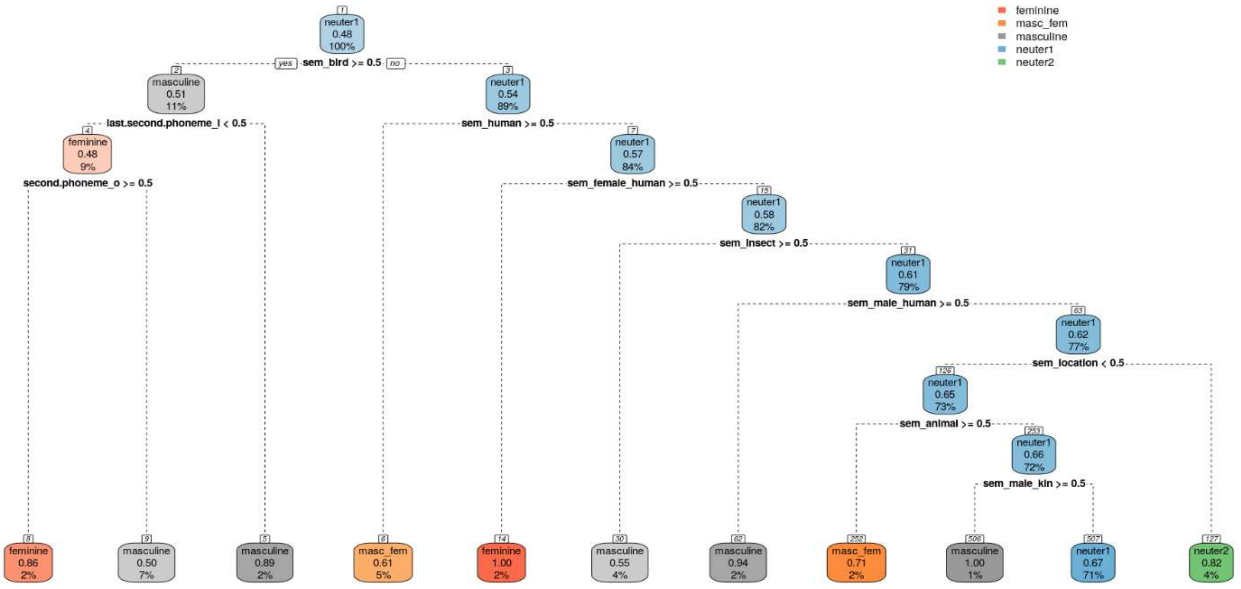

Fig7. The decision tree of gender assignment in Mian

The buckets at the bottom indicate the gender assigned by the computational classifier, the accuracy of each decision and the ratio of the data that is covered by this decision. Taking again node 5 as an example, the nouns fulfilling the criteria in node 1 and 2 are assigned to the masculine gender. This decision affects 2% of the entire dataset and the accuracy of the decision is 89%. For those 2% of nouns, 89% are interpreted correctly as masculine, but the other 11% are actually not masculine nouns.

The evaluation of the predictions based on this tree is displayed in Figure 8. The best overall performance (f-score) is found with the gender neuter 1. This is not surprising since neuter 1 is the biggest gender category in the data. It thus has more training data, and a higher probability to get chosen by chance. Neuter 2 is the second biggest gender category but the decision tree performs poorly on it. Most of the neuter 2 nouns are interpreted as belonging to neuter 1. A possible explanation to this erroneous classification is the partial semantic overlap between the nouns of the two genders. As an example, both genders have nouns referring to inanimate count nouns. Moreover, both genders also have nouns referring to masses and liquids (see Section 3.2).

