A typology of classifiers and gender: From description to computation

Marc Allassonnière-Tang

ACTA UNIVERSITATIS UPSALIENSIS
Studia Linguistica Upsaliensia
23

A typology of classifiers and gender

From description to computation

Marc Tang

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in Humanistiska teatern, Thunbergsvägen 3H, Uppsala, Saturday, 9 March 2019 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Sebastian Fedden (University Sorbonne nouvelle-Paris III).

Abstract
Tang, M. 2019. A typology of classifiers and gender. From description to computation. *Studia Linguistica Upsaliensia* 23. 78 pp. Uppsala: Acta Universitatis Upsaliensis.
ISBN 978-91-513-0507-3.

Categorization is one the most relevant tasks realized by humans during their life, as we consistently need to categorize the things and experience that we encounter. Such need is reflected in language via various mechanisms, the most prominent being nominal classification systems (e.g., grammatical gender such as the masculine/feminine distinction in French). Typological methods are used to investigate the underlying functions and structures of such systems, using a wide variety of cross-linguistic data to examine universality and variability. This analysis is itself a classification task, as languages are categorized and clustered according to their grammatical features. This thesis provides a cross-linguistic typological analysis of nominal classification systems and in parallel compares a number of quantitative methods that can be applied at different scales.

First, this thesis provides an analysis of nominal classification systems (i.e., gender and classifiers) via the description of three languages with respectively gender, classifiers, and both. While the analysis of the first two languages are more of a descriptive nature and aligns with findings in the existing literature, the third language provides novel insights to the typology of nominal classification systems by demonstrating how classifiers and gender may co-occur in one language in terms of distribution of functions. Second, the underlying logic of nominal classification systems is commonly considered difficult to investigate, e.g., is there a consistent logic behind gender assignment in language? is it possible to explain the distribution of classifier languages of the world while taking into account geographical and genealogical effects? This thesis addresses the lack of arbitrariness of nominal classification systems at three different scales: The distribution of classifiers at the worldwide level, the presence of gender within a language family, and gender assignment at the language-internal level. The methods of random forests, phylogenetics, and word embeddings with neural networks are selected since they are respectively applicable at three different scales of research questions (worldwide, family-internal, language-internal).

Keywords: Classifiers, Gender, Nominal classification, Functions, Random Forests, Phylogeny, Word Embeddings, Neural Networks

Marc Tang, Department of Linguistics and Philology, Box 635, Uppsala University, SE-75126 Uppsala, Sweden.

ISSN 1652-1366
ISBN 978-91-513-0507-3
urn:nbn:se:uu:diva-366598 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-366598)

Man rår ej för att glädjetårar rinner,
när man bland vänner är

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I Eliasson, Pär and Marc Tang. 2018. The lexical and discourse functions of grammatical gender in Marathi. *Journal of South Asian Languages and Linguistics*, 5(2). doi.org/10.1515/jsall-2018-0012

II Saikia, Pori and Marc Tang. (Submitted). Nominal classification in Assamese: An analysis of function. In Marc Tang and Marcin Kilarski (eds.), *Nominal Classification in Asia: Functional and diachronic perspectives*. Amsterdam: John Benjamins.

III Tang, Marc and Marcin Kilarski. (Submitted). Functions of gender and numeral classifiers in Nepali. *Poznan Studies in Contemporary Linguistics*.

IV Tang, Marc and One-Soon Her. (in press). Numeral base, numeral classifier, and noun: Word order harmonization. *Language and Linguistics*, 22(1).

V Tang, Marc and One-Soon Her. (in press). Insights on the Greenberg-Sanches-Slobin Generalization: Quantitative typological data on classifiers and plural markers. *Folia Linguistica*.

VI Her, One-Soon, Marc Tang and Bing-Tsiong Li. (in press). Word order of numeral classifiers and numeral bases: Harmonization by multiplication. *Language Typology and Universals*.

VII Her, One-Soon and Marc Tang. 2018. A statistical explanation of the distribution of sortal classifiers in languages of the world via computational classifiers. *Journal of Quantitative Linguistics*. doi.org/10.1080/09296174.2018.1523777

VIII Tang, Marc and Michael Dunn. (Submitted). Measuring the phylogenetic signal of grammatical gender in Indo-Aryan languages. *Language Dynamics and Change*.

IX Basirat, Ali and Marc Tang. (in press). Linguistic information in word embeddings. In Randy Goebel, Yuzuru Tanaka, and Wolfgang Wahlster (eds.), *Lecture notes in artificial intelligence*. Dordrecht: Springer.

X Basirat, Ali and Marc Tang. 2018. Lexical and Morpho-syntactic Features in Word Embeddings: A Case Study of Nouns in Swedish. *Proceedings of the 10th International Conference on Agents and Artificial Intelligence - Volume 1: NLPinAI*, 663-674. doi.org/10.5220/0006729606630674

Reprints were made with permission from the publishers.

Acknowledgments

Writing this thesis was a challenging but enjoyable task thanks to the amazing team I have been working with. Each discussion I had with colleagues and friends at Uppsala University and other institutions around the globe was always able to cheer me up and give me energy and willingness to carry on.

First of all, I am extremely thankful for all the supervision and help I received. Every knock on the door was always welcomed by a cheerful "Come in!" and a big smile, regardless of how busy that day could have been. Comments and suggestions were always constructive and pushed me to investigate further. When I obviously did not think through my research question thoroughly, I was gently suggested to re-organize my thoughts, which was very helpful in improving the quality of research while maintaining a strong motivation to do so.

I am also extremely thankful with regard to all the suggestions and advices I received from my colleagues and friends. The endless discussions and the creative drawings in our office were very productive and gave birth to many research projects that we so far managed to complete and expand. The clear and simple explanation of complex algorithms during some of these conversations encouraged me to learn more computational methods and played an important role in the structure of this thesis. The fruitful sharing at the Uppsala Working Group on Empirical linguistics and the GIS for language study network also let me learn much more than I could have done by myself.

The positive interaction with colleagues always had a quick-charge effect on me and was able to boost back my energy when it was at a low point (coffee probably also played a minor role in this). Administratively, every question I had was consistently answered and explained. Initiatives were also welcomed and discussed. Special thanks to the LingSing choir and its organizers as well as active participants who invested a lot of their personal time to provide happy memories to all members of the department! Likewise, the pudding group and the tough vikings deserve a special note. Even though the members often depart for their own new journey, our regular contact and chat is a great source of happiness. Outside of Uppsala, the long conversations at various conferences not only created (and continue to create) research projects but also involved very entertaining discussions that we hopefully will continue in future meetings.

I also thank my parents, my brother, and my grandmother who always approved my decisions in their respective way. Last but not least, my wife and our two girls M&M have always been supportive since we have been living together. This milestone could not have been reached without their affection.

Contents

1. Introduction

Categorization is one the most relevant tasks realized by humans during their life, as we consistently need to categorize the things and experience that we encounter (Lakoff and Johnson, 2003, p. 162-163). Such need is reflected in language via various tools, the most prominent being nominal classification systems (e.g., grammatical gender). To understand the underlying functions and structures of such systems, typological methods investigate their universality and variability across languages based on a wide variety of cross-linguistic data (Clahsen, 2016, p. 599). Interestingly, this process of analysis is by itself a classification task, as languages are categorized and clustered according to their grammatical features. This thesis thus tries to undergo a cross-linguistic typological analysis of nominal classification systems while running in parallel a typological analysis of the quantitative methods that can be applied in this cross-linguistic typological analysis. Due the format of compilation, this section of the thesis only provides an overview of the topics involved, whereas detailed methods and results are reported separately in each individual paper.

1.1 Current trends in linguistics

Thanks to the growth of technology, the gathering and comparison of cross-linguistic data became much easier and more efficient in the recent decades. Such change of environment also had an impact on research trends, as the construction of digitalized databases is becoming increasingly important for researchers and research-funding institutions. Comparative analyses based on this new type of data also had a significant effect on theoretical definitions and theories in linguistic. Taking the topic of nominal classification as an example, previous approaches commonly considered languages in terms of binary features, e.g., does a language have grammatical gender or not? Under such view, well-known gender languages such as German, French, Spanish, among others are easily classified as gender languages; whereas languages such as Mandarin Chinese, Korean, and Japanese are typically labeled as non-gender languages. However, recent comparative analyses have shown that a wide variety of languages are not that conveniently assigned to one of the two categories. By way of illustration, Assamese (Eastern Indo-Aryan) marks biological gender (masculine/feminine) on personal pronouns and a few adjectives but does not mark it on the verbs (Kalita, 2003). Should it be counted as

having a grammatical gender system or not? Another question also relates to how many gender systems are found in a language? While the distinction seems relatively simple in French with biological gender, some languages show a much more complex picture. As an example, Nepali (Northern Indo-Aryan) marks animacy on personal pronouns but labels adjectives with biological gender (Acharya, 1991). The productivity and stability of grammatical gender may also vary according to register, e.g., in Sinhala (Dhivehi-Sinhala), grammatical gender is more consistently marked in written texts rather than in oral speech (Gair, 2007, p. 775).

Observations of this kind have shown the diversity of languages and recent studies adopted the view of continuum with regard to nominal classification (Grinevald and Seifart, 2004). Languages with different parameters are positioned on a lexical-grammatical continuum according to their level of grammaticalization. Under such view, French would be considered to have a gender system more grammaticalized than Assamese. The same logic has been applied in the analysis of concurrent gender systems in one language. This approach allows the grading of a gender system in terms of canonicity (Corbett and Fedden, 2016; Fedden and Corbett, 2017), e.g., the French gender system would be considered more prototypical than in Assamese. While previous research has shown that these concepts of lexical-grammatical continuum and canonicity are applicable in terms of qualitative analysis, few studies have applied them to large set of languages as databases are commonly built on more general features such as number of genders due to limitation of data and resources (Corbett, 2013b; Di Garbo, 2014). One of the main aims of this thesis is therefore to apply both binary and continuum approaches in the building of databases for nominal classification and investigate what information different quantitative methods can retrieve from these two types of data.

With regard to quantitative methods, the rise of statistics and programming language has been observed within various scientific fields such as natural sciences. These methods have also been recently introduced to linguistics. While extreme caution is advised with regard to the selection of their input and the interpretation of their output, their scientific potential is undeniable and they can definitely bring novel insights and supports to linguistics. However, statistics and programming language are commonly perceived as a difficult and unpleasant subject (Albarracin et al., 2017; Ben-Zvi and Garfield, 2004; Carnell, 2008; Gould, 2010), which results in students and researchers being generally reluctant to invest an extensive amount of time and resource on programming and statistics since they are already suffocating under the current workload of their projects (Buckley et al., 2015; MacInnes, 2009). Such decision commonly originates from a low probability in terms of return of investment. Students and researchers are not willing to invest time and energy in learning a new method that may not be directly applicable to their research, especially if the data used in the learning process are not directly related to their work. The importance of context and practice is thus one of the major

keys to engage students and researchers in learning/using programming and statistics (Brown, 2017; Wild and Pfannkuch, 1999). Numerous pedagogical publications have addressed such need by providing learning materials directly related to linguistics, e.g., *How to do linguistics with R* by Natalia Levshina (2015). Another main aim of this thesis is therefore to use different types of quantitative methods and demonstrate their strengths and weaknesses when combined with linguistic data.

1.2 Materials and methods

In this section, we summarize the motivation for selecting nominal classification (mostly) in Indo-Aryan as a subject of case study. Moreover, we also explain the reason for choosing the three methods of random forests, phylogenetics, and word embeddings with neural networks. The details of individual languages and methods are not explained in this section, since they are included in Chapter 2 and 3 of the thesis.

1.2.1 Why nominal classification?

Linguists are investigating nominal classification systems (how languages classify nouns of their lexicon) due to their various lexical and pragmatic functions as well as correlation with cognitive and cultural facets of human behavior. Within languages of the world, the two most common nominal classification systems are grammatical gender and sortal classifiers. As an example, French possesses grammatical gender and categorizes all nouns into masculine or feminine. This categorization is reflected grammatically via agreement with other elements of the clause (Corbett, 1991), c.f., *un grand ballon* (one.MASC.SG big.MASC.SG ball) 'a ball' and *une grande table* (one.FEM.SG big.FEM.SG table) 'a table'. On the other hand, Mandarin Chinese apply sortal classifiers and classify objects according to their specific inherent features such as shape, c.f., *yi4 ke1 qiu2* (one CLF.ROUND ball) 'a ball' and *yi4 zhang1 zhuo1zi0* (one CLF.2D table) 'a table'[1].

Languages may rely on different types of nominal classification systems to express several lexical and discourse functions (Contini-Morava and Kilarski,

[1]Sortal classifiers are different from measure words and measure terms (Kilarski, 2014, p. 9). Measure words do not categorize but denote the quantity of the entity named by noun, e.g., in Mandarin Chinese *san1 ping2 shui3* (three MENS.BOTTLE water) 'three bottles of water'. Sortal classifiers, on the opposite, do not provide such information of quantity; they classify a noun inherently and designate semantic features inherent to the noun. Measure terms refer to phrases in English such as *three bottles of water*. In English, the measure term 'bottles' carries plural marking and requires the insertion of 'of' before the following noun. Such syntactic requirements are not observed with measure words, the two categories of measure words and measurer terms are thus distinguished.

2013). By way of illustration, the use of different sortal classifiers on the same noun may indicate different referents in Mandarin Chinese, c.f., *yi2 li4 yu4mi3* (one CLF.GRAIN corn) 'a corn niblet', *yi4 gen1 yu4mi3* (one CLF.STICK corn) 'a corn cob', *yi4 zhu1 yu4mi3* (one CLF.PLANT corn) 'a corn plant', and *yi4 ke1 yu4mi3* (one CLF.TREE corn) 'a corn plant'. Moreover, the categories of object concepts that are encoded by languages suggest that the neural underpinnings of such concepts are not only affected by universal tendencies but also by cultural idiosyncrasies (Kemmerer, 2017a,b). For instance, Hindi (Indo-Aryan) and Pnar (Austroasiatic) speakers maintain a patrimonial and matrimonial kinship system respectively. Grammatical gender systems of both languages share universal tendencies based on human cognition, i.e. associating long, thin, vertical objects with the masculine grammatical gender whereas round, flat, horizontal ones are associated with the feminine grammatical gender. However, these grammatical gender systems also distinguish between different sociocultural values as in the Hindi language, objects of large size are generally assigned to the masculine gender; whereas in Pnar, large sized objects tend to be associated with feminine gender. All these factors combined together show that nominal classification is a highly cross-disciplinary topic worth investigating, e.g., it is related to linguistics, psychology, neuroscience, sociology, among others.

1.2.2 Why Indo-Aryan (and beyond)?

With regard to their spatial distribution, gender and classifiers almost display a complementary distribution. Grammatical gender languages are mostly found in Africa, Australia, Europe, Oceania, and part of the Americas (Aikhenvald, 2000, p.78); whereas classifier languages are commonly located in Asia and sporadically attested in Africa, Europe, and the Americas (Figure 1.1).

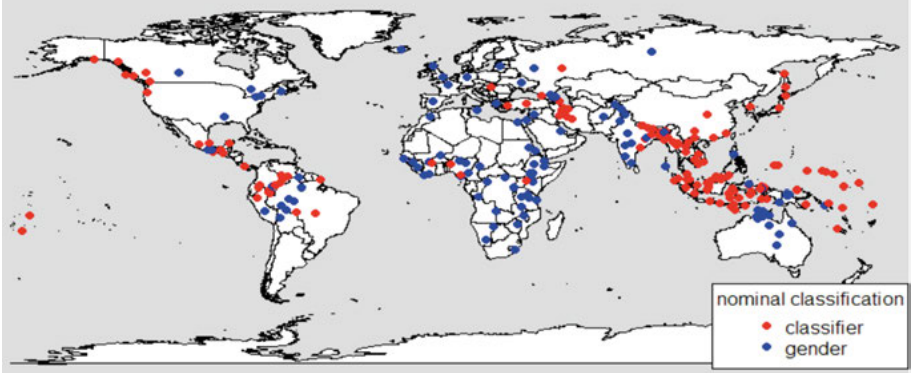


Figure 1.1. Spatial distribution of gender and classifiers (Corbett, 2013a; Gil, 2013).

Recent studies of nominal classification have mostly scrutinized languages from Europe, South America, and Papua New Guinea with a qualitative approach. For instance, languages of these regions have been investigated in terms of concurrent nominal classification systems, e.g., Minangkabau (Austronesian) rely on both noun classifiers and sortal classifiers (Marnita, 1996), while Paumari(Arawan) has a masculine/ feminine gender system plus a human/ non-human gender system (Fedden and Corbett, 2017). Other studies have also been done on co-existing classifier and gender systems in South American languages, e.g., Palikur Aikhenvald (2000). This thesis focuses on another region possessing languages with complex nominal classification systems. The northern region of India is the meeting point between Indo-European and Sino-Tibetan languages, which possess prototypical examples of gender and sortal classifier systems, respectively. This condition makes the languages of this zone unique in the sense that they show a co-occurrence of sortal classifiers and grammatical genders. Such concurrent systems are uncommon statistically (Sinnemaki, 2018). Gender generally occur in inflectional languages while classifiers tend to occur in languages with less complex inflectional morphology. The co-existence of the two systems in the same language creates a conflict between these opposing tendencies. Moreover, the way genders and classifiers categorize nouns is drastically different: gender languages such as French categorize all nouns into either masculine and feminine, while classifier languages such as Mandarin Chinese class nouns according to the inherent properties of their referents, e.g., animacy and shape. The co-existence of these systems in the same language thus creates a unique patchwork of semantic and morphosyntactic properties.

1.2.3 Why these three methods?

Linguistic analysis can target different types of data. Macro-analyses typically involve large datasets of languages that are annotated with various linguistic features. The World Atlas of Language Structures (Dryer and Haspelmath, 2013) is a good example of such approach. This type of analysis generally includes languages from various families and provides an overview of a specific grammatical feature in languages of the world. While the strength of this method is its ability to provide a big picture of a linguistic feature, this is also its weakness. When including a big quantity of languages, it is extremely difficult to design values that can accurately mirror the observations in every language. As an example, a database on word order between the subject, the verb, and the object may point out the dominant order in a language but can hardly take into account all the possible variations in speech as some languages have a flexible word order without a dominant pattern, e.g., in German the dominant order is subject-verb-object in main clauses without an auxiliary and subject-object-verb in subordinate clauses and clauses with an

auxiliary (Dryer, 2013). Micro-analyses targeting one specific language can scrutinize its grammatical features and generate a more faithful representation; nevertheless, it lacks the insights of comparative studies.

We propose three different methods that are suitable for three different types of data with different scale. First, random forests is a computational classifier that can be used to investigate the probabilistic universals proposed by linguists in languages worldwide. Second, phylogenetic models can target an intermediate size of data with language-family-internal analysis and take into account factors such as Galton's problem (Levinson et al., 2011, p. 511). Probabilistic universals can therefore be evaluated in terms of language evolution, whereas concentrating on a smaller set of languages also augments the precision and accuracy of the grammatical features. Finally, word embeddings and neural networks can be used as a tool for language-internal analysis. Word embeddings extract the linguistic information from large size corpora as word vectors, which can then be fed to the neural network classifier. While these three methods are by no means restricted to a specific scale of analysis, we can demonstrate how they could be applied on different types of linguistic data. Most of the analyses in this thesis are done via the programming language R (R-Core-Team, 2018) due to its open-access availability, but the same methods can equally be found in other programming languages.

1.3 Aims and contributions

This thesis has three major aims, which are listed as follow. First, it provides an analysis of nominal classification systems (i.e., gender and classifiers) via the description of three languages with respectively gender (I), classifiers (II), and both (III). While the first two papers are more of a descriptive nature and align with findings in the existing literature, the latter provides novel insights to the typology of nominal classification systems by demonstrating how three nominal classification systems (two gender systems and one sortal classifier system) may co-occur in one language in terms of distribution of morphosyntactic structure and functions (III). Second, the underlying logic of nominal classification systems is commonly considered difficult to investigate and/or inexistent, e.g., is there a consistent logic behind gender assignment in language? is it possible to explain the complementary-like distribution between gender and classifiers in language of the worlds while taking into account geographical and genealogical effects? This thesis addresses the lack of arbitrariness of nominal classification systems at three different scales: The distribution of classifiers at the worldwide level (IV,V,VI,VII), the presence of gender within a language family (VIII), and gender assignment at the language-internal level (IX,X). Third, this thesis introduces new applications of quantitative methods from biology and computer science to answer the questions mentioned previously. The methods of random forests

(IV,V,VII), phylogenetics (VIII), and word embeddings with neural networks (IX,X) are selected since they are respectively applicable at three different scales of research questions (worldwide, family-internal, language-internal). These methods have already been applied in the linguistic literature, but not in the subjects involved in this thesis. Random forests has mostly been used in corpus analysis but not on probabilistic universals, phylogenetics analyses were generally based on cognate-coded data instead of grammatical features, whereas word embeddings and neural networks have been widely used in computational linguistics but less in general linguistics.

2. A descriptive account of gender and classifiers

The papers in this section follow a synchronized structure. Each paper first presents an overview of its nominal classification system (i.e., gender and/or classifier), then the main lexical and discourse functions of these systems are displayed. The three papers are of a descriptive nature, but provide novel data to studies of nominal classification. Paper I and II on Marathi and Assamese bring additional data of prototypical gender and classifier systems in the Indo-Aryan language family, whereas Paper III shows concrete examples of how gender systems and classifiers may co-exist in one language in terms of structure and function. All three described languages originate from the Indo-Aryan language family, their respective location is shown in Figure 2.1.

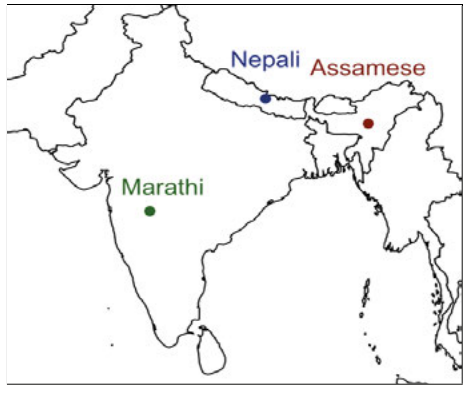


Figure 2.1. Location of the three languages included in the current chapter.

Assamese belongs to the Eastern Indo-Aryan subgroup and is thus found in North-East of India. Nepali is affiliated to the Northern Indo-Aryan subgroup and is located in Nepal. Marathi is labeled as a member of the Southern Indo-Aryan subgroup and is spoken in the South-West of India. These three languages are geographically ranging from East to West and represent very different types of nominal classification systems, even though they belong to the same Indo-Aryan language family. They have been chosen to represent the three main points of the lexical-grammatical continuum of nominal classification systems. Moreover, the nominal classification systems of these languages have been described as an overview in previous studies, but they have not been scrutinized in terms of functions. We adopt the framework of

functional typology (Contini-Morava and Kilarski, 2013) to fill this gap. In functional typology, the typical functions of nominal classification systems are categorized into lexical and discourse functions. Lexical functions relate to expansion of the lexicon, differentiation of referents, individuation, and ascription of properties. On the other hand, discourse functions include reference identification, reference management, and re-presentation. We analyze if these functions are fulfilled by the nominal classification system of each of the three languages. Such analysis may then be used as references for future linguistic studies, e.g., in pragmatics. In the following sub-sections, we summarize the structure and functions of nominal classification system of each language, whereas full details are found in each individual paper.

2.1 Prototypical gender: Marathi (I)

The Marathi language is mostly spoken in the state of Maharashtra in India, and is ranked within the top five major languages in India in terms of speaker population. Due to its versatile linguistic environment (e.g., Indo-Aryan languages to the North and Dravidian to the South), Marathi has developed some non-Indo-Aryan phonological or syntactic features (Dhongde and Wali, 2009, p. 2), but its gender system has being preserved. While previous studies mostly focused on morphosyntax and general descriptions (Apte, 1962; Gupte, 1975; Joshi, 1993; Nayudu, 2008; Wali, 1989; Dhongde and Wali, 2009; Wali, 2006), we aimed at providing a more detailed description and functional analysis of grammatical gender in Marathi.

2.1.1 The gender system of Marathi

Nouns are obligatorily classified as masculine, feminine, or neuter in the grammatical gender system of Marathi (Dhongde and Wali, 2009, p. 40). Grammatical gender refers to grammatical agreement (Corbett, 1991). The use of different lexical items to distinguish between the biological gender of referents is not sufficient to represent grammatical gender, c.f., *kutraa* ‘dog(male)’ and *kutrii* ‘dog(female)’. Grammatical gender typically requires a stable grammatical agreement between the controller (i.e., the noun) and other elements in the clause. The domain of agreement is generally considered as asymmetric and local, e.g., if there is grammatical gender agreement between the noun and the adjective, the gender marking of the adjective solely depends on the noun and not vice-versa. Moreover, the agreement is consistently found within the boundaries of a specific syntactic phrase, such as the noun phrase (Corbett and Fedden, 2016, p. 499). As an example in (1), gender in Marathi is mirrored through grammatical agreement on the possessive pronouns, adjectives, and verbs. The three sentences display a similar structure; nevertheless, the form of the demonstrative, the adjective, and the verb varies according to

the masculine (3a), feminine (1b), or neuter (1c) gender of the noun that is being referred to. For instance in (1a), the masculine gender of the noun *mitra* 'friend (male)' is reflected through grammatical agreement on the first person possessive pronoun *maajhaa* (1.POSS.M), the adjective *baraa* (good.M) and the verb *distaa* (look.PRS.M).

(1) Gender agreement on the verb with present tense in Marathi

a. *maajh-aa mitra bar-aa disto.*
POSS-1SG.M friend(SG.M) good-SG.M look.PRS.3SG.M

'My friend(male) looks good.'

b. *maajh-ii maitriin bar-ii diste.*
POSS-1SG.F friend(SG.F) good-SG.F look.PRS.3SG.F

'My friend(female) looks good.'

c. *maajh-a ghar bar-a dista.*
POSS-1SG.N house(SG.N) good-SG.N look.PRS.3SG.N

'My house looks good.'

Grammatical gender in Marathi complies with the three features of canonical gender system (Corbett and Fedden, 2016; Fedden and Corbett, 2017). First, the gender values match agreement classes, i.e., we observe three distinct marking for three different gender classes without overlap. By way of illustration, the agreement pattern displays three different agreement markers for three different gender classes, c.f., *maajhaa* (POSS.1SG.M), *maajhii* (POSS.1SG.F), and *maajha* (POSS.1SG.N). Second, the gender of nouns is constant and invariable in Marathi. For instance, the grammatical gender of the noun form *ghar* (house) is consistently neuter and does not vary arbitrarily. Third, "gender can be read unambiguously off the lexical entry of a noun" (Corbett and Fedden, 2016, p. 527). In other words, the gender of a noun is inferred from its grammatical agreement rather than pure lexical form. As an example, even though the nouns *kutra* 'dog' and *mitra* 'friend' share the same word-final vowel, their different gender (i.e., neuter and masculine) can be inferred from the agreement pattern with other elements in a clause, c.f., (1a) and (1c). An overview of the gender agreement in Marathi is shown in Table 2.1.

Table 2.1. *Overview of grammatical gender marking in Marathi*

		Singular	Plural
Verb	Past	+	+ (3rd)
	Present	+	-
	Future	-	-
Adjective	Variable	+	+
	Invariable	-	-
Pronoun	Personal (3rd)	+	+
	Demonstrative	+	+
	Possessive	+	+

Verbs generally mark gender agreement in the past tense and present tense. The target of agreement (i.e., subject or object) may vary with regard to transitive verbs, but the agreement itself remains mandatory. Intransitive verbs agree with their subjects and transitive verbs usually agree with their objects instead in the past tense. However, the latter may be neutralized in cases of emphasis on the object by means of adding the suffix *-laa* on the object. With regard to adjectives, phonological constraints result in the lack of agreement on consonant-final adjectives, while the other adjectives systematically agree with their subject. Finally, pronouns commonly mark gender (with exception of personal pronouns, which only mark gender on the third person), regardless of grammatical number.

2.1.2 Functions of gender in Marathi

The stability of the gender system can provide various tools for Marathi speakers to expand the referential power of the lexicon and facilitate the process of discourse. As summarized in Table 2.2, most lexical and discourse functions listed by functional typology are equally attested in Marathi. Grammatical gender in Marathi may expand the power of the lexicon through the use of diminutives or nominalizers, c.f., *aarsaa* 'mirror (masculine)' and *aarsii* 'small mirror (feminine)'. Moreover, grammatical gender can help to differentiate among animate and inanimate referents, e.g., by distinguishing the biological sex among animates. Furthermore, grammatical gender in Marathi equivalently contributes to the function of individuation, as mass nouns tend to be affiliated to the neuter gender. Finally, speakers of Marathi may also apply different grammatical gender on a noun to convey their subjective attitude toward the referent of the noun, e.g., neuter gender may be used to indicate distance and/or indifference with a human referent, as in *te veda asa karta* (that.N idiot so do.PRS.3SG.N) 'that idiot is doing like that'.

Table 2.2. *Functions of grammatical gender in Marathi*

Type	Function	Example in Marathi
Lexical	Expansion of lexicon	- Nominalization - Differentiation in terms of size
	Differentiation of referents	- Indication of biological sex among animates - Expression of semantic distinction among inanimates
	Individuation	- Tendency to associate mass nouns to the neuter gender
	Ascription of properties	- Expression of the speaker's distance from animates - Indication of crudeness for abstract concepts
Discourse	Reference identification	- Identification of referents via anaphora and deixis
	Reference management	- Neuter tends to be affiliated with the general sense - Difference of gender marking with regard to emphasis
	Re-presentation	- Indication of changes in the speaker's perspective during discourse

With regard to discourse functions, grammatical gender in Marathi is used to identify referents along discourse. For instance, grammatical gender may help to track referents and avoid constant repetition of the noun form. Moreover, grammatical gender may equally indicate the prominence of the referent within discourse, i.e., different gender may convey different degree of distance with the referent and infers the status of the referent in the discourse context. Last but not least, the lexical function of ascribing properties may be applied across discourse and indicate a change of attitude from the speakers toward the referent.

2.2 Prototypical classifier: Assamese (II)

Assamese belongs to the Indo-Aryan group of the Indo-European language family and has approximately 13 million people as first language speakers according to the 2011 census of India. It is mostly spoken in the North-East of India within the state of Assam, and also serves as a lingua franca to various communities across North East India (Goswami and Tamuli, 2003; Moral, 1997). What makes Assamese worth studying regarding nominal classification is the fact that while Indo-European languages are commonly using grammatical gender, Assamese (along with its Indo-Aryan neighbors such as Bengali) uses classifiers instead.

2.2.1 The classifier system of Assamese

Classifiers in Assamese occur in a bound form and their use with numerals or with nouns when the numerals are absent in a noun phrase is mandatory. However, they are not an indigenous feature of the language, as the earliest references to the usage of sortal classifiers in Assamese can only be traced back to the first half of the fourteenth century (Barz and Diller, 1985, p. 169). The current inventory of classifiers in Assamese can be divided into seven types according to their meaning, as shown in Table 2.3. Classifiers for animates distinguish between biological sex and social status, with a general classifier used for non-human animates. Classifiers for inanimate, on the other hand, are generally categorized according to shape and size, which is also a common semantic parameter in nominal classification systems (Kemmerer, 2017a, p. 412). As a reminder, we only include here sortal classifiers and do not discuss measure words or measure terms.

Table 2.3. *Inventory of sortal classifiers in Assamese*

Classifier	Category	Meaning
- ta	general	
-zon	human	masculine
-zoni	human	feminine
-zona	honorific	deities/saints
-goraki	honorific	humans highly respected
- khon	shape	flat
-dal	shape	long and round
-khila	shape	flat and flexible
-pat	shape	flat and narrow
-ti	size	small and appealing
-zupa	plant	bushy
-pah/ -pahi	plant	flower
-khini	mass	

The general classifier *-ta* commonly refers to inanimate and non-human animates, as in *tini-ta mekuri* (three-CLF.GENERAL cat) ‘three cats’. Moreover, the general classifier *-ta* also occurs with all inanimate nouns that do not take a specific classifier, e.g., nouns such as *kompjuter* ‘computer’ rely on the general classifier as in *e-ta kompjuter* (one-CLF.GENERAL computer) ‘a computer’ since they are new additions to the Assamese vocabulary and no specific classifiers have been assigned to them. Regarding human animates, Assamese has specific classifiers with either male or female referents. *-zon* is used with male humans, while the classifier *-zoni* is used for female humans, c.f., *sari-zon manuh* (four-CLF.MALE.HUMAN human) ‘four men’ and *sari-zoni manuh* (four-CLF.FEMALE.HUMAN human) ‘four women’. When expressing respect toward the referent, the gender neutral honorific classifiers *-goraki* and *-zona* are used instead. *-goraki* is appropriate for conversations and texts, whereas *-zona* is specifically related to deities and saints (Borah,

2012, p. 301), c.f., *sari-goraki hikhjok* (four-CLF.HON teacher) ‘four teachers’ and *sari-zona guru* (four-CLF.HIGH.HON religious.teacher) ‘four (religious) teachers’. As for inanimates, they are mostly classified by shape. For instance, the classifier *-dal* is used for round and long entities, as in *sari-dal rosi* (four-CLF.LONG.ROUND rope) ‘four ropes’. Occasionally, shape classifiers may also apply to non-human animates, as long as the referent shares the specific feature of the classifier, e.g., *du-dal kesu* (two-CLF.LONG.ROUND earthworm) ‘two earthworms’.

Two classifiers slightly diverge from the common categories. First, the diminutive classifier *-ti* refers to any object that is very small and appealing in nature, c.f., *e-zon lora* (one-CLF.MALE.HUMAN boy) and *e-ti lora* (one-CLF.SMALL boy) ‘a cute little boy’. Second, the classifier-*khini* is used to indicate that the referred noun is uncountable, as in *pani-khini* (water-CLF.MASS) ‘the water’. This is unusual considering that classifier languages are generally using measure words for mass nouns rather than assigning them to a specific sortal classifier (Her, 2012a, 1677). The literature also diverges according to the status of this classifier (Borah, 2012; Goswami and Tamuli, 2003), further investigation within the language is thus required to clarify its theoretical label.