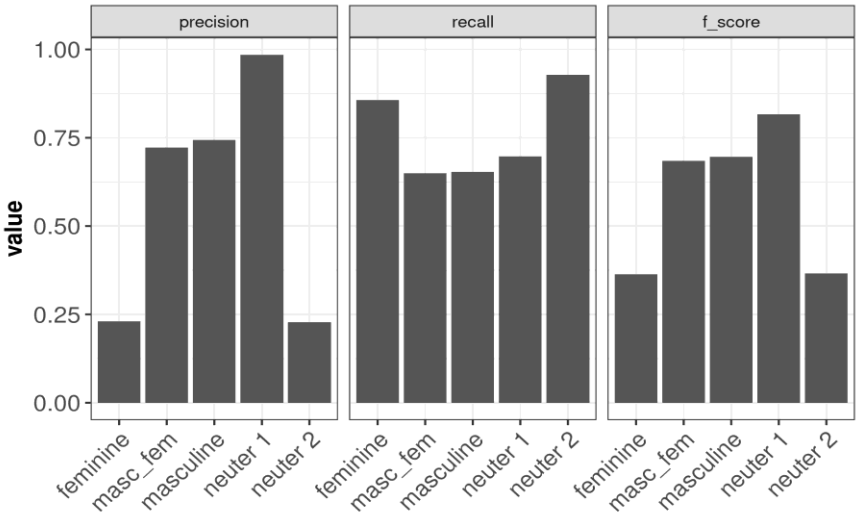

Fig8. The classification performance of the decision tree

To visualize the interaction across different gender categories, the confusion matrix based on the test set is provided in Table 7. A confusion matrix shows how the predictions of the model fit with the actual values in the data. The rows represent the actual values and the columns are the predicted values. The values on the diagonal are the correct predictions. For instance, 129 nouns are predicted correctly as neuter 1. A detailed analysis of the errors indicate that the model is having difficulties to distinguish between nouns from neuter 1 and neuter 2. A similar issue is found between the masculine and feminine genders. Feminine nouns tend to be erroneously classified as masculine, which also explains the high recall for feminine but the low precision for masculine.

Table 7. The confusion matrix of the decision tree based on the test set. The rows represent the actual values and the columns are the predicted values.

	feminine	masc_fem	masculine	neuter 1	neuter 2
feminine	6 (2.2%)	1 (0.4%)	12 (4.4%)	7 (2.5%)	0 (0.0%)
masc_fem	0 (0.0%)	13 (4.7%)	3 (1.1%)	2 (0.7%)	0 (0.0%)
masculine	1 (0.4%)	6 (2.2%)	32 (11.6%)	4 (1.5%)	0 (0.0%)
neuter 1	0 (0.0%)	0 (0.0%)	1 (0.4%)	129 (46.9%)	1 (0.4%)
neuter 2	0 (0.0%)	0 (0.0%)	1 (0.4%)	43 (15.6%)	13 (4.7%)

The second computational classifier based on random forests conducts a similar analysis but generates a sample of 500 randomized trees, which diminish the risks of overfitting the data. As an example, the test set based on one decision tree encounters the risk of being biased by a specific group of nouns occurring within the random split of the data between the training and test sets. The random forests classifiers randomize the tokens and the variables, which is expected to diminish the risk of accidental biases. The performance of the random forests computational classifier is shown in Figure 9. The best overall classification performance (f-score) is still found with neuter 1, which again, is not surprising due to the large data size of neuter 1. The performance on the other gender categories is much more stable than with only using one decision tree. This is also expected since the random forests randomizes the data.