Finally, the word order of classifiers may vary according to definiteness. As demonstrated in (2), the classifier *-khon* is used for objects of flat shape such as beds. The occurrence of the noun *bisona* ‘bed’ before the numeral and the classifier implies a definite reading in (2a), while the indefinite reading appears in (2b) with the structure of [Numeral-Classifier Noun]. Such phenomenon is quite common among South-East Asian and Southern Sinitic languages (Bisang, 1999; Li and Bisang, 2012).

(2) Sample of sortal classifiers in Assamese

a. *bisona* *du-khon* *anilu.*
bed two-CLF.FLAT bring.PST.1.

‘I brought the two beds.’

b. *du-khon* *bisona* *anilu.*
two-CLF.FLAT bed bring.PST.1

‘I brought two beds.’

As a summary, the contemporary sortal classifier system in Assamese includes thirteen sortal classifiers and categorizes referents by the following criterion: human animates are classified according to their biological sex and social status, while non-human animates and inanimate nouns occur with either the general classifier or shape/size/plant/mass classifiers.

2.2.2 Functions of classifiers in Assamese

All the lexical and discourse functions of classifiers in Assamese may be outlined in Table 2[1]. Lexical functions relate to the effect of classifier on the noun in the domain of classifier structures, while discourse functions include the discourse context. As an example, the lexical function of ascribing properties allows the speaker to convey her/his attitude toward the referent by using different classifiers on the noun. For instance, a baby may be referred to via the size classifier *-ti* instead of the conventional general classifier *-ta* or the male human classifier *-zon* to show that the speaker views the baby as small and cute, i.e., *kesuwa-ti* (baby-CLF.SMALL) 'the cute little baby'. The lexical function of ascribing properties is thus restricted to one clause. However, if the speaker uses the function of ascribing properties several times during discourse it becomes the discourse function of re-presentation. As a reminder, these functions are also applicable in languages with grammatical gender, as shown in Table 2.2 with Marathi.

Table 2.4. *Functions of classifiers in Assamese*

Type	Function	Example in Assamese
Lexical	Differentiation of referents	Different classifiers on the same noun may link to different referents
	Individuation	Count nouns apply sortal classifiers while mass nouns use measure words
	Ascription of properties	Different classifiers on the same noun may show the attitude of the speaker
Discourse	Reference identification	Classifiers help referent tracking in anaphora, deixis and disambiguation
	Reference management	Different orders of numeral, classifier and noun convey definiteness
	Re-presentation	Different classifiers on the same noun in discourse show changes of attitude

The sortal classifier system in Assamese is generally similar to other classifier languages, in exception to two functions: individuation and definiteness marking. As for individuation, in classifier languages such as Mandarin Chinese, mass nouns are categorized via measure words. However, the Assamese classifier system provides two options. Mass nouns may apply measure words but they also have a specific classifier for mass. The existence of such classifier may have implication in the count/mass distinction across languages. As for definiteness marking, the use of word order to mark definiteness is not present in all classifier languages of the world. Mandarin Chinese is often viewed as a typical classifier language (Zhang, 2013, p. 1-2) but does not rely on word order to mark definiteness, i.e., the construction [Noun Classifier] would be

[1]The lexical function of expansion of the lexicon is not listed since classifiers generally come from nouns and thus cannot serve to expand the lexicon.

ungrammatical, **ben shu* (CLF.VOLUME book). This differentiation may be due to areal variation or different development processes of classifiers, i.e., item-oriented versus category-oriented (Bisang, 1999, p. 159), since definiteness marking via classifiers is also present in other languages such as Wu and Cantonese (Li and Bisang, 2012).

2.3 Mixed: Nepali (III)

Nepali is mostly spoken in Nepal, in which Indo-Aryan languages are found in the south-western part, while Tibeto-Burman languages are spoken in the north-eastern region. The status of Nepali as the official language of Nepal and its use as lingua franca in the entire area result in a superficial knowledge of the language among L2 speakers. Nepali is thus subject to strong influence from other languages. As an example, even in the capital Kathmandu, where Nepali is used in government matters, commerce and communication, the Tibeto-Burman language Newari also had a long literary tradition, resulting in influence on Nepali. Similarly, in the Darjeeling area, Nepali is spoken alongside Bangla and Tibetan. Although both areas are acknowledged as literary centers of Nepali, even there Nepali has been subject to influence from other languages.

This variation is reflected in the gender systems and classifier systems within languages of the area. Generally speaking, grammatical gender is present in most Indo-European languages of Nepal. However, due to intense contact with Sino-Tibetan languages lacking this feature, gender agreement is realized less consistently in some languages (Weidert, 1984, p. 205). For example, gender agreement in modern Maithili only appears in certain tenses, e.g., the past tense and formal registers (Yadav, 1996, p. 63-64). The opposite development may also occur, as illustrated by the influence of Hindi on written Nepali, where "... the continued marking of the feminine in verbal forms, and the persistence of feminine endings for some nouns and adjectives may be attributable at least in part to these strong features of Hindi grammar." (Riccardi, 2003, p. 545). Considerable variation is also found in classifier systems. Only Newari is attested to possess a rich classifier system (Kiryu, 2009), while the inventories of classifiers in most other languages are restricted. For example, Awadhi distinguishes five classifiers (Barz and Diller, 1985, p. 162) and Bhojpuri has two (Verma, 2007). Other languages such as Tamangic languages are reported to have a classifier-like construction with numerals, but usually with only one classifier, as in Eastern Tamang *gor som jha* (CLF three son) 'three sons' (Lee, 2011, p. 32).

With regard to Nepali, there is a striking lack of agreement as to the actual properties of gender and classifiers. This concerns in particular the number of genders and classifiers. For example, the number of genders stated in the literature varies between two (Acharya, 1991, p. 99; Matthews, 1998, p. 23-

28; Poudel, 2010), four (Manders, 2007, p. 52) and eleven (Pokharel, 2010, p. 40). As regards sortal classifiers, the commonly attested inventory only includes two human/non-human classifiers (Acharya, 1991, p. 100; Matthews, 1998, p. 54; Riccardi, 2003, p. 559-560). However, more recent descriptions suggest that the number of classifiers is actually much higher. For example, Pokharel (Pokharel, 2010, p. 53) claims that Nepali has developed more than 200 sortal classifiers due to language contact with Tibeto-Burman and Austro-Asiatic languages. This discrepancy found in the literature on Nepali can be attributed to the areal variation mentioned above. Similarly to the other languages of Nepal, there is considerable variation in the complexity and expression of both gender and sortal classifiers. In view of these issues, while there is a long tradition of research on language contact in South Asia that has dealt with such phenomena (Emeneau, 1956; Priestly, 1983; Barz and Diller, 1985), the status of both gender and sortal classifiers in Nepali is controversial and calls for further research.

2.3.1 The gender systems of Nepali

In our study, we analyze nominal classification in Nepali based on Fedden and Corbett (2017)'s typology of concurrent nominal classification systems and interpret Nepali as having two gender systems distinguishing between the masculine/feminine and human/non-human oppositions plus a sortal classifier system. With regard to the masculine/feminine gender system, nouns denoting male animates are masculine, nouns denoting female animates are feminine, and the residue composed of inanimates is assigned to the masculine gender. The masculine/feminine agreement is found on adjectives, verbs, possessive adjectives, ordinal numbers, and the general classifier. An example is illustrated in (3), where the possessive adjectives, the adjectives, and the verbs distinguish between masculine and feminine forms.

(3) Gender agreement in Nepali

a. *mer-o* *ramr-o* *keto* *nepali* *bolcha*
POSS-1SG.M beautiful-M boy(SG.M) Nepali speak.PRS.3SG.M

'My handsome boyfriend speaks Nepali.'

b. *mer-i* *ramr-i* *keti* *nepali* *bolche*
POSS-1SG.F beautiful-F girl(SG.F) Nepali speak.PRS.3SG.F

'My beautiful girlfriend speaks Nepali.'

The human/non-human gender system is found in third person pronouns, c.f., *u* 'he/she' and *tyo* 'it'. Both male and female humans are referred to via the human pronoun, e.g., *u ramr-i che* (he/she beautiful-F be.PRS.3SG.F) 'she is beautiful' and *u ramr-o cha* (he/she beautiful-M be.PRS.3SG.M) 'he is hand-

some'. In those examples, while the pronoun does not change form between male and female human referents, the masculine/feminine agreement on the adjective and the verb indicates that the pronouns in the two examples are having referents of the opposite sex. The same distinction can be found when using the non-human pronoun to refer to non-human animates. For instance, the following sentences could be used when referring to a female buffalo and a male buffalo respectively, c.f., *tyo ramr-i che* (it beautiful-F be.PRS.3SG.F) 'it is beautiful' and *tyo ramr-o cha* (it beautiful-M be.PRS.3SG.M) 'it is beautiful'.

While the human/non-human gender system is only pronominal, the masculine/feminine gender system involves grammatical agreement, which may be subject to reduction represented by the lack of feminine agreement in informal speech (Matthews, 1998, p.150; Riccardi, 2003, p.555; Upadhyay, 2009, p.575). This is especially found within the use of Nepali as a second language by speakers of languages without gender. For instance, this type gender reduction is much more frequent among the sociolects of Nepali spoken by bilingual Tibeto-Burman speakers and their monolingual children, e.g., in the speech of the Darjeeling dialect where the majority of the speakers come from a Tibeto-Burman family background (Pokharel, 2010, p.56).

2.3.2 The classifier system of Nepali

Most available descriptions of Nepali only mention two sortal classifiers: *jana*, which occurs with human countable nouns, and *wota*, which occurs with non-human countable nouns (Acharya, 1991, p. 100). We further developed observations in recent studies that suggested a much larger classifier inventory (Pokharel, 1997) and analyzed ten sortal classifiers. They include a general and a human classifier, together with a number of classifiers for inanimates that involve contrasts based on shape, dimensionality and material. The ten classifiers are illustrated in Table 2.5 along with their respective semantic domains and examples.

Table 2.5. *Inventory of sortal classifiers in Nepali*

Classifier	Category	Example
wota	general	book, car, shop, telephone
jana	human	man, woman, uncle, aunt
dana	round fruits	apple, grape, orange
sinka	long object	noodle, bamboo
ghoga	long plant	maize
geda	grain	mustard, maize, rice
koso	natural capsule	banana, bean, pea
khili	artificial capsule	cigarette, betel nut
pana	two-dimensional	paper
than	two-dimensional and large	old hand-made paper

Most of the classifiers occur with nouns for natural objects such as plants, fruits and food products. In contrast, nouns for modern objects such as computers and phones can only take the general classifier, e.g., *tin wota kampyutara* (three CLF.GENERAL computer) 'three computers'. The difference in the complexity of semantic categorization for traditional and modern objects thus suggests that the semantics (and the inventory) of classifiers in Nepali is subject to variation, which is probably why the existing literature had divergent observations with regard to the size of the classifier inventory in Nepali.

Sortal classifiers in Nepali typically occur in the context of quantification with the ordering of Numeral-Classifier-Noun, e.g., *tin jana manche* (three CLF.HUMAN man) 'three men', *tin dana syaauu* (three CLF.ROUND.FRUIT apple) 'three apples', and *tin khili cuurot* (three CLF.ARTIFICIAL.CAPSULE cigarette) 'three cigarettes'. However, special attention is required to the following points. First, the general classifier *wota* displays quite a particular morphosyntactic behavior. It occurs with numerals either independently or as fused with a numeral, while other classifiers only occur in an independent form, c.f., *ek wota chora* (one CLF.GENERAL son) 'a son' and *eu-ta chora* (one-CLF.GENERAL son) 'a son'. The general classifier in Nepali also exhibits masculine/feminine agreement in both independent and fused forms (Pokharel, 2010, p. 42-43), c.f., *tin wot-a keto* (three CLF.GENERAL-M boy) 'three boys' and *tin wot-i keti* (three CLF.GENERAL-F girl) 'three girls'. Second, similarly to the areal variation discussed previously with reference to gender, there are analogous forms of variation in the inventory and expression of sortal classifiers. There is thus a tendency to use the masculine form of the general classifier with all non-human animates in colloquial speech in rural areas, e.g., *tin wot-a bhainsi* (three CLF.GENERAL-M female.buffalo) 'three (female) buffaloes' instead of *tin wot-i bhainsi* (three CLF.GENERAL-F female.buffalo) 'three (female) buffaloes'. In addition, there is also variation in the inventory size of sortal classifiers. Our preliminary data shows that more extensive inventories of classifiers are found in the east of Nepal in the area where Tibeto-Burman languages are spoken (Noonan, 2003, p. 75). In contrast, in the west, closer to Indo-Aryan languages, as in central Nepal in Kathmandu, Nepali speakers use fewer classifiers, mainly the general classifier *wota* and the human classifier *jana*.

Further research is necessary to determine whether the variation found within gender systems and sortal classifier systems among speakers from different regions of Nepal is due to an urban vs. rural or east vs. west distribution, since the former would imply differentiation between standard and non-standard/spoken varieties, while the latter could be primarily attributed to language contact.

2.3.3 Functions of gender and classifiers in Nepali

We first turn to a comparison of the functions expressed by gender systems and sortal classifiers. The respective functions are summarized in Table 2.6. With regard to the gender systems, the masculine/feminine and human/non-human systems are represented in the same column since they do not represent overlap of forms (pronouns vs. agreement markers) or semantics (masculine/feminine vs. human/non-human). Furthermore, the human/non-human gender system is only found on pronouns. We thus focus on the interaction between the masculine/feminine gender system and the classifier system.

Table 2.6. *Functions of gender systems and classifiers in Nepali*

Type	Gender systems	Classifiers
Lexical functions		
Expansion of the lexicon	Yes (expression of size among inanimates with masculine/feminine gender markers)	No (classifiers are not affixed to nouns)
Differentiation of referents	Yes (indication of masculine/feminine among animates)	Yes (expression of size and shape among inanimates)
Individuation	No (lack of uses of gender to distinguish between individuated and non-individuated senses of nouns)	Yes (classifiers individuate all nouns for the purpose of quantification)
Ascription of properties	Yes (expression of the speaker's attitude towards animates by gender shift between masculine and feminine)	Yes (expression of degrees of respect towards animate referents by classifier choice)
Discourse functions		
Reference identification	Yes (use of the masculine/feminine and human/non-human gender systems to identify and disambiguate referents)	Yes (use of classifiers to identify and disambiguate among inanimate referents, and of the general classifier among both animate and inanimate referents)
Reference management	Restricted (borderline examples among non-human animates)	Yes (use of the presence or choice of a general/specific classifier to signal the discourse status of the referent)
Re-presentation	Yes (use of the masculine/feminine gender to indicate a change in the speaker's perspective towards an animate referent) sortal classifiers	Yes (use of a different classifier to indicate a change in the speaker's perspective towards an animate or inanimate referent)

A complementary distribution of functions between the masculine/feminine gender and sortal classifiers occurs in the case of two lexical functions and one discourse function. First, in lexical expansion, gender is used to express size differences among inanimates while sortal classifiers are not applicable for this function since they can convey size but cannot be used to create new lexical items. At the same time, such uses of gender are untypical since they occur among inanimates rather than animates as in the other functions and the realization of agreement depends on style. Second, in individuation, classifiers individuate both animate and inanimate nouns for the purpose of quantification, while no examples have been attested of the use of gender to indicate degrees of individuation. And third, with regard to reference management, the presence of a classifier or the choice between a general and specific classifier can be used to signal the discourse status of the referent, while only borderline examples have been found of related uses of gender, where the grammatical gender of nouns for feminine non-human animates depends on their prominence in discourse.

In contrast, examples of functional overlap in terms of the classified nouns are provided by the lexical functions of differentiating referents and attributing properties as well as the two discourse functions of reference identification and re-presentation. However, in these cases we deal with two types of differences with respect to the types of classified nouns and the meanings that are expressed. First, with regard to the type of classified nouns, in differentiating referents, gender markers are used to indicate sex among animates while sortal classifiers are used to indicate size and shape distinctions among inanimates. Analogously, in reference identification, both the masculine/feminine and human/non-human gender systems can be used to introduce and identify referents, with the former further being used to disambiguate among animate referents. In contrast, sortal classifiers are used for reference identification predominantly among inanimate referents, with the masculine and feminine forms of the general classifier also used for animate referents. Second, with regard to the expressed meanings, even though both the masculine/feminine gender and classifiers are used to indicate the speaker's attitude towards an animate referent, they express different meanings, i.e., affection vs. respect. Likewise, in re-representation of referents, the choice of both a different gender, i.e., masculine vs. feminine, and a different classifier can be used to indicate a change in the speaker's perspective, with the use of gender restricted to animates and classifiers used both with animates and inanimates. However, gender and classifiers express different meanings regarding animate referents, i.e., affection vs. respect, analogously to the expression of affective meanings mentioned above. Therefore, these examples do not qualify as functional overlap in the narrow sense, which leads us to conclude that also in these cases we deal with cases of functional differentiation.