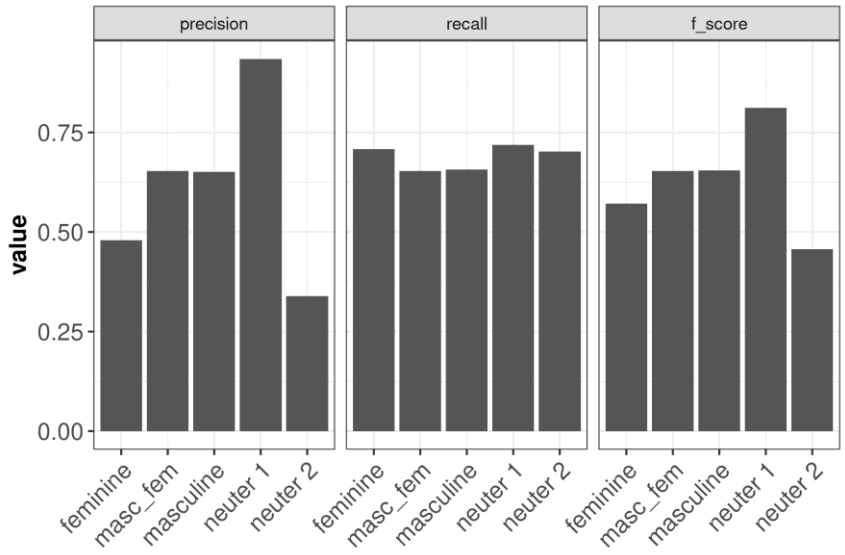

Fig9. The classification performance of the random forests

The error analysis of the random forests computational classifier is displayed in Table 8. Some tendencies found with the single-tree-based computational classifier are also found with the random forests classifier. A portion of feminine nouns is again erroneously affiliated to the masculine gender. Moreover, two thirds of the neuter 2 nouns are also erroneously affiliated to the neuter 1 gender.

Table 8. The confusion matrix of the decision tree based on random forests. The rows represent the actual values and the columns are the predicted values.

	feminine	masc_fem	masculine	neuter 1	neuter 2
feminine	46 (5.0%)	3 (0.3%)	34 (3.7%)	12 (1.3%)	1 (0.1%)
masc_fem	6 (0.7%)	34 (3.7%)	6 (0.7%)	6 (0.7%)	0 (0.0%)
masculine	12 (1.3%)	14 (1.5%)	86 (9.4%)	18 (2.0%)	2 (0.2%)
neuter 1	0 (0.0%)	1 (0.1%)	3 (0.3%)	413 (45.0%)	25 (2.7%)
neuter 2	1 (0.1%)	0 (0.0%)	2 (0.2%)	126 (13.7%)	66 (7.2%)

The randomization process also allows the extraction of the importance of the variables. The individual importance of the variables (i.e., the formal and semantic features) are assessed via the conditional permutation-based variable importance. If a variable is consistently helpful in predicting the family affiliation in most of the data subsets, it indicates that this variable has a high importance for the classification task. First, the frequency and the mean of the minimal depth for each variable within all the 500 trees generated by the random forests classifier are visualized. The minimal depth indicates how far the node with a specific variable is from the root node. As an example from Figure 5, the semantic feature of 'bird' is the root node, which equals to a minimal depth of zero. If a variable is frequently close to the root node, it is considered to have a high importance. The minimal depth of the top ten most important variables is shown in Figure 10. Three of the five top variables are semantic, which indicates that semantic features play a more important role in gender assignment in Mian. The gap between the first five variable increases quickly whereas the gap between the sixth and the tenth variable is small. This shows that there is no big difference of importance after the first few variables.