2.4 Summary

The comparison between prototypical gender/classifier languages (Marathi and Assamese) and a language with a mixed system (Nepali) provides new evidence regarding morphosyntactic and functional properties of complex systems of nominal classification. In the first place, we contribute to ongoing research by presenting a case study that has been regarded as rare and untypical within not only traditional typologies (Dixon, 1982a,b) but also more recent approaches (Fedden and Corbett, 2017). Further, a functional analysis of concurrent nominal classification systems allows us to evaluate more general principles of functionality that have been proposed with regard to grammatical categories. The complementary distribution of gender vs. classifiers in languages of the world can be interpreted in terms of a complexity trade-off. Since both types of systems have related functions, the fact that they are rarely combined in the same language can be explained in terms of economy and distinctiveness as avoidance of multiple patterns in the same functional domain (Zipf, 1949; Hawkins, 2004). What we have examined with Nepali is a situation where two gender systems and a sortal classifier system exceptionally combine in a single language. Still, even in this case the co-occurrence of the three systems can be accounted for in functional terms. Even though such a situation results in apparent redundancy, an analysis of lexical and discourse functions shows that the three systems actually combine in a way that obeys the principles of economy and distinctiveness. Both gender systems and classifiers in Nepali are used to categorize all nouns, but they still have a largely complementary functional distribution, where a) a function may be expressed by only one system; b) gender systems and classifiers may be functionally exploited with different types of nouns, e.g., animate vs. inanimate; and c) both gender systems and classifiers may be functionally exploited for the same function in the same category of nouns; in which case, they convey different meanings.

3. Quantitative methods in linguistics

Each paper of this chapter investigates linguistic hypotheses via computational methods and starts with a brief overview of the theoretical background for the research question. Then, the source of data is summarized along with a short explanation of the related computational method. Afterward, the output of the experiment is listed as an individual section and summarized in the conclusion.

The first paper shows how the algorithm of random forests can provide additional insights to probabilistic universals in linguistics. The second paper demonstrates how phylogenetic inferences can be helpful to infer the evolution of grammatical features within a language family. Finally, the third paper investigates the automatic recognition of grammatical gender within a language by combining methods from computational linguistics and general linguistics. The three linguistic hypotheses have been selected to represent research questions at different scales: Languages worldwide, language family internal, and within an individual language.

3.1 Probabilistic universals of sortal classifiers worldwide (IV, V, VI, VII)

We investigate the distribution of sortal classifiers in languages of the world by applying existing probabilistic universals in the computational classifier algorithm of random forests. Previous studies demonstrated that the structure of numeral systems and morphosyntactic plural markers have individually strong predictive power with regard to the usage of sortal classifiers in language. We use these two factors as explanatory variables to train the computational classifier of random forests and evaluate the accuracy of their predictive power when selecting the existence/absence of sortal classifiers as response variable.

3.1.1 Classifiers, plural markers, and multiplicative bases

Several theoretical approaches have been proposed to explain the distribution of sortal classifiers within languages of the world (Greenberg, 1990a,b; Chierchia, 1998; Borer, 2005). The connection between sortal classifiers and multiplication originates from an observation on word order. In an enumerative construction composed of numeral, classifier, and noun; the noun is never cross-linguistically attested to intervene between the numeral and the classifier

(Aikhenvald, 2000, p.104-105; Greenberg, 1990b, p.185; Peyraube, 1998; Wu et al., 2006). As an example, constructions such as [NUM CLF N] or [N NUM CLF] are commonly found in languages such as Mandarin Chinese or Thai but no languages show the [CLF N NUM] or [NUM N CLF] patterns. Two recent hypotheses build on previous studies and add mathematical concepts to the discussion (Her and Lai, 2012; Her, 2017; Tang, 2017). Under such view, sortal classifiers are considered to form a multiplicative structure with the numerals and bear the mathematical value of one along with the semantic feature used to highlight the following referent. For instance, the sortal classifier *tiao2* (CLF.LONG) in *san1 tiao2 sheng2zi0* (three CLF.LONG rope) 'three ropes' functions as a multiplicand with the value of one and forms a multiplicative structure with the numeral three, c.f., *san1 tiao2 sheng2zi0* (three CLF.LONG rope) = three times one rope = three ropes[1].

Studies on sortal classifiers also pointed out the complementary-like distribution of sortal classifiers and plural markers, which is commonly referred to as the Greenberg-Sanches-Slobin generalization (Greenberg, 1990b; Sanches and Slobin, 1973). It initially states that if a language uses sortal classifiers in its basic structure of quantitative expressions, then the noun is normally not marked for number in the same structure (Greenberg, 1990b, p.177), since classifiers and plural markers belong to the same syntactic category (Borer, 2005; Her, 2012b). For instance, Mandarin Chinese uses sortal classifiers in quantitative expressions, the nouns following the numeral and the classifier are therefore generally not marked by plural. This generalization involves complementary distribution but not collective exhaustivity (Fromkin et al., 2011), i.e., sortal classifiers and plural markers tend not to occur together; however, it does not imply that either one of the two is always found in languages of the world. By way of illustration, a classifier language commonly lacks plural marking, but languages without plural marking do not necessarily have classifiers (Doetjes, 2012, p.2566). Moreover, the generalization does not forbid the co-occurrence of sortal classifiers and plural markers in the same language; nevertheless, it does predict that if both structures are allowed in the same language, they are not likely to co-occur in the same clause (T'sou, 1976, p.1216) [2].

[1] As a disclaimer, mensural classifiers such as *san1 ping2 shui3* (three MENS.BOTTLE water) 'three bottles of water' in Mandarin Chinese are different from sortal classifiers. This paper only discusses sortal classifiers. For further references on this distinction, please refer to Aikhenvald (2000) and Her (2012a).

[2] Several languages (e.g., Hungarian, Mandarin Chinese, Persian, among others) are found attested with both sortal classifiers and plural markers, but they are generally considered as not real-exceptions due to the optional nature of sortal classifiers and/or plural markers in the targeted languages (Ghomeshi, 2003; Gerner, 2006; Bisang, 2012; Doetjes, 2012). Recent studies suggest that the theoretical definition of sortal classifiers and plural markers is the main explanation to these apparent counter-examples (Tang et al., 2018). On one hand, sortal classifiers should be differentiated from other types of classifiers such as noun classifiers and verbal classifiers (Aikhenvald, 2000; Dixon, 1986; Grinevald, 2015). On the other hand,

Following the combination of these two approaches, the first proposed hypothesis states that the existence of sortal classifiers necessarily implies that the language has a multiplicative numeral system (Her, 2017; Her et al., 2018). In other words, sortal classifiers require the concept of multiplication to form a multiplicative structure; sortal classifiers can thus appear in a language only if a counting system with a multiplicative structure is already present. However, this relation is unidirectional, as the existence of multiplicative numerals does not automatically imply that a language has sortal classifiers. The second hypothesis suggests that morphosyntactic plural markers (e.g., -s in English) are in complementary-like distribution with sortal classifiers since the two elements represent the same formal underlying category (Tang et al., 2018). This functional account unifies plural markers and sortal classifiers as multiplicands that bear the value of one and syntactically mark the countability of nouns. It is thus unlikely to have both plural markers and sortal classifiers in the same language. Should this occur, the two grammatical elements are then expected to be in complementary-like distribution in the noun phrase.

As a summary, two factors have been proposed to predict the distribution of sortal classifiers in language: the absence/occurrence of multiplicative bases and morphosyntactic plural markers. The merge of the two probabilistic universals would result in the following statements: Since sortal classifiers and morphosyntactic plural markers belong to the same category of multiplicand, both of them entail that a language has multiplicative bases. This relation is only unidirectional, the presence of multiplicative bases therefore does not imply that a language necessarily has sortal classifiers and/or morphosyntactic plural markers. Finally, sortal classifiers and morphosyntactic plural markers tend not to co-occur in the same language; if they do, they are expected to not appear within the same structure.

3.1.2 Source of data

The dataset comprises of a sample of 400 languages weighted according to geographical and genealogical factors. For instance, since the Austronesian family accounts for 17.14% (1262/7363) of languages in the world (Lewis et al., 2009; Simons and Fennig, 2018), the same ratio is applied in the dataset (19.00%, 76/400). Likewise for geographical factors: Since the Pacific region accounts for 18.74% (1380/7363) of the languages worldwide, a similar ratio is found in the dataset (18.50%, 74/400). This dataset is not an absolute representative of all 7363 languages of the world, but it is estimated to be sufficient for macro-analyses. A visual representation of the 400 languages is shown in Figure 3.1.

---

only morphosyntactic plural markers (Kibort and Corbett, 2008) should be counted in the generalization, i.e., morphosemantic nominal plural markers such as collective or associative plurals (Rijkhoff, 2000; Vogel and Comrie, 2000) should be excluded.

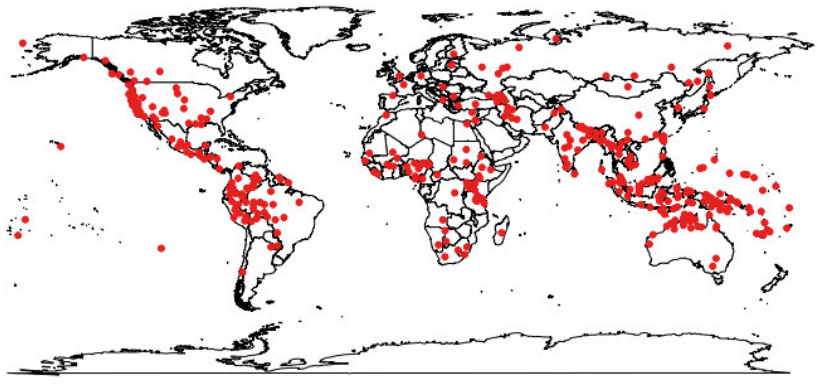


Figure 3.1. Spatial distribution of the 400 languages in the data set

Each language in the dataset is annotated in terms of the features listed in Table 3.1. The features may be divided in the two main categories of grammatical information and metadata. Grammatical information relates to whether the language has morphosyntactic plural markers, multiplicative bases, and sortal classifiers. Metadata refers to the precise location of the language, along with its continent and genus affiliation. The last two features are included to assess the potential areal and genealogical effect on the distribution of sortal classifiers. Both genus and locations are extracted from the World Atlas of Language Structures, whereas the information of continent is based on Ethnologue. The categorical variables of `continent` and `genus` are converted into dummy variables to avoid loss of information. For instance, the categorical variable `continent` is represented by five variables instead of one, c.f., `continent_Africa`, `continent_Americas`, `continent_Asia`, `continent_Europe`, and `continent_Pacific`. Mandarin Chinese is located in Asia and thus has the value of `1` for `continent_Asia` and `0` for the four other dummy variables related to continent.

Table 3.1. *Features encoded in the 400 languages of the dataset*

Feature	Content
`morphosyntactic_plural`	Binary value of presence/absence (yes/no)
`multiplicative_base`	Binary value of presence/absence (yes/no)
`sortal_classifier`	Binary value of presence/absence (yes/no)
`longitude`, `latitude`	Point-coded location of the language from WALS
`continent`	Africa/Americas/Asia/Europe/Pacific
`genus`	Genus classification of the language from WALS

As a general example, French is annotated as `yes` for morphosyntactic plural, `yes` for multiplicative bases, and `no` for sortal classifiers. Morphosyntactic plural since grammatical number is found in French, e.g., *ils sont ici* (they be.PRS.PL here) 'they are here'. As for multiplicative bases, they are

equally present in French, e.g., in *deux cents* (two hundred) 'two hundred' the multiplicand is represented by *cent* 'hundred'; while sortal classifiers are not found in French. With regard to the metadata, French is genealogically affiliated to the `Romance` genus and pinpointed in `continent_Europe` geographically. The annotation of grammatical information is limited in the sense that it is restriction-type features. By way of illustration, the productivity of sortal classifiers is not distinguished cross-linguistically; thus, Chinese with obligatory sortal classifiers has the same value as Hungarian with optional sortal classifiers. Likewise in terms of inventory size and frequency across spoken and written data. Gradient data would probably provide additional insight to the subject (Corbett and Fedden, 2016; Grinevald, 2000) but for the current purpose of investigating the general distribution of grammatical features, this coding is considered sufficient.

3.1.3 The method: Random forests

The algorithm of random forests generates two main outputs: Conditional inference recursive partitioning trees and conditional permutation variable importance. Conditional inference tree is a method of regression and classification based on binary recursive partitioning (Breiman et al., 1984), which is widely used in data mining and machine learning (Chen and Ishwaran, 2012, p.324) and has recently being applied in the field of linguistics (Levshina, 2015; Tagliamonte and Baayen, 2012). As a general method, the data is recursively partitioned in a binary pattern to form homogeneous groups. During this process, the model uses a bootstrap sample of the original data and randomly selects a subset of variables for each split instead of using all variables, so that the variance of the output is maintained as low as possible. The algorithm stops the partitioning process when no variables may split the data with statistical significance. Based on the generated trees, the algorithm can then depict the relative importance of the predictors via conditional permutation-based variable importance, i.e., it allows us to rank the individual importance of variables. This ranking is obtained via random permutation in the out-of-bag data of the tree, from which the estimate of prediction error is calculated. The importance of a variable is thus the average difference between the estimate and the out-of-bag error without permutation. The larger the importance of a variable, the more predictive it is. As a summary, inference trees show how the variables interact with each other and their statistical significance within the data, whereas the importance of variable displays their relative ranking in terms of influencing power.

The main advantage of random forests is the use of permutations when retrieving p-values. The labels of data points are reshuffled randomly and the statistical test is applied for each shuffled data. The result is statistically significant if the proportion of the permutations providing a test statistic greater than

or equal to the one observed in the original data is smaller than the significance level. This methodology can handle data with small quantity of observations and large number of possibly correlated variables, which usually represents a difficulty for conventional statistical tests (Tagliamonte and Baayen, 2012). Moreover, recursive partitioning can bypass several distributional assumptions and handle more easily the presence of outliers (Levshina, 2015, p.292).

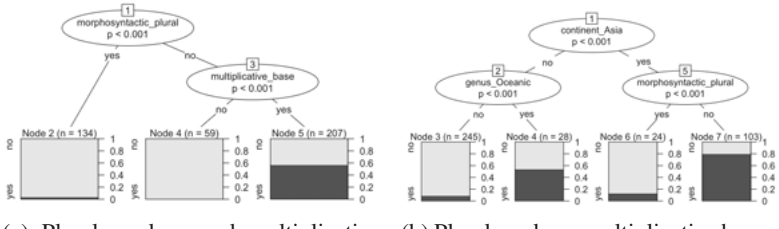
The output of random forests can be evaluated by three methods: The index of concordance C, the Rand index, and the f-score. The *index of concordance C* is a generalization of the area under the receiver-operating characteristic curve (Harrell, 2001). It quantifies how the model discriminates the values of the response variable. The C-index ranges between 0 and 1, a value equals to 0.5 shows a by-chance classification performance, whereas a value above 0.7 represents acceptable performance and above 0.8 indicates a good performance. The *Rand index* commonly generates similar output with the C-index and refers to the overall predictive accuracy of the model and is calculated by dividing the total number of correctly retrieved tokens by the total number of retrieved tokens (Rand, 1971). Then, the detailed performance are investigated category-internally to assess if one of the value of the response variable represented more difficulties for the classifiers, e.g., were classifier languages easier to identify than non-classifier languages. The two values of *precision* and *recall* are thus generated. Precision evaluates how many tokens are correct among all the output of the classifier, whereas recall quantifies how many tokens are correctly retrieved among all the expected correct output. The two measures assess different facets of the output, and are then combined into the f-score, which is equal to the harmonic mean of the precision and recall, i.e., 2(recall*precision)/(recall+precision) (Ting, 2010). Finally, in case the quantity of classifier and non-classifier languages is unbalanced within the dataset, we can use the rule of majority label prediction (*Zero rule*) as a benchmark of accuracy. As an example with our data, since more non-classifier languages than classifier languages are attested in the dataset (69.75%, 279/400), the computational classifier could reach a prediction precision of 69.75% just by labeling all the 400 languages as non-classifier languages. We thus expect that the use of morphosyntactic plural markers and multiplicative bases as explanatory variables should at least exceed the accuracy of 69.75%.

3.1.4 Results

The calculations are realized via the packages `randomForestExplainer`, `rms`, `randomForest`, and `party` (Harrell, 2015; Hothorn et al., 2006; Liaw and Wiener, 2002; Paluszynska, 2017) from R (R-Core-Team, 2018). First, in order to clarify the complex interaction of the predictors evaluated by the random forests, we tested the statistical model of conditional inference tree with sortal classifiers as response variable and the parameters of numeral bases

plus morphosyntactic plurals as explanatory variables. Then, we added the geographical and genealogical factors as explanatory variables to investigate their interactive and individual effect on the prediction of sortal classifiers in language. Finally, we extracted the importance of each variable from the random forests.