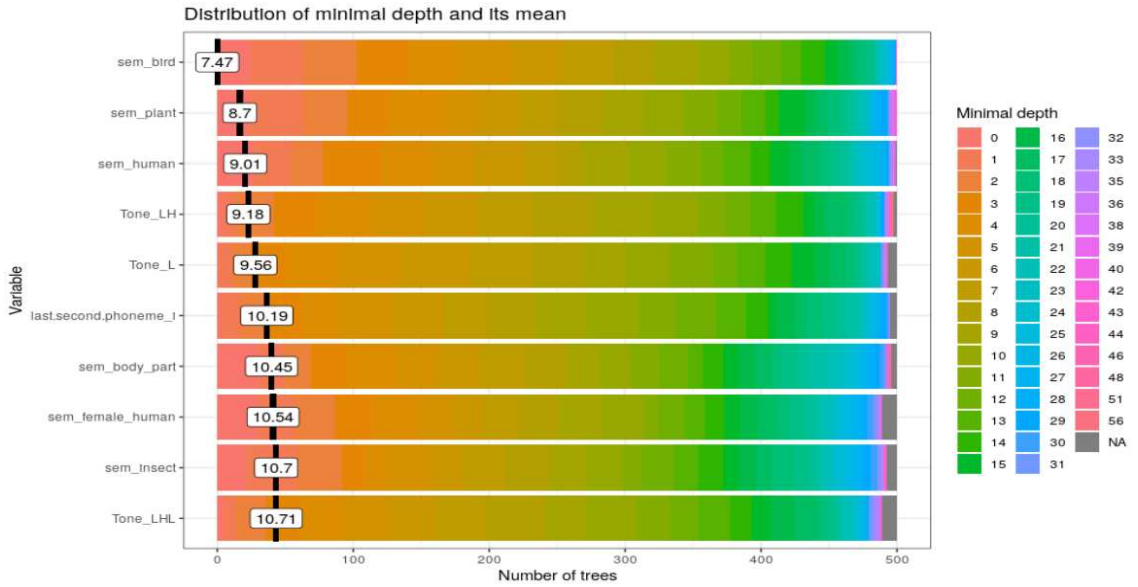

Figure 10. The distribution of the minimal depth and its mean

Similar results are found when using other measures. In Figure 11, the variables are ranked according to their effect on the accuracy and the homogeneity of the nodes. On the one hand, the mean decrease of accuracy indicates how worse the model performs without each variable. A high decrease infers that the variable has a strong predictive power. On the other hand, the mean decrease of the Gini coefficient indicates how each variable contributes to the purity of the nodes and the end of the tree. A high decrease of the Gini coefficient when removing a variable indicates that this variable has a strong predictive power and therefore a high importance. In both measures, the top ten ranked variables are all semantic. These results suggest that semantic features play a much more important role than formal features with regard to the gender assignment of nouns in Mian.

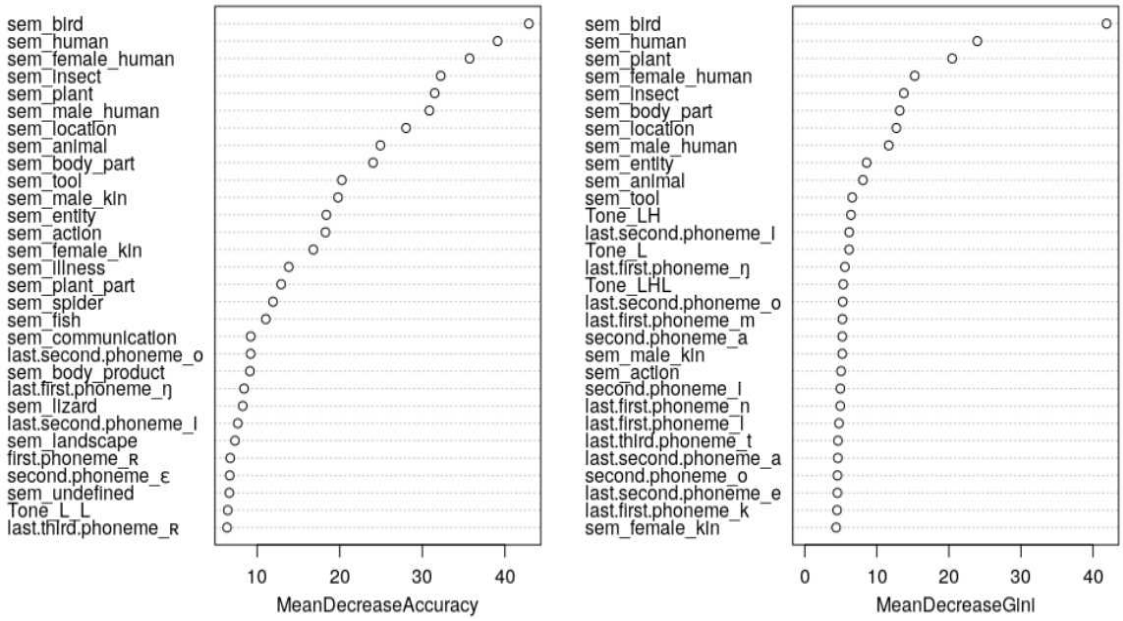

Figure 11. The accuracy and the purity of the nodes (Mean decrease of accuracy and Mean decrease of Gini coefficient)