Figure 3.2 displays the conditional inference trees obtained via Monte Carlo simulations. The variables that are statistically significant are listed in the upper nodes, which are able to divide the data into several buckets. The buckets are colored according to the ratio of classifier languages. For instance in Figure 3.1.4, Node 4 does not contain classifier languages and is thus in gray, whereas Node 5 contains approximately 60% of classifier languages colored in black. The Figure shows that if a language does not have morphosyntactic plural (Node 1 to Node 3) and does have multiplicative bases (Node 3 to Node 5), it is statistically highly significant ($p < 0.001$) that it is going to have classifiers. In other cases, it is unlikely to have classifiers (e.g., if the language has morphosyntactic plural, or if the language does not have morphosyntactic plural but does not have base). However, when we include the geographical and genealogical factors (Figure 3.1.4), we see a strong geographical effect as the `continent` factor is located at the top of the root. In other words, the model can identify the majority of the classifier languages just by selecting languages located in Asia. For languages found in Asia, the interaction observed in Figure 3.1.4 still holds as languages with morphosyntactic plurals tend not to have sortal classifiers ($p<0.001$). However, for languages not affiliated to the `Asia` region, the effect of genus seems to be stronger than the effect of morphosyntactic plurals. Most classifier languages outside of Asia are mostly found in the `Oceanic` genus (i.e., the Austronesian language family), the conditional inference tree thus displays that this feature is by itself sufficient to identify classifier languages outside of Asia with high precision. Finally, the variable of multiplicative bases is not shown in the tree, which means that its predictive power is weakened when we take into account geographical and genealogical factors.



(a) Plural markers and multiplicative bases as explanatory variables.

(b) Plural markers, multiplicative bases, continent, and genus as explanatory variables.

Figure 3.2. Conditional inference tree with sortal classifiers as response variable.

The C-statistics (0.82/0.85) and Rand indexes (76.5/85.5) of both models reach excellent discrimination and show higher accuracy than the Zero rule (69.75%). An improvement in the predictive power of the model is found when geographical and genealogical factors are included. Likewise for Precision and Recall (Table 3.2), we observe a major improvement in the recall of non-classifier languages (67.7 to 87.7) and the precision toward identifying classifier languages (56.5 to 74.1) when including geographical and genealogical factors.

Table 3.2. *Precision and recall from the conditional inference trees.*

	Without continent and genus		With continent and genus	
	no classifiers	with classifiers	no classifiers	with classifiers
Recall	67.7%	96.7%	87.7%	80.1%
Precision	97.9%	56.5%	91.1%	74.1%
F-score	80.1%	71.3%	89.4%	76.9%

The analysis by conditional inference tree in Figure 3.2 showed the most relevant variables when considering the interaction of all the variables. Nevertheless, we still need to investigate the individual importance of each variable, i.e., a variable could have a strong effect but not be shown on the conditional inference tree due to a slight difference of predictive power with the listed variables or a weakened effect when interacting with other variables. The predictors include the features listed in Table 3.1, i.e., `morphosyntactic_plural`, `multiplicative_base`, `continent`, and `genus`. Figure 3.3 shows the frequency of minimal depth for each variable across all the trees generated by the random forests and its mean.

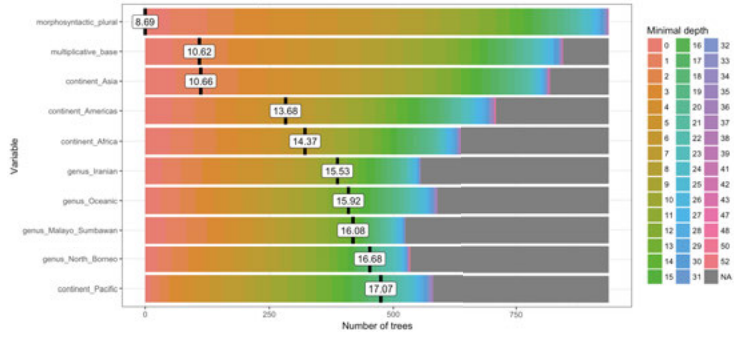


Figure 3.3. Distribution of the ten variables with the smallest mean minimal depth

The minimal depth refers to how far is the node with the variable from the root node. A small value indicates that the variable is frequently represented as the root node (or a top node in the tree) and is thus more important. We only list here the ten variables with the smallest mean minimal depth. `Morphosyntactic_plural` is by far the most important variable, followed

by multiplicative_base and continent_Asia. Some predictivity is detectable for other geographical and genealogical factors, but their minimal depth is relatively bigger than the top three variables. We may therefore infer that even though multiplicative bases are not showing on the conditional inference tree of Figure 3.2, the variable still plays a significant role in the distribution of sortal classifiers in languages of the world, whereas the areal effect of 'classifiers in Asia' is once more observed.

This observation is equally attested in different measures. Figure 3.4 shows the importance of variables sorted according to their effect on the accuracy and purity of nodes. The mean decrease of accuracy refers to how worse the model performs without each variable; a high decrease indicates that the variable has a strong predictive power. The mean decrease of the Gini coefficient shows how each variable contributes to the homogeneity of the nodes and the end of the tree, i.e., can this variable contribute to clearly-separated buckets. Again, a high decrease of Gini coefficient when removing a variable indicates that this variable has a strong predictive power and therefore a high importance. In both measures, the variables morphosyntactic_plural, multiplicative_base, and continent_Asia are consistently at the top, which further supports our observations in Figure 3.3. Moreover, all measures also show that the variable of morphosyntactic plurals is stronger in terms of predictive power than the variable of multiplicative bases.

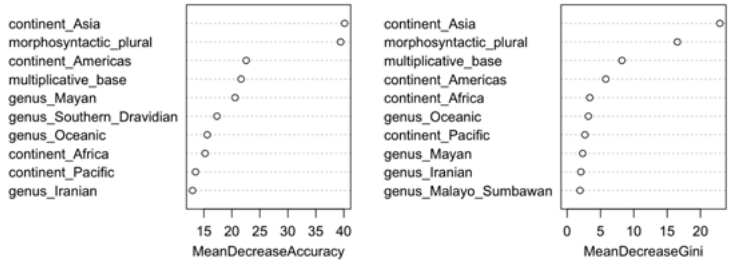


Figure 3.4. Importance of the variables with sortal classifiers as response variable and morphosyntactic plural markers, multiplicative bases, continent, and genus as explanatory variables

Finally, an overview of the importance of variables is displayed in Figure 3.5. The x-axis represents the mean minimal depth of each variable, the y-axis points out the frequency that a variable is used to split the root node, and the size of the bubbles indicates the total number of nodes that use the variable for splitting. The top ten important variables are labeled and highlighted in blue. The three variables being used the most as root nodes and being included the most frequently across all the generated trees are still morphosyntactic_plural, multiplicative_base, and continent_Asia.

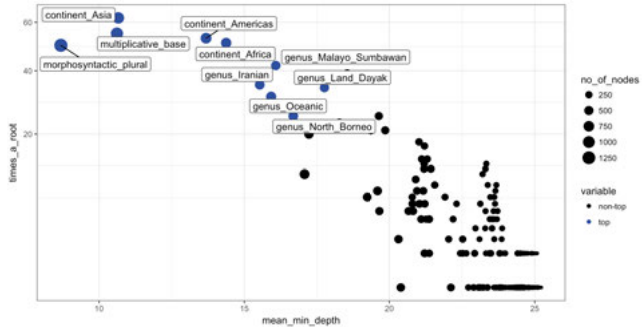


Figure 3.5. Multi-way importance plot of the variables

As a summary, the variables of morphosyntactic plurals and multiplicative bases can predict the occurrence/absence of sortal classifiers in language with high precision. Among these two variables, morphosyntactic plurals shows stronger predictive power than multiplicative bases. Adding geographical and genealogical factors as variables improves the performance of the model and demonstrates that sortal classifiers are subject to a strong areal affect as most classifier languages are found in Asia, whereas the genealogical effect is of a minor nature.

3.1.5 Conclusion

In this study we demonstrated how computational methods could be applied to linguistic hypotheses. Specifically, the model of random forests was able to reveal the interaction pattern of linguistic variables along with their individual importance under various measures. Such a methodology allows a multi-faceted approach of linguistic theories and provides a ranking of variables in terms of importance rather than an arbitrary clear-cut division. Our results are partially consistent with existing linguistic hypotheses as multiplicative bases and morphosyntactic plural markers have a strong predictive power with regard to the absence/occurrence of sortal classifiers in a language, even when taking into account geographical and genealogical effects. However, the correlation between morphosyntactic plurals and multiplicative bases is not as strong as theoretically expected. Moreover, our results may relate to more than one linguistic theory that can explain the correlation patterns identified in this study. As an example, it applies equally to the Greenberg-Sanches-Slobin generalization or the count-mass hypothesis (Chierchia, 1998). Further features are thus required to investigate the individual predictive power of each theory.

3.2 Phylogenetic signals of grammatical gender in Indo-Aryan (VIII)

We investigate the evolution of grammatical gender in Indo-Aryan languages using phylogenetic comparative methods. 44 presence-absence features pertaining to grammatical gender have been compiled for 48 Indo-Aryan languages. The grammatical gender features relate to gender marking on the verbs, adjectives, personal pronouns, demonstrative pronouns, and possessive pronouns. The results of our Bayesian Reverse Jump Hyper Prior analysis, which infers the evolutionary dynamics of changes between feature values, are consistent with historical linguistic and typological studies on gender systems in Indo-Aryan languages. Finally, we are able to demonstrate via a conditional inference tree how the main sub-groups of the Indo-Aryan family are distinguishable from their characteristic configurations of grammatical gender features. An analysis using Mantel correlograms shows a possible effect of language contact on the typology of grammatical gender systems in Indo-Aryan languages.

3.2.1 The canonicity of nominal classification systems

The recent trend in the field is to avoid considering grammatical gender as a binary feature of absence/existence in a language, but rather treating it as a cline (Grinevald, 2015), or as set of sub-features which together approach more or less to the prototype (As in the Canonical Typology approach, Corbett and Fedden, 2016). For instance, French and German may be easily annotated as having grammatical gender in contrast to languages such as Chinese that use other systems of nominal classification; however, languages which only have a pronominal gender system (e.g., English) may be included as having gender as well, if the definition of a gender language is solely the existence of gender marking in a language. Yet, stating overtly that the grammatical gender system in French and English are identical is not reflecting accurately the state of the two languages. Another example would be the comparison between Hindi and Swedish. Hindi distinguishes masculine and feminine whereas Swedish adopts the uter/neuter categories. In Hindi, gender is marked on the verbs and adjectives, while Swedish marks gender on articles and adjectives, but not on verbs. In Hindi, only certain adjectives are marked with gender while in Swedish the majority of adjectives bear gender marking. Both languages would receive the same label if only general features such as absence/existence or number of gender were considered, even though their grammatical structure is fundamentally different. Our goal is thus to demonstrate how quantitative methods may be used to measure the overt complexity of grammatical gender systems and provide a concrete basis for comparison across languages (Bisang, 2014; Corbett and Fedden, 2016). We select the Indo-Aryan language family as a case study and gather data on morphosyntactic gender features in

each language. We then apply Bayesian phylogenetic methods to investigate the historical processes leading to the current typology of grammatical gender systems in Indo-Aryan, as well as identifying which languages and gender features have been influenced by non-phylogenetic factors. This methodology allows us to capture a detailed representation of gender systems in every individual language, rather than to treat the entire gender system as a single category.

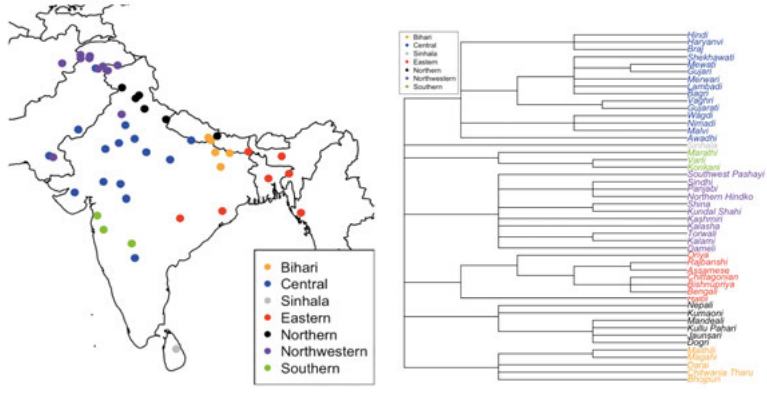
We select the Indo-Aryan language family as a case study and gather data on morphosyntactic gender features in each language. Then, we apply Bayesian phylogenetic methods to investigate the current clustering and historical evolution of grammatical gender systems in Indo-Aryan. Furthermore, we equally pinpoint which languages and gender features are representing noise in our current data and require additional definition. Such methodology allows us to capture a detailed representation of the gender systems in each individual language, rather than to treat the entire gender system as a single category.

The main contributions of this paper are as follows: First, the recent trend of studies on grammatical gender suggests a gradient approach that considers grammatical gender as a continuum is better than treating it as a binary linguistic feature (Grinevald, 2000; Corbett and Fedden, 2016). However, no quantitative study has yet applied such approach due to the lack of available sample. Our study therefore provides quantitative data to investigate the efficiency of the canonical approach on a larger scale. Second, phylogenetic comparative methods (Dunn, 2015) can be used to infer patterns and test hypotheses about the dynamics of evolutionary change within a family (Levinson et al., 2011; Dunn et al., 2011). In this study we investigate the diachrony of gender systems in Indo-Aryan languages, showing how typological profiles change over time. Third, we use partial Mantel correlograms to measure possible geographic factors contributing to the typological distribution of gender systems, and we use conditional inference trees to examine which particular features of gender systems are most relevant for establishing a subgroup-level typological profile.

3.2.2 Source of Data

Our data contains a weighted subset of the 216 languages found in the Indo-Aryan language group (Hammarstrom et al., 2018). As commonly performed in previous studies (Dunn et al., 2013; Kolipakam et al., 2018), languages from a sufficient number of varieties were selected to quantitatively represent all the previously reported major subgroups of Indo-Aryan. For instance, the Bihari sub-group accounts for nearly 11% (23/216) of all Indo-Aryan languages. The Bihari languages we have chosen thus equally represents 11% (5/48) of our dataset. The same logic applies for other major groups and sub-groups of Indo-Aryan languages. However, we excluded certain languages due to their

status of outliers in terms of geographical distance with other languages. As an example, we did not include the Romani languages, Lomraven, Fiji Hindi, Carribean Hindustani, and Andaman Creole Hindi. The result of this filter is summarized in the spatial and phylogenetic overview shown in Figure 3.6. The main Indo-Aryan sub-groups are Bihari, Dhivehi-Sinhala, Indo-Aryan Central zone, Indo-Aryan Eastern zone, Indo-Aryan Northern zone, Indo-Aryan Northwestern zone, and Indo-Aryan Southern zone.



(a) Spatial distribution (b) Glottolog's phylogenetic tree

Figure 3.6. Overview of the subset of 48 Indo-Aryan languages.

Each of the 48 languages were annotated according to 44 binary features on grammatical gender and three features of metadata extracted from Glottolog (longitude, latitude, language family). The gender features were extracted from language grammars and consultation with linguists working on the language. These features may be divided into three categories: type of gender, gender marking on the verb, and gender marking on adjectives and pronouns. The first category contains four features that indicate if the language distinguishes between biological gender, neuter gender, animacy, and/or humanness. Each feature may carry the value of `1` (yes) or `0` (no). For instance, the grammatical gender system of Hindi (Indo-Aryan Central zone) is solely based on biological gender (masculine/feminine) (Agnihotri, 2007). Hindi is thus annotated as `1` for biological gender and `0` for neuter gender, animacy, and humanness. The second category includes 16 features with regard to gender marking on the verb. These features are generated by the parameters of tense (past/ present/ future), number (singular/ plural), argument (subject/ object), and gender type (biological/ neuter/ animacy, humanness). By way of illustration, four features relate to the present tense: Does the verb mark gender (regardless of which type) on the present singular? does the verb mark gender on the present plural? does the mark the gender of the subject in the present tense? does the verb mark the gender of the object in the present tense? The same logic applies for the past and future tense. For instance,

Assamese (Indo-Aryan Eastern zone) is annotated as `0` for all four features since the language does not mark gender on the verb (Kalita, 2003). The third category of 24 features relates to gender marking on the adjectives and pronouns (possessive, demonstrative, personal). Each of the four grammatical category is labeled according to number (singular/ plural) and gender type (biological, neuter, animacy, humanness). As an example, the six features of personal pronouns are: Do the personal pronouns mark gender in singular? do the personal pronouns mark gender in plural? do the personal pronouns mark biological gender (regardless of singular or plural)? do the personal pronouns mark neuter gender? do the personal pronouns mark gender based on animacy? do the personal pronouns mark gender based on humanness? As an example, personal pronouns in Sinhala (Dhivehi-Sinhala) mark humanness, animacy and biological gender in both singular and plural (Gair, 2007, p.783). The value of the features for personal pronouns are thus `1 1 1 0 1 1`.

As shown in Figure 3.7, most Indo-Aryan languages mark biological gender, whereas rare cases are attested for neuter, animacy, or humanness. Verbs and adjectives are the most commonly marked syntactic categories, while only near half of the languages mark gender on pronouns. Interestingly, most cases of non-biological gender marking are attested on the demonstrative and personal pronouns. This symmetry is expected as third person pronouns and demonstrative pronouns are generally used interchangeably in most Indo-Aryan languages. The diversity of gender marking on pronouns is also supported by the agreement hierarchy as personal pronouns are the most likely to be influenced by grammatical agreement with semantic justification (Corbett, 1979, 2012).

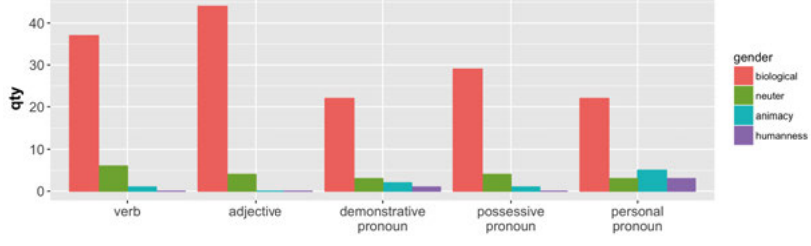
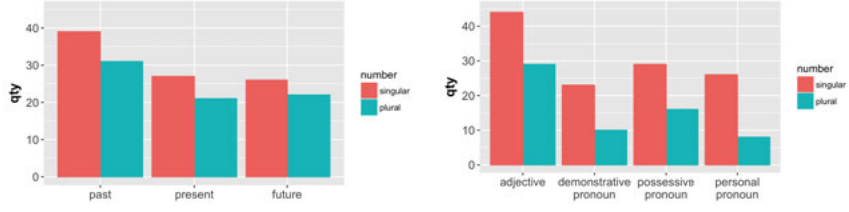


Figure 3.7. The distribution of gender type across syntactic categories

With regard to the interaction of gender and number. Almost half of the languages in our dataset mark gender on past, present, and future tense in both singular and plural (Figure 3.2.2). The past tense tends to be marked more consistently than present or future, and gender is slightly less marked in plural than in singular. Such observation is not typologically rare, as gender is frequently neutralized in the plural number (Corbett, 1991, p.155). Within the other syntactic categories, adjective is frequently marking gender, whereas it is less common in the three types of pronouns analyzed in our study (Figure

3.2.2). We equally observe the tendency that gender is neutralized in plural number. However, this tendency is especially salient in personal pronouns and demonstrative pronouns. As for adjectives and verbs, we also note that the feminine gender tend not to show gender distinction. As an example in Wagdi (Indo-Aryan Central zone), demonstrative pronouns have different forms for masculine singular *pelo* and masculine plural *pela* but use the same form for feminine singular and plural *peli*. Likewise for verb agreement, e.g., "Wagdi perfective suffixes are phonologically identical to those of Marwari and Gujarati, and can be generalized as follows: *-o* is the marker of masculine singular, *-a* of masculine plural; and *-i* of feminine singular and plural" (Phillips, 2012, p.70,89).



(a) Verb (b) Other syntactic categories

Figure 3.8. Overview of the gender-number interaction across syntactic categories.

As a summary, our data shows the tendency of Indo-Aryan languages to commonly mark biological gender. Moreover, verbs and adjectives are more frequently marking gender compared to pronouns. Finally, gender neutralization in the plural number occurs more frequently for pronouns and adjectives than for verbs. These observations correlates with previous studies by providing concrete quantitative data to the discussion.

3.2.3 The method: Phylogenetic inferences

Phylogenetic comparative techniques from evolutionary biology have recently started to be applied to language data, since these novel methods are able to bypass the difficulties caused by the non-independence of features produced by evolutionary processes (Galton's problem Mace and Holden, 2005). In such analyses a tree (itself produced by Bayesian phylogenetic inference) provides the context from inferring the evolutionary processes undergone by the target features.

In the present analysis we use Bayesian Monte Carlo Markov Chain (MCMC) phylogenetic inference to generate a sample of trees that are consistent with genealogical constraints derived from the Glottolog reference phylogeny. In technical terms, we are inferring a tree sample directly from the (highly specified) topological prior. This gives us tree which is both consistent with the

presumed historical linguistic tree, and which also has some estimate of uncertainty where the glottolog tree is unresolved.

Based on the sample of trees generated from the MCMC, we can extract a maximum clade credibility tree that reflects best the clustering of languages included in our dataset. To generate such tree, the algorithm goes through the tree set and assess the likelihood of each binary split based on how many times this split occurs in the data. Then, the product of the likelihoods of each split in the tree is calculated and each tree is assigned a score based on this product. The tree with the highest score is selected as the maximum clade credibility tree and considered as the most representative tree in the sample. The maximum clade credibility tree can then be compared with the current tree obtained from historical linguistics and their divergence may be scrutinized.

Finally we carried out an MCMC analysis using Reverse Jump Hyperprior (RJHP) to investigate the evolution of our typological feature values mapped onto the phylogenetic tree. This gives us the probability of change between the different values of each feature and allows us to infer evolutionary patterns. In other words, based on the currently observed grammatical gender features in our dataset, we may infer how likely and in which direction these features are subject to change diachronically.

3.2.4 Results

Within this section, we first measure the phylogenetic signal for each gender feature present in our dataset. Then, we use RJHP to assess how the gender features with a strong phylogenetic signal in Indo-Aryan are more likely to have evolved across time. The Bayesian phylogenetic inference was carried out using MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003), and the results are evaluated with Tracer (Rambaut et al., 2018), while the phylogenetic comparative analyses are realized via the `phangorn` (Schliep, 2011), `phytools` (Revell, 2012), and `ape` (Popescu et al., 2012) packages in R (R-Core-Team, 2018). Our MCMC run based on the constraints from the Glottolog tree generated a set of 3000000 generations. We removed 1500000 trees as burn-in, and extracted one tree every 1500 trees, which resulted in our final sample of 1000 trees (ACT = 1500, Effective Sample Size = 1000). To visualize how the sample of trees matches with the Glottolog tree, every tree of our sample is scored by the product of the likelihood of the splits observed in each individual tree. The tree with the highest score is selected as the maximum credibility (MCC) tree that represents the best the data. Figure 3.9 displays the Glottolog tree and the MCC tree of our sample plotted together. The parallel lines between the two trees show that the main language groups are represented correctly, e.g., Varli, Marathi, and Konkani from the Southern sub-group are found in similar clusters in both the Glottolog tree and the MCC tree.

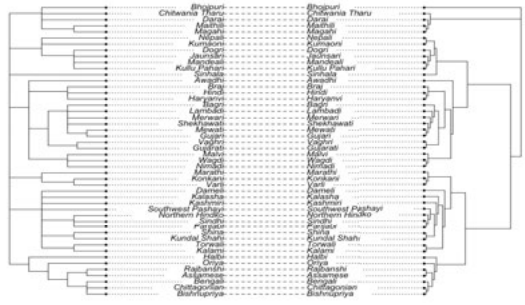


Figure 3.9. The Indo-Aryan tree from Glottolog (left) and the maximum credibility tree from our subset of 48 Indo-Aryan languages (right) plotted as facing phylogenies with links between tips

The phylogenetic signal of gender features is evaluated first to provide a statistical measure of how well do individual gender features match with the language phylogeny proposed by historical linguistics. By way of visualization, Figure 3.10 displays the Glottolog tree of the 48 Indo-Aryan languages with tip labels annotated according to the gender features encoded in our data.

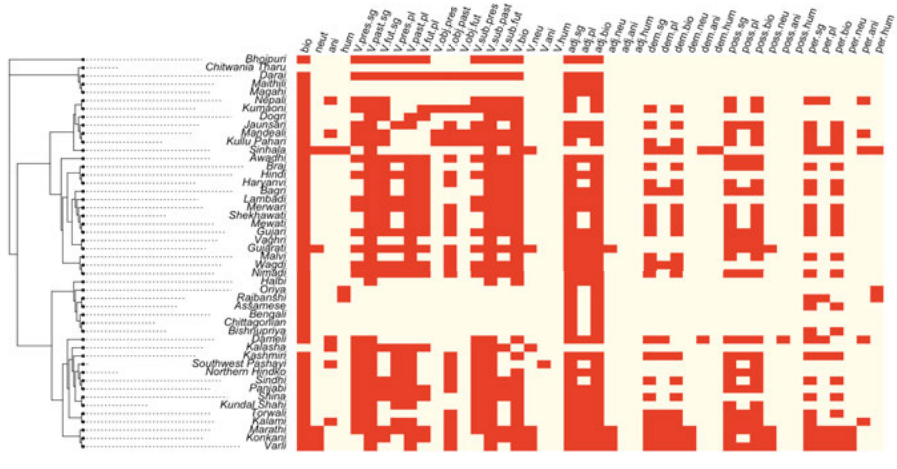
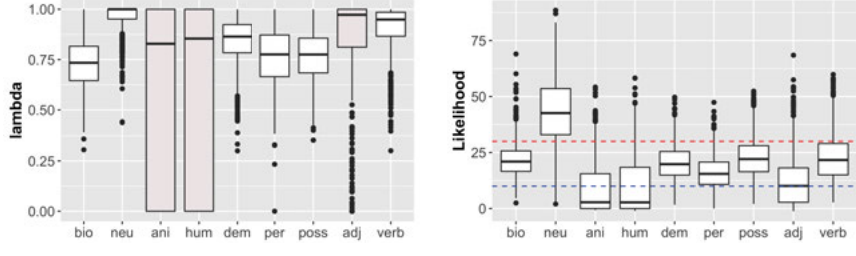


Figure 3.10. Heatmap visualization for the phylogenetic signal of gender features within the 48 Indo-Aryan languages. Red indicates presence of the feature and yellow refers to absence of the feature.

A phylogenetic signal is considered to be strong if languages with the same value of features are close to each other genealogically. For example, languages commonly mark gender on the verb, with the exception of a few specific subgroups such as Eastern Indo-Aryan. These sub-groups without gender marking can be easily identified as the yellow areas in the heatmap within the columns related to verb (`V.pres.sg` to `V.hum`). We thus except a strong phylogenetic signal for gender on verbs. On the other hand, gender marking on

personal pronouns does not show regular patterns, as languages for the same subgroup do not tend to share the same value (column `per.sg` to `per.hum`). We therefore except a weak phylogenetic signal for gender on adjectives.

We conduct the hypothesis test for statistical significance according to Pagel's lambda (Pagel, 1999; Blomberg et al., 2003; Ives et al., 2007) to assess statistically the phylogenetic signal of the 44 gender features. This method proposes that assuming a pure Brownian model of evolution, the phylogenetic relation between species matches with the expected covariance matrix of their traits. In our case, the phylogenetic relation between languages matches with the variation of their gender features. However, when non-phylogenetic factors have an impact on trait evolution, the estimated influence of the phylogeny is down-weighted. The lambda coefficient represents this weight by incarnating the transformation of the phylogeny with the best fit of trait data to a Brownian model of evolution. Lambda varies between zero and one, the closer to zero thus indicates phylogenetic independence while a lambda close to one (or slightly above one[3]) shows strong phylogenetic signal (Freckleton et al., 2002). Based on the lambda of each feature, its likelihood ratio may be calculated by doubling the difference between the the log-likelihood of the optimal value of lambda (LogL) and the log-likelihood if lambda is zero (LogL0), i.e., 2*(LogL-LogL0). Figure 3.11 shows the phylogenetic signal of the main categories of gender features in our dataset. `Biological`, `neuter`, `animacy`, and `humanness` indicate the type of gender used by the language, e.g., how strong is the phylogenetic signal of using biological gender within the 48 Indo-Aryan languages. `dem`, `per`, and `poss` refer to gender marking on various types of pronouns. Finally, `adj` and `verb` show the phylogenetic signal on adjectives and verbs.



(a) Range of Lambda and P-value (b) Likelihood ratio

Figure 3.11. Visualization of Lambda and Likelihood ratio of the phylogenetic signal of gender within the 48 Indo-Aryan languages. The gray box plots in (3.11a) show a p-value higher than 0.05. The blue and red lines in (3.11b) indicate strong and very strong likelihood ratio.

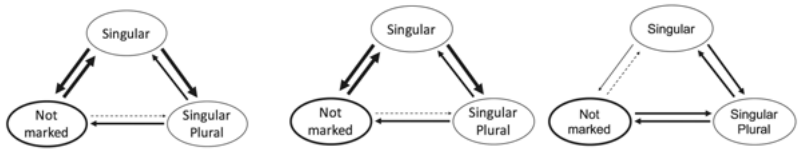
[3]Pagel's lambda can exceed one if the similarity of the analyzed traits exceed the expectations of the Brownian model of evolution

The lambda value of animacy-based gender, humanness-based gender, and gender marking on adjectives fluctuate much more than other the other grammatical categories (Figure 3.11a). This variation is mirrored as we cannot reject the null hypothesis of no association for these three categories. The likelihood ratio in Figure 3.11b is conventionally interpreted as follows: <3 infers weak evidence, 3-10 shows moderate evidence, 10-30 (blue line) means strong evidence, > 30 (red line) refers to very strong evidence (Jeffreys, 1961). For instance, the test indicates a strong evidence of phylogenetic signal for biological gender within the 48 Indo-Aryan languages we surveyed. The mean likelihood ratio of biological gender is 21.590 (median = 20.944), which means that data occurred 21 times more likely under the alternative hypothesis of phylogenetic signal than under the null hypothesis of no phylogenetic signal. As also observed in Figure 3.11a, the influence of the phylogeny is low for the categories of animacy-based gender, humanness-based gender, and gender marking on adjectives.

Our measurements of phylogenetic signal are consistent with the existing linguistic literature. Gender distinction based on humanness and animacy has a weak phylogenetic signal. This is expected since grammatical gender based on humanness is more a Dravidian feature than an innate Indo-Aryan feature (Kilarski, 2014, p.13). As for animacy, Proto-Indo-European had animacy marking and some small traces of that gender system may have remained on the pronouns, which are the most likely to retain gender marking. For instance, Kalami (Indo-Aryan Northwestern zone) only marks animacy on the oblique case of third person singular personal pronouns (Baart, 1999, p.39). On the other hand, the phylogenetic signal of biological and neuter gender is ,as expected, strong since the Proto-Indo-Aryan languages such as Vedic are attested to have masculine, feminine, and neuter gender (MacDonell, 1999; Beekes, 2011). Gender marking on verbs, demonstratives, personal pronouns, and possessives equally shows a strong phylogenetic signal. We thus proceed with the diachronic analysis on these grammatical categories that have a strong phylogenetic signal.

Then, we perform the reverse jump hyperprior (RJHP) that evaluates which patterns of transition are more common within the observed data. For instance, based on how gender is marked today on the demonstratives, we may infer how gender marking on the demonstratives is more likely to have evolved across history. One main advantage of the RJHP is that it determines the amount of heterogeneity from the data by assessing all possible combinations and account for the uncertainty in the amount of heterogeneity while other phylogenetic parameters of interest are evaluated and bypasses the need for complex selection procedures (Green, 1995; Gowri-Shankar and Rattray, 2007). As an example with gender marking on the demonstratives from our data, the RJHP includes the scenarios of reversed change between states, i.e., the model not only scores the probability that the demonstratives lose or acquire gender marking but also evaluates cases such as the demonstratives have loss gender

marking but re-acquired it at a later stage or vice-versa (unlike maximum likelihood method, which would only involve the first type of scenarios). A sample of the results from our MCMC analysis with RJHP are shown in Figure 3.12. The thickness of the arrow is correlated with the transition rate between the three categories of marking, while the category with the biggest letter size indicates the most likely ancestral state. For instance in Figure 3.12a, the most likely ancestral state is without gender marking (N). Then, it is more probable to acquire gender marking in singular (S) and then on plural also (A). However, it is equally possible to experience loss of gender marking from singular (S) and singular/plural (A), whereas acquiring (or re-acquiring) gender marking on the demonstrative from scratch to marked on both singular and plural is unlikely. The transition pattern of demonstrative (Figure 3.12a) and personal (Figure 3.12b) pronouns is synchronized as demonstrative pronouns are commonly used as third person pronouns in Indo-Aryan.



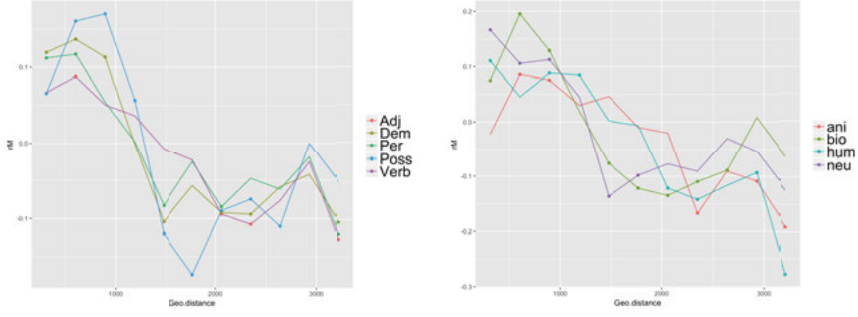
(a) Demonstrative pronouns (b) Personal pronouns (c) Possessive pronouns

Figure 3.12. Summary of transition rates from reverse jump hyper prior with regard to the evolution of gender marking on different types of pronouns in the Indo-Aryan languages from our dataset.