Since the three measures also show variation within their results, an overview of the top ten variables in each measure is shown in Table 9. Seven variables are found consistently within all three measures: sem_bird, sem_plant, sem_human, sem_body_part, sem_female_human, and sem_insect. If only considering two measures at a time, the variables found in the last two

measures (Mean decrease of accuracy and Mean decrease of Gini coefficient) are almost identical, which indicates that their contribution is stable with regard to the classification task.

Table 9. The top ten variables for different measures of importance (variables found in all three measures in bold)

	Minimal depth	Accuracy	Purity
1	**sem_bird**	**sem_bird**	**sem_bird**
2	**sem_plant**	**sem_human**	**sem_human**
3	**sem_human**	**sem_female_human**	**sem_plant**
4	tone_LH	**sem_insect**	**sem_female_human**
5	tone_L	**sem_plant**	**sem_insect**
6	Last.second.phoneme_l	sem_male_human	**sem_body_part**
7	**sem_body_part**	sem_location	sem_location
8	**sem_female_human**	sem_animal	sem_male_human
9	**sem_insect**	**sem_body_part**	sem_entity
10	tone_LHL	sem_tool	sem_animal

5. Discussion

The results from the application of the three machine learning methods argue against any analysis of Mian based on an equal role for formal and semantic features. In table 9 semantic features are in the overwhelming majority for all three measures of importance. From the decision tree in Fig. 7 we can see that semantic features partition the assignment space at the higher level and that formal features have only a fine discriminatory role to play within one, semantically defined, class (birds). Furthermore, it is unlikely that an argument based on the markedness of the gender values for Mian would allow us to avoid this conclusion: in the confusion matrices for both the decision tree (table 7) and the random forests (table 8) neuter 2 is predicted for some neuter 1 nouns, and neuter 1 is predicted for some neuter 2 nouns; and masculine is predicted for some feminine nouns, while feminine is predicted for some masculine nouns. This indicates that a markedness hierarchy for the values would not be in a position to resolve conflicts of assignment.

Our motivation for using machine learning on a language for which the published analysis (Fedden 2011) treats its gender system as semantically based was to determine whether there were form-based generalisations that had been overlooked. The machine learning analyses, however, essentially confirm Fedden's (2011) original treatment, but with an interesting – if marginal – form-based generalization within one semantic domain. This indicates that we can certainly find systems in which semantics dominate, and it also suggests that the method we have presented could be used to test analyses of mixed systems based on semantic or formal features, so as to determine how these interact. It should also be noted that the semantic categories drawn on for machine learning are not specific to these analyses, but have also been used in Evans, Brown and Corbett's (2002) analysis of Bininj Gun-wok.

There are, of course, limitations in our study. The dictionary used only includes 917 nouns. While it is relatively large for an underdocumented language, it is still small compared to an entire lexicon. Additional data should be added to avoid the risks of inflating the importance of some semantic features. As an example, body parts account for 13.3% (122/917) of the Mian data. This ratio is likely to be inflated in comparison to the ratio found in other languages for which large-scale dictionaries with more than 20,000 of nouns are available. More semantic features should also be added or extracted automatically to be comparable with the

automatically extracted formal features. Additional experiments are required on other languages to investigate the effect of the data size on the results. Furthermore, given the data available we have not been able to determine the effect of token word frequency in relation to the role of form-based and semantic assignment of gender in Mian.