The transition of possessive pronouns (Figure 3.12c) indicates that the most likely ancestral state is also the absence of gender marking. Interestingly, the most probable transition pattern then points to a state in which gender is marked on both singular and plural possessive pronouns, and a possible following loss of gender marking on the plural possessive pronouns. This observation differs from the transition pattern suggested for demonstratives and personal pronouns, i.e., from absence of gender marking to gender marking on the singular and then on the plural. We speculate that such phenomenon is due to the fact that possessive pronouns are typically formed by personal pronouns combined with genitive case markers in Indo-Aryan languages, i.e., personal pronouns do not mark gender (as explained in the paragraph above) while gender marking on the genitive case marker occurred at later stage in Indo-Aryan history, c.f., the genitive marker *-caa* (GEN.MASC.SG) in Old Marathi was derived from focus markers in Old and Middle Indo-Aryan, which later came to mark for gender (Master, 1964; Peterson, 2017). This combination thus automatically results in both singular and plural possessive pronouns marking gender symmetrically since they are formed in the same manner, whereas gender neutralization on the plural may occur at a later stage. For

instance in Nepali, the personal pronouns distinguish animacy but not biological gender, c.f., *u* 'he/she' and *tyo* 'it'. However, possessive pronouns in Nepali are formed by suffixing the genitive case marker to personal pronouns; and the genitive case marker in Nepali marks gender, i.e., *ko* 'GEN.MASC.SG' and *ki* 'GEN.FEM.SG'. The possessive pronouns in Nepali thus mark gender in the singular, e.g., *usko* 'his'. However, gender is neutralized in the plural in Nepali, as the genitive case marker does not mark gender in its plural form *kaa* (Acharya, 1991, p.112-115).

To further verify if language contact has more effect in a specific category of gender features (e.g., adjectives). A Mantel correlogram is performed to evaluate the trend of geographical influence (Legendre et al., 2015). The results are shown in Figure 3.13. The x-axis indicates the geographical distance between languages in kilometers. The y-axis represents the Mantel correlation coefficient, which ranges between -1 and 1. The closer to 0 the weaker the correlation, whereas the closer to 1 or -1 indicates stronger positive or negative correlation. Colored dots refer to classes with statistically significant correlation coefficient ($p<0.05$). The horizontal lines represent the expectation of the Mantel statistic under no spatial autocorrelation.



(a) Grammatical categories (b) Gender types

Figure 3.13. Partial Mantel correlograms representing the correlation coefficient (y-axis) between linguistic distance and geographical distance (x-axis) when controlling for phylogeny.

The correlogram by main categories of features (Figure 3.13b) shows that gender types along with gender marking on adjectives and possessive pronouns are less stable than gender marking on verbs, possessives, and demonstratives. For instance, the correlation coefficient of gender marking on adjectives (blue) is relatively low and fluctuates more compared to other categories. Likewise in terms of probability, while the effect of geographical factors is identified as weak but statistically significant for verbs, personal pronouns, and demonstratives; this is not the case for adjectives and possessive pronouns. Such observation suggests that there is more noise in the data of adjectives and

possessive pronouns. Further analysis is thus required with regard to these specific categories.

Finally, we also investigate the predictive power of the gender features with regard to their current language family association. In other words, we noticed a strong phylogenetic signal of gender features; however, this does not necessarily imply that these features match the individual labels assigned by the Glottolog tree. It is possible that languages with similar gender features are clustered together as the same language sub-group but the detailed branching structure within the cluster may slightly differ from the interpretation of Glottolog's tree, as shown in Figure 3.9. We thus run the statistical test of conditional inference tree via Monte Carlo simulations (Breiman, 2001) to evaluate the predictive power of the 44 gender features with regard to the language sub-groups (Bihari, Central, Eastern, Sinhala, Northern, Northwestern, Southern). Figure 3.14 displays the results of our analysis. The relevant features are labeled in the main three nodes whereas the output of classification is indicated via different buckets (Node 4 to 7) at the bottom of the graph. Different bars refer to the six language sub-groups involved in the analysis. The conditional inference tree can easily identify the Bihari, Eastern, and Southern sub-groups. For instance, if a language marks neuter on the demonstrative, it is very likely to belong to Southern Indo-Aryan, whereas a language not marking neuter on demonstratives and not marking gender on the verb in the past singular or the adjective plural is very likely to belong to Eastern Indo-Aryan.

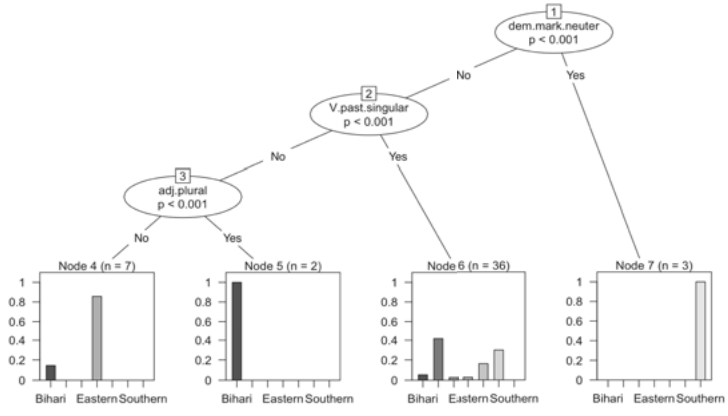


Figure 3.14. Conditional inference tree with language families as response variable and 44 gender features as explanatory variables

Nevertheless, the conditional inference has difficulties distinguishing between Northern, Northwestern and Central Indo-Aryan, as the three sub-groups are merged as one category in node 6. This has an impact on the predictive power of the model as its Rand index (accuracy) is only 54.16%, i.e., if the model was asked to predict the language family of one of the 48 Indo-Aryan languages from our dataset based on the 44 gender features, it could just get

near-half of the families correctly. This observation matches with our previous findings and further show that Northern, Northwestern, and Central Indo-Aryan are less distinguishable in terms of gender features, which is probably due to their small geographical and phylogenetic distance, along with the political landscape of the region in comparison to the Eastern Indo-Aryan languages (Peterson, 2017, p.233).

As a summary, we have investigated in this section the phylogenetic and geographical signals of different gender features. Our analysis shows that while gender features reflect accurately one aspect of the current convergence and divergence of the Indo-Aryan languages in our dataset, they may not perfectly mirror current phylogenetic trees such as the Glottolog tree, which are based on various types of features (e.g., phonology, morphology, syntax, among others). Nevertheless, the main purpose of this paper was to apply a concrete measure-scale for grammatical gender on quantitative data and phylogenetic methods. This aim is considered reached and successful since we were able to show that Bayesian phylogenetic methods could provide interesting results in terms of clustering and generate additional insight to the evolution of gender systems within the Indo-Aryan language family. Finally, we also scrutinized the relative importance of factors encoded in the data via Mantel tests and conditional inference trees, which allowed us to pinpoint which features and languages should be further investigated.

3.2.5 Conclusion

The evolution of grammatical gender in the Indo-Aryan family is relevant to studies of the Indo-European family and to research related to nominal classification in general. The main purpose of this paper was to better understand the evolutionary dynamics of grammatical gender modeled as a fine-grained combination of typological features rather than as a binary trait. Second, the results of phylogenetic comparative methods replicated the existing linguistic knowledge on the diachrony of grammatical gender systems in Indo-Aryan. Third, we also discovered additional transfer patterns for grammatical gender systems in Indo-Aryan, e.g., the gender transfer pattern of verbs. Fourth, the measurement of phylogenetic signal combined with partial Mantel correlograms and conditional inference trees was able to identify which gender features and which language subgroups were more likely to have been influenced by language contact. This was especially relevant due to the location of the Indo-Aryan language family on the crossroads of Indo-European,Sino-Tibetan, Dravidan, and Austro-Asiatic languages, which employ different systems of nominal classification.

3.3 Grammatical gender identification in Swedish (IX, X)

Word embeddings combined with various types of computational classifiers (e.g., neural networks) reflect one (of many) aspect(s) available to language processing in the human mind (Collobert et al., 2011; Mikolov et al., 2013; Pennington et al., 2014). Yet, there is still a need for better understanding of the information contained in word embeddings. We select a linguistically motivated classification of words (grammatical gender in Swedish) as a case study to compare the knowledge provided by linguistic theories and the information encoded into basic statistical structures such as word embeddings. Moreover, we also include analyses of word frequency and errors. Such steps are not commonly observed in computational linguistic methods, but we show that they can provide additional insights to this type of study.

We propose the following research questions: 1) Can word embeddings alone provide sufficient information for neural networks to identify perfectly grammatical gender in Swedish? An intuitive guess would be positive since the gender of a noun should be predictable from its co-occurrence statistics (e.g., neuter nouns tend to co-occur with determiners and adjectives in the neuter inflection). 2) If the classification output still contains errors, what types of errors are made and can we explain these errors from a linguistic perspective? 3) Can the analysis of word frequency provide additional insights to the use of word embeddings?

3.3.1 Grammatical gender in Swedish

Grammatical gender in Swedish is an inherent property of every noun, which is not expressed overtly on the noun unless it combines with other elements and agrees with them. As demonstrated in 4, nouns in Swedish are divided into neuter and uter (common). The two categories are thus reflected on the determiners and adjectives respectively.

(4) Gender agreement in Swedish (Indo-European)

a. *Ett* *stor-t* *äpple.*
SG.NEUT big-SG.NEUT apple.SG.NEUT

'A big apple.'

b. *En* *stor-∅* *häst.*
SG.UTER big-SG.UTER horse.SG.UTER

'A big horse.'

Uter in Swedish historically derives from a fusion of feminine and masculine gender. Old Swedish originally retained a three gender system including

masculine, feminine and neuter, as other ancient Indo-European languages (Luraghi, 2011, p.437). However, “linguistic change led to a merger between many morphological gender forms at the end of the Middle Ages, and masculine and feminine forms could not always be discriminated" (Andersson, 2000, p.552), eventually resulting in the two-gender system of modern Swedish. This diachronic change lead to a rather unbalanced distribution of nouns between uter and neuter. Further details are shown in the following Section.

While grammatical gender assignment on nouns is commonly viewed as arbitrary (Andersson, 1992; Teleman et al., 1999), contradictory observations are made in Swedish. Dahl (2000, p.586) points out that animate nouns strongly tend to be affiliated to the uter gender, especially “all non-pejorative, classificatory nouns denoting adult human beings, a qualified majority of all other human nouns and a majority of all other animate nouns". Apart from the historical explanation that uter combined masculine and feminine, which originally included animates of both biological gender, additional evidence for such tendency are brought via an analysis of pronouns and gender agreement. First, uter indefinite pronouns used without a noun are interpreted as referring to animates, c.f., *Jag såg någon* ‘I saw someone’ vs. *Jag såg något* ‘I saw something’. Second, in noun-phrase external agreement, uter forms are preferred with human referents even if the head noun of the noun-phrase is lexically neuter, e.g., in *ett ungt statsråd* ‘a young government minister’ biological gender tends to override grammatical gender in terms of complement and pronominal reference (Holmes and Hinchliffe, 2013, 98). Hence, “there is in fact a general rule assigning uter gender at least to human nouns and noun phrases referring to persons, with exceptions that are probably no more serious than in most gender systems" (Dahl, 2000, p.586-587).

A broad version of the rule would be to assign uter gender to animates and neuter gender to inanimates, while explaining the exceptions via a leakage of inanimates to uter gender. Such a hypothesis is also supported by Fraurud (2000, p.191), who observed the tendency of count/mass division between uter and neuter nouns. Nouns referring to concrete and countable entities are more likely to be uter while abstract or collective meanings are associated to neuter. As an example, “possible people containers" denoting locations or organizations are perceived as collective units and tend to be neuter (Fraurud, 2000, p.203). Some of these speculations will be compared with our findings via the computational analysis.

3.3.2 Source of data

Our model relies on two main sources of data: A raw corpus and a dictionary. Both data in this research originate from the Swedish Language Bank (Språkbanken) located at the University of Gothenburg: a corpus of Swedish raw sentences without part of speech tagging and a list of nouns affiliated

to grammatical gender. The corpus originates from Swedish Wikipedia at Wikipedia Monolingual Corpora, Swedish web news corpora (2001-2013), and the Swedish Wikipedia corpus collected by Språkbanken[4]. These types of corpora are commonly applied in computational analysis (Erk, 2012) and were judged suitable for our analysis. First, with regard to the raw corpus, the OpenNLP sentence splitter and tokenizer are used for normalization. By way of illustration, we replace all numbers with a special token NUMBER and convert uppercase letters to lowercase forms. Second, the list of nouns and their affiliated grammatical gender is extracted from the SALDO (Swedish Associative Thesaurus version 2) dictionary[5], which includes five categories: *uter*, *neuter*, *plural*, *vacklande* (variable) and *blank* (unassigned nouns). An overview of the distribution is displayed in Table 3.3.

Table 3.3. *Gender of nouns in Swedish based on SALDO*

CODE	GENDER	FREQUENCY	PERCENTAGE	EXAMPLE
u	uter	61745	69.83	lampfot, vagga
n	neuter	25148	28.44	adverb, pendelur
p	plural	333	0.38	anor, makaroner
v	vacklande	764	0.86	bukspott, kolesterol
	blank	437	0.49	fotboja, puma

The categorization of SALDO is "quite generous" and includes various potential forms and categories (Borin et al., 2008, 27), i.e., nouns mostly occurring in plural forms are listed as the separate type *plural* and nouns attributed to two gender according to speaker variation are also affiliated to the class *vacklande*. Moreover, some nouns are annotated as *blank* if their gender was "indeterminate" (Borin et al., 2008, 27). In our analysis, we include *uter* and *neuter* since only these two classes fulfill the conditions of grammatical gender. Moreover, the overall frequency and quantity of the *plural*, *vacklande* and *blank* nouns is relatively low. We thus consider that removing these patterns of variation from our data does not affect the analysis. Finally, due to the high ratio of compounds in Swedish (Carter et al., 1996; Ostling and Wiren, 2013; Ullman and Nivre, 2014), we excluded nouns with a frequency lower than 100 occurrences within our corpus. The filtered list of nouns contains 21,162 nouns and is shown in Table 3.4.

Table 3.4. *Uter and neuter nouns in Swedish based on SALDO*

CODE	GENDER	FREQUENCY	PERCENTAGE	EXAMPLE
u	uter	15002	70.89	lampfot, vagga
n	neuter	6160	29.11	adverb, pendelur

[4] `https://spraakbanken.gu.se/eng/resources/corpus`
[5] `https://spraakbanken.gu.se/eng/resource/saldo`

We observe a substantial reduction of the list of nouns in terms of size. Nevertheless, the general ratio of uter and neuter nouns is maintained as 70% – 30%. An additional verification in terms of frequency shows that the distribution of uter and neuter nouns is equally represented among high and low frequency words. As shown in Figure 3.15, the y-axis indicates the ratio of uter (white) and neuter (gray) nouns, while the x-axis refers to the 21,162 nouns included in our study partitioned into ten groups according to their descending frequency. For instance, both the uter-neuter ratio of the most frequent 2100 words (1) and the less frequent 2100 (10) are close to 70% – 30%.

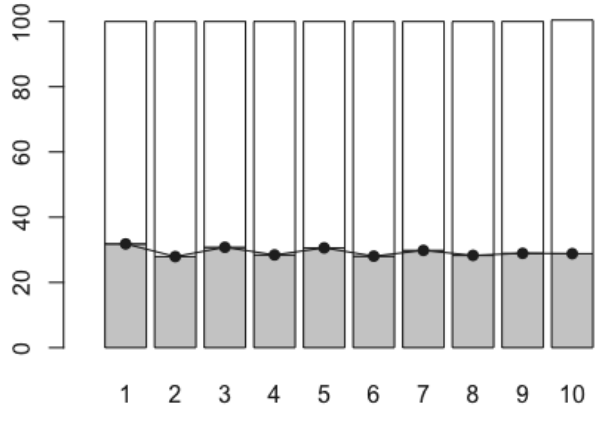


Figure 3.15. Distribution of uter (white) and neuter (gray) nouns with regard to frequency. The y-axis indicates the total ratio. The x-axis represents the nouns of the corpus partitioned into ten groups according to their descending frequency

The balance between neuter and uter nouns does not deviate from the general ratio attested in the entire lexicon, as the average of the uter-neuter balance across the ten groups is 70.70% – 29.30% with a standard deviation inferior to 1.35%. We therefore estimate that our filtering does not negatively affect the accuracy of our experiment. Furthermore, the distribution of uter and neuter nouns is expected to reflect the general tendency of language use within the corpus we apply in our study. This step is extremely important to decide the ratio of tokens from each category. By way of illustration, if uter and neuter nouns were equally distributed within the most frequent 2100 words, using the ratio of 70% – 30% may not reflect accurately what type of input data is commonly available to identify grammatical gender in Swedish.

3.3.3 The method: Word embeddings and neural networks

The recognition of grammatical gender on Swedish nouns may be categorized hypothetically in three possible approaches: selection by chance, scrutiny of the word itself, and analysis of the surrounding context. Selection by chance

refers to the majority label prediction (i.e., Zero rule) (Nasa and Suman, 2012), in which the baseline performance in terms of accuracy is set as the simplest classification method that relies on the majority and ignores all predictors. As suggested by the strategies employed by L2 adult learners of Swedish, guessing that a noun is uter provides a high chance of success since 71.06% of the nouns in Swedish are uter. Hence, a computational model is expected to at least surpass 71.06% of accuracy to be worth using. Second, the form of the word itself may provide hints. Several morphological regularities are attested, e.g., some derivational suffixes usually refer to a specific gender (*-eri* '-ing' for neuter, *-(h)et* '-ness/-(ab)ility' for uter). Moreover, phonological tendencies are also attested due to historical reasons, e.g., words in *-a* and *-e* tend to be uter. However, exceptions are frequent and gender assignment is still considered as opaque by linguists. We thus do not take into consideration scrutiny of the word (Nastase and Popescu, 2009) and target the analysis of the surrounding context via word embeddings.

Our procedure for the use of word embeddings is shown in Figure 3.16. The cylinders refer to the data sources and the rectangles refer to the processes. The entire experiment consists of three main steps. First (embedding), a corpus (raw sentences with segmented words) is fed to the word embedding model, which assigns a vector to each word according to its contexts of occurrence, i.e., which words are preceding and following. In our study, such vector representation is generated via the RSV (Real-valued Syntactic Word Vectors) model for word embeddings (Basirat and Nivre, 2017), which is an automatic method of word embedding based on the structure of GloVe (Pennington et al., 2014). In the second step (labeling), the list of word vectors associated with the nouns are labeled with their grammatical gender based on the dictionary. In the third step (classification), this list is then divided into three equivalent disjoint sets, namely train, development, and test sets (Bishop, 2006). The training set (80%, 16915/21162) is used by neural networks to generate different parameters of classifiers to handle the task of gender recognition. The development set (10%, 2104/21162) is used to tune the hyper-parameters of the word embedding model. Regarding context type, we investigate the three available options, i.e., forward, backward and both. As for context size, we include the following settings: 1,2,3,4,5 words. With regard to dimensionality, the typical values used in the literature are in the range of 5,10,50,100, and 200. We set the dimensionality as 50 to represent a balance between processing time and precision (Melamud et al., 2016). Finally, the test (10%, 2143/21162) set is used to measure the performance of neural networks. All words are randomly selected in their base format with no morphological inflection and all sets contain an equivalent distribution of uter and neuter nouns, i.e., the three partitions contain the same ratio of 70%-30% between uter and neuter nouns. As a summary, provided partial information on the gender of nouns in a language, neural networks may be able to predict the gender of other nouns (or novel nouns) in the same language.

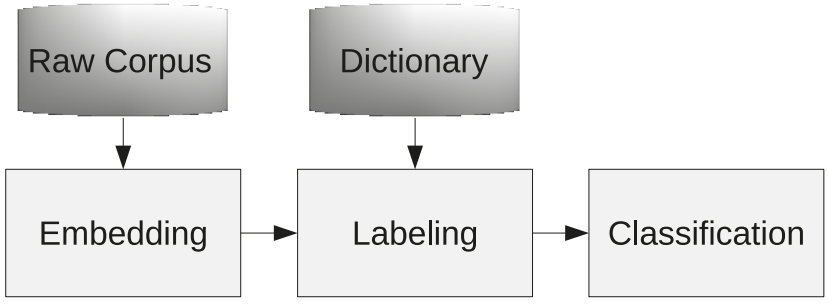


Figure 3.16. The process of predicting nominal classes from word embeddings

Recent research in computational linguistics focused on the performance of word embedding models with regard to classifying task, i.e., are the word vectors generated by word embedding models sufficient for the classifiers (e.g., neural network) to perform a classifying task with accuracy. Topics related to linguistics involved the differentiation of count and mass nouns (Katz and Zamparelli, 2012), the distinction of common and proper nouns (Lopez, 2008), opinion mining and sentiment analysis in texts (Pang and Lee, 2008), topic tracking in modern language-use via analysis of web-retrieved corpora (Baeza-Yates and Ribeiro-Neto, 2011), restoration of case marking (Baldwin et al., 2009), among others. The identification of grammatical gender has been investigated in a broad picture along with other semantic and syntactic linguistic features such as count/mass and common/proper (Basirat and Tang, 2018a,b). These studies demonstrated that grammatical gender could be classified with high accuracy via neural networks based on information extracted from word embeddings. However, tuning the quantity of neurons could not further improve the performance. Moreover, the effect of word frequency was not taken into account and no detailed error analysis was provided. We thus aim at filling these two gaps and suggest potential solutions to increase the classification accuracy.

3.3.4 Results

We first display the results of the development set according to the parameters of the word embedding model, context type and context size. Then, based on the tuning from the development set, we run neural networks on the test set to evaluate the performance of the model. The output of neural networks during the development set is assessed with the *F-score* (Ting, 2010), which is based on the weight of *Precision* and *Recall*. Finally, we provide additional analyses to evaluate the effect of word frequency and the patterns of classification errors.

Two major observations can be extracted from the development set and be summarized in Figure 3.17. First, the error rate of recognizing neuter nouns

is positively correlated with context size, as all three window types perform at their best with window size set as one. We suspect that this effect is caused by the increase of irrelevant information when window size expands, i.e., increasing the window size includes larger syntactic domain and incorporate words that may be uninformative or confusing when predicting the grammatical gender of the target noun. By way of illustration, in a sentence composed of a subject-noun, verb, and object-noun, the grammatical gender of the object-noun may differ from the subject-noun. A larger window size would thus take into account information about the gender of both nouns and encounter difficulties to predict the gender of the subject-noun. Second, neural networks generate the best performance when setting the context size as one in terms of asymmetric backward context. Following our previous explanation, the symmetric context type takes into account both preceding and following words, which can combine confusing information. As for the asymmetric-forward context type, its poor performance is also expected in terms of language structure: in syntactic-head-final languages such as Swedish, syntactically relevant information tend to be in the preceding position (Broekhuis, 2011).

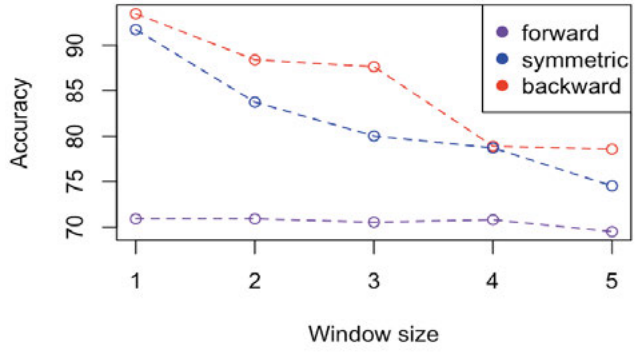


Figure 3.17. Overall performance of neural network with different context type.

Based on the the parameters of window size one with asymmetric-backward window type, the performance of neural networks reaches an f-score of 92.02% (Table 3.5). This exceeds by far the majority label prediction of 71.0%. Yet, the values of precision, recall, and f-score are all higher for uter nouns; which means that neuter nouns were harder to identify for neural networks both in terms of positive predictive value and sensitivity.

Table 3.5. *The performance of neural networks on grammatical gender prediction.*

	PRECISION	RECALL	F-SCORE
Neuter	88.70%	84.16%	86.37%
Uter	93.34%	95.40%	94.36%
Overall	91.98	92.12	92.02

To visualize how neural networks conceive gender of nouns in Swedish, we plot the spatial representation generated by neural networks in Figure 3.18. Such visualization is obtained by reducing the 50 dimensions included in our experiment settings to a two-dimensional representation via the tSNE software (Maaten and Hinton, 2008). First, this two-dimensional space reflects the unbalanced distribution between uter and neuter nouns (70.89% and 29.10%) as the cluster formed by uter nouns (green) outsize the agglomeration of neuter nouns (blue). Second, uter and neuter nouns are scattered in two different areas, which implicates that they can be distinguished according to semantic and/or syntactic features present in the language. Third, the errors of neuter nouns misinterpreted as uter nouns (black triangle) are mostly located in the uter nouns cluster (green). In other words, the model had difficulties recognizing neuter nouns which were located within the space of uter nouns, and vice-versa. If gender was not encoded according to certain semantic and/or syntactic factors, we would expect to see uter and neuter nouns randomly scattered. However, we observe the opposite, since uter and neuter nouns do form different clusters in Figure 3.18. This demonstrates that semantic and syntactic regularities are embedded in the language and are meaningful to differentiate between uter and neuter nouns in Swedish. Hence, we expect that the errors generated by the model are linguistically motivated. By way of illustration, the errors are expected to be Swedish words which have a semantic or syntactic overlap between uter and neuter.

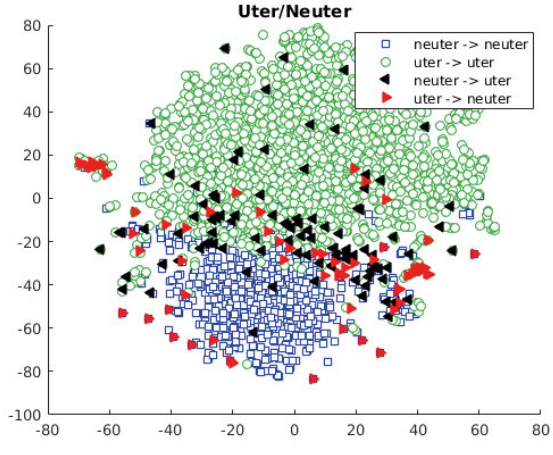


Figure 3.18. tSNE representation of the word vectors classified by neural networks with respect to their grammatical gender. "X > Y" means the noun belonging to category X is classified as Y

We equally need to evaluate the confidence level of the model along with its performance. Even though neural networks could identify correctly 92.02% of the test set, it is necessary to analyze if such task was relatively easy in terms

of decision process. Figure 3.19 shows the histogram of the entropy from the output of neural networks. The entropy scales the uncertainty involved to identify the noun classes. By way of illustration, high values of entropy can be interpreted as more uncertainty in the output of the classifier, which shows the weakness of the information provided by the word vectors with regard to nominal classes. A histogram skewed toward left shows the high certainty of the classifier for a particular nominal class, e.g., most words classified with an entropy close to zero implies that neural networks were highly confident when labeling the gender of the noun. However, if the histogram is skewed toward right, the classifier is uncertain about its outputs.

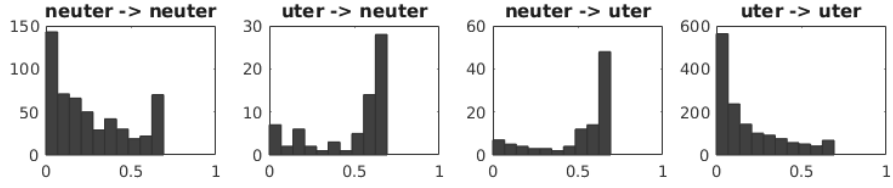


Figure 3.19. The histogram of the entropy of the output from neural networks with regard to grammatical gender. The y-axis indicates the amount of words from the test set, whereas the x-axis refers to the entropy.

The most left and right histogram displays a left-oriented skewness. Neural networks were thus relatively confident when classifying correctly the nouns according to their gender. Moreover, the middle graphs representing the erroneous output of neural networks are skewed toward the right. Neural networks were uncertain when classifying certain nouns, which resulted in a false identification of gender. We expect that the entropy is representative of the models precision: a lower entropy equals a low level of uncertainty when classifying nouns according to their gender. This hypothesis is verified in Figure 3.20, where we visualize that the mean and median entropy of the errors (0.50) is much higher than the mean entropy of the correct outputs (0.20). The *non-parametric approximative two-sample Fisher-Pitman permutation test* shows that the null hypothesis of no-association between the two categories can be rejected at a statistically significant level with regard to probability and indicates a strong negative correlation (z = -16.6, $p < 0.001$)[6].

[6]We apply the non-parametric approximative two-sample Fisher-Pitman permutation test due to the fact that we cannot statistically reject the null hypotheses of non-homoscedastic variance and unequal sample size in our data

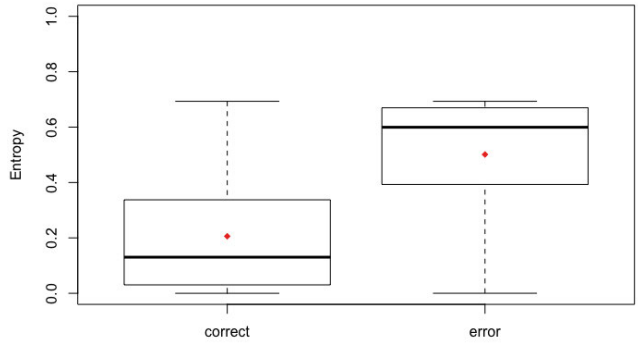
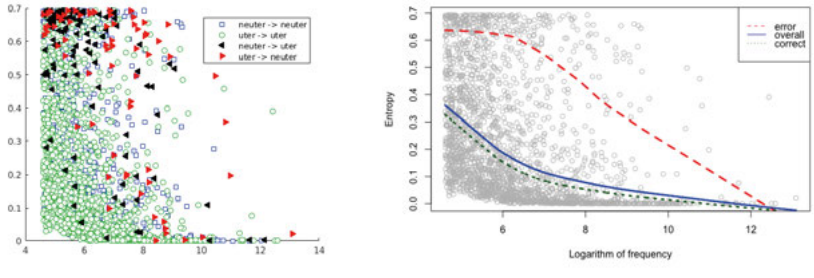


Figure 3.20. Box plot of entropy in correct and erroneous outputs of neural networks with regard to grammatical gender

However, an alternative explanation of such distribution could be related to frequency. An intuitive guess would be that the vectors of high-frequency nouns include more information since the nouns are represented by more tokens in the corpus. In such case, the semantic and syntactic information would be not be relevant with regard to nominal classification. We thus visualize in Figure 3.21a the general distribution of the test set. If the accuracy of neural networks was purely based on word-frequency, we would expect high entropy for low-frequency word and vice-versa. The left-skewed pattern of tokens of errors apparently support such hypothesis. Nevertheless, we may equally find that most of the low-frequency words are also classified correctly by neural networks. Therefore, we expect that frequency should not have a strong effect size. The output of the *Kendall's tau non-parametric correlation test* (Abdi, 2007) supports such speculation as the negative correlation between entropy and frequency is statistically significant but moderately strong (z = -25.395, *tau* = -0.3663, $p < 0.001$).



(a) Distribution of words (b) Regression analysis

Figure 3.21. Distribution of the test set with regard to entropy and frequency. The y-axis indicates the entropy, while the x-axis refers to the natural logarithm of frequency.

This weak correlation between entropy and frequency is further proven by the following observations. The association between the two variables has a non-linear monotonic nature, i.e., the lines in Figure 3.21b show that the increase of frequency may include quite a large quantity of nouns without any significant decrease in terms of entropy. However, after a certain level of frequency, the entropy drops relatively fast. The effect of frequency is small within the low-frequency nouns whereas a stronger effect size is observed within the high-frequency words. Following the assumptions of Zipf's law (Zipf, 1935), we observe that the majority of the nouns are found under the frequency logarithm of eight (86.65%, 1857/2143). Thus, a re-run of Kendall's tau test with solely the subset of nouns with frequency logarithm below eight illustrates that the correlation between entropy and frequency is less strong within tokens of correct classification (z = -20.419, *tau* = -0.3292 , $p < 0.001$). Such effect is even more salient with regard to the errors (z = -3.6542, *tau* = -0.2079, $p < 0.001$), as the *tau* coefficient decreases and the probability of the null hypothesis augments. As a summary, the visualization of semantic space and the statistical analysis between frequency and entropy demonstrated that frequency only had a weak effect size on the classification task. Neural networks were able to recognize the gender of nouns based on semantic and syntactic context information retrieved from word embeddings.

Following our second research question, word embeddings combined with neural networks could capture with high accuracy (92.02%) the grammatical gender of the nouns in Swedish. Yet, this result is not as 'perfect' as expected. We thus provide a linguistic categorization of all the errors made by neural networks during the test set. Table 3.6 displays the distribution of the errors among the main and sub-categories, along with examples.

Table 3.6. *Errors of neural networks in the test set*

CATEGORY	QUANTITY	RATIO	EXAMPLE
Noise	17	9.94%	
different gender in dictionary and corpus	11	6.43%	tidsplan
proper name	6	3.51%	rosengård
Bare noun	44	25.73%	
abstract noun	10	5.85%	fjärilsim
fixed usage	12	7.02%	pistolhot
mass	22	12.87%	fosfat
Polysemy	110	64.33%	
different meanings with different gender	10	5.85%	vad
different parts of speech	100	58.48%	kaukasiska
Total	171	100%	

Our analysis shows that the errors can be labeled with the following three categories: noise, bare nouns, and polysemy. First, noise is defined as a contradiction between the gender annotated in the dictionary and the gender

observed in corpus. Second, bare nouns refer to nouns which are only used in an isolated form. Third, polysemy includes nouns that may indicate two or more referents labeled with divergent gender or different parts of speech. Most of the errors were related to noise in the raw data or cases of polysemy with regard to the targeted nouns. By way of illustration, one word form may have more than two referents, which are respectively uter and neuter. Moreover, one word form may refer to a noun and an adjective depending on the context. This type of problem is expected to be solved by adjusting the parameters of the model, e.g., avoid a simple binary choice between uter and neuter during the classification process. On the other hand, neuter nouns were harder to identify than uter nouns. We speculate that this is related the the fact that most uter nouns are related to animate and countable nouns, which rarely occur as bare nouns. Therefore, word embeddings can retrieve more information from the surrounding context of the noun. As for neuter nouns, they are commonly abstract and mass nouns (Dahl, 2000; Fraurud, 2000), which were more likely to represent difficulties in classification as these two types of nouns generally occur in bare forms and do not provide sufficient clues in word embeddings. Mass nouns were even harder to identify since they often undergo semantic conversion toward