An interesting area for future work is the difference of transparency between the formal and semantic features. By this we mean that the effect of semantic features is easily traceable whereas formal features have more complex patterns that may be difficult to identify and require a larger data size. If both formal and semantic features result in a predicting accuracy of 80%, but a simple model can get 80% for semantic features while a complex model is needed for formal features to get 80%, even if both features reach an accuracy of 80%, the semantic information is much more transparent and easier to extract. In our analysis, we only looked at the final accuracy, but for future work it will be interesting to weight the accuracy according to the complexity of the models, and this will also be beneficial when evaluating competing analyses of individual languages.

It is also important to move beyond well-known Indo-European gender systems, which represent only a small part of the typological space, so as to test hypotheses that have been developed largely with them in mind, and in providing tested empirical evidence that the semantic-dominance hypothesis is real for some languages with gender we are contributing to a wider investigation of nominal classification. Recent work (Fedden & Corbett 2018) has looked at the key dimensions of semantics and exponence to describe the typological space of both classifier and gender systems. In examining the role of semantic and formal features in the gender system of one language we have demonstrated the importance of quantifying the degree to which semantic features may dominate, rather than assuming that there is a type 'gender' that can be easily described as an optimal combination of form-based and semantic assignment. In demonstrating empirically that this assumption is unwarranted we are also better placed to cover the full range of systems and to consider how they may arise.

6. Conclusion

We extracted information on formal and semantic features from a dictionary of Mian in order to test the original analysis of the gender system as being predominantly semantic. The information was fed to tree-based and neural-network-based computational classifiers. The performance of the computational classifiers was used as an indicator of the value of the type of information (semantics or phonology) in the gender assignment of nouns. The results show that semantic features consistently provide more information than formal features when it comes to predicting the grammatical gender of Mian nouns. Thus, they provide evidence confirming the status of semantics-dominant systems in some languages.

The methods used here enable us to understand in greater detail the role of semantics and form in noun-classification systems such as gender and, if applied to a larger sample of languages, have longer-term promise for evaluating hypotheses on gender assignment such as the *semantic-dominance hypothesis* and the *equality hypothesis*. Our results also suggest that the level of transparency should also be taken into account in future work when assessing the role of semantics and form.

Acknowledgements: To be added

References

Baerman, Matthew, Dunstan Brown, and Greville G Corbett. 2005. *The Syntax-Morphology Interface: A Study of Syncretism*. Cambridge: Cambridge University Press.

Banerjee, Mousumi, Ying Ding, and Anne Michelle Noone. 2012. "Identifying Representative Trees from Ensembles." *Statistics in Medicine* 31 (15): 1601–16. https://doi.org/10.1002/sim.4492.

Breiman, Leo, Jerome Friedman, Charles J Stone, and Richard Olshen. 1984. *Classification and Regression Trees*. New York: Taylor & Francis.

Comrie, Bernard. 1999. "Grammatical Gender Systems: A Linguist's Assessment." *Journal of Psycholinguistic Research* 28 (5): 457–66. https://doi.org/10.1023/A:1023212225540.

Contini-Morava, Ellen, and Marcin Kilarski. 2013. "Functions of Nominal Classification." *Language Sciences* 40 (November): 263–99. https://doi.org/10.1016/j.langsci.2013.03.002.

Corbett, Greville G. 1982. "Gender in Russian: An Account of Gender Specification and Its Relationship to Declension." *Russian Linguistics* 6 (2): 197–232.

———. 1991. *Gender*. Cambridge: Cambridge University Press.

———. 2012. *Features*. Cambridge: Cambridge University Press.

———. 2013a. "Number of Genders." In *The World Atlas of Language Structures Online*, edited by Matthew S Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology.

———. 2013b. "Systems of Gender Assignment." In *The World Atlas of Language Structures Online*, edited by Matthew S Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Corbett, Greville G, and Sebastian Fedden. 2016. "Canonical gender." *Journal of Linguistics* 52 (3): 495-531.

Corbett, Greville G, and Norman Fraser. 2000. "Gender Assignment: A Typology and a Model." In *Systems of Nominal Classification*, edited by Gunter Senft, 293–325. Cambridge: Cambridge University Press.

Corbett, Greville G, Sebastian Fedden, and Raphael Finkel. 2017. "Single versus concurrent systems: Nominal classification in Mian." *Linguistic typology* 21 (2): 209-260.

Corteen, Emma. 2019. "The Assignment of Grammatical Gender in German: Testing Optimal Gender Assignment hypothesis." PhD dissertation, Cambridge: Cambridge University.

Dobrin, Lise. M. 1999. "Phonological Form, Morphological Class, and Syntactic Gender: The Noun Class Systems of Papua New Guinea Arapeshan." PhD dissertation, Chicago: University of Chicago.

Donohue, Mark. 1997. "Tone Systems in New Guinea." *Linguistic Typology* 1 (3): 347–86.

Dowle, Matt, and Arun Srinivasan. 2019. "Data.Table: Extension of Data.Frame." *R Package Version* 1.12.2. https://CRAN.R-project.org/package=data.table.

Dryer, Matthew S. 2013. "Prefixing vs. Suffixing in Inflectional Morphology." In *The World Atlas of Language Structures Online*, edited by Matthew S Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Dryer, Matthew S, and Martin Haspelmath. 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Evans, Nicholas, Dunstan Brown and Greville G. Corbett. 2002. "The semantics of gender in Mayali: Partially parallel systems and formal implementation". *Language*, 78: 111-155.

Fedden, Sebastian. 2011. *A Grammar of Mian*. Berlin: Walter de Gruyter.

Fedden, Sebastian (n.d.) Mian dictionary, Word and Excel files.

Fedden, S., & Corbett, G. G. (2018). Extreme classification. *Cognitive Linguistics, 29*(4), 633-675. doi:http://dx.doi.org/10.1515/cog-2017-0109

Fisher, Ronald A. 1936. "The use of multiple measurements in taxonomic problems." *Annual Eugenics*, 7 (3): 179–188.

Haykin, Simon. 1998. *Neural Networks: A Comprehensive Foundation*. New Jersey: Prentice Hall PTR.

Healey, Alan. 1964. "A Survey of the Ok Family of Languages, Reconstructing Proto-Ok." PhD dissertation, Canberra: Australian National University.

Hockett, . 1958. "A course in modern linguistics". *Language Learning*, 8(3-4): 73-75.

Kemmerer, David. 2014. "Word Classes in the Brain: Implications of Linguistic Typology for Cognitive Neuroscience." *Cortex* 58 (September): 27–51. https://doi.org/10.1016/j.cortex.2014.05.004.

———. 2017. “Categories of Object Concepts across Languages and Brains: The Relevance of Nominal Classification Systems to Cognitive Neuroscience.” *Language, Cognition and Neuroscience* 32 (4): 401–24. https://doi.org/10.1080/23273798.2016.1198819.

Kibrik, Aleksandr E., K. I. Kazenin, E. A. Ljutikova & S. G. Tatevosov (eds.). 2001. Bagvalinskij jazyk: Grammatika: Teksty: Slovari [Bagvalal: Grammar, texts, dictionaries]. Moscow: Nasledie.

Kuhn, Matt, and Davis Vaughan. 2019. “Parsnip: A Common API to Modeling and Analysis Functions.” *R Package Version* 0.0.3.1. https://CRAN.R-project.org/package=parsnip.

Kuhn, Max, Fanny Chow, and Hadley Wickham. 2019. “Rsample: General Resampling Infrastructure.” *R Package Version* 0.0.5. https://CRAN.R-project.org/package=rsample.

Kuhn, Max, and Hadley Wickham. 2019. “Recipes: Preprocessing Tools to Create Design Matrices.” *R Package Version* 0.1.6. https://CRAN.R-project.org/package=recipes.

Lakoff, George, and Mark Johnson. 2003. *Metaphors We Live By*. London: The University of Chicago Press.

Liaw, Andy, and Matthew Wiener. 2002. “Classification and Regression by RandomForest.” *R News* 2 (3): 18–22.

Manova, Stela. 2015. *Affix ordering across languages and frameworks.* Oxford: Oxford University Press.

Milborrow, Stephen. 2019. “Rpart.Plot: Plot Rpart Models: An Enhanced Version of Plot.Rpart.” *R Package Version* 3.0.8. https://CRAN.R-project.org/package=rpart.plot.

Nesset, Tore. 2006. “Gender Meets the Usage-Based Model: Four Principles of Rule Interaction in Gender Assignment.” *Lingua* 116 (9): 1369–93. https://doi.org/10.1016/j.lingua.2004.06.012.

Nichols, Johanna. 1996. “Head-Marking and Dependent-Marking Grammar.” *Language* 62: 56–119.

Paluszynska, Aleksandra, and Przemyslaw Biecek. 2017. “RandomForestExplainer: Explaining and Visualizing Random Forests in Terms of Variable Importance.” *R Package Version* 0.9. https://CRAN.R-project.org/package=randomForestExplainer.

Parks, Randolph W., Daniel S. Levine, and Debra L. Long, eds. 1998. *Fundamentals of Neural Network Modeling: Neuropsychology and Cognitive Neuroscience*. Computational Neuroscience. Cambridge, Mass: MIT Press.

Pawley, Andrew. 2005. “The Chequered Career of the Trans New Guinea Hypothesis: Recent Research and Its Implications.” In *Papuan Pasts: Cultural, Linguistic and Biological Histores of Papuan-Speaking Peoples*, edited by Andrew Pawley, Robert Attenborough, Jack Golson, and Robin Hide, 67–108. Canberra: Pacific Linguistics.

Prince, Alan, and Paul Smolensky. 1993. *Optimality hypothesis: Constraint Interaction in Generative Grammar*. Boulder: Rutgers University of Colorado.

R-Core-Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rice, Curt. 2006. “Optimizing Gender.” *Lingua* 116 (9): 1394–1417. https://doi.org/10.1016/j.lingua.2004.06.013.

Ross, Malcolm. 2005. “Pronouns as a Preliminary Diagnostic for Grouping Papuan Languages.” In *Papuan Pasts: Cultural, Linguistic and Biological Histories of Papuan-Speaking Peoples*, edited by Andrew Pawley, Robert Attenborough, Jack Golson, and Robin Hide, 15–66. Canberra: Pacific Linguistics.

Sánchez-Gutiérrez, Claudia, Hugo Mailhot, Hélène Deacon, and Maximiliano Wilson. 2018. “MorphoLex: A derivational morphological database for 70,000 English words.” *Behavior Research Methods* 50 (4): 1568-1580.

Seifart, Frank. 2010. “Nominal Classification.” *Language and Linguistics Compass* 4 (8): 719–36.

Tagliamonte, Sali A, and 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: 135–78.

Tang, Marc. in press. "A simple introduction to programming and statistics with decision trees in R" *Teaching Statistics*.

Therneau, Terry, and Beth Atkinson. 2019. "Rpart: Recursive Partitioning and Regression Trees." *R Package Version* 4.1-15. https://CRAN.R-project.org/package=rpart.

Ting, Kai Ming. 2010. "Precision and Recall." In *Encyclopedia of Machine Learning*, edited by Claude Sammut and Geoffrey I. Webb, 781–781. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-30164-8_652.

Wickham, Hadley. 2017. "Tidyverse: Easily Install and Load the Tidyverse." *R Package Version* 1.2.1. https://CRAN.R-project.org/package=tidyverse.

Wurm, Stephen. 1982. *Papuan Languages of Oceania*. Tubingen: Narr.

Zaliznjak, Andreij. 1973. "On Ponimanii Termina 'Padež' v Lingvističeskix Opisanijax." In *Problemy Grammatičekogo Modelirovanija*, edited by Andreij Zaliznjak, 53–87. Moscow: Nauka.

Zasorina, Lidija N. (ed.) 1977. *Chastotnyj slovar' russkogo jazyka*. Moscow: Russkij Jazyk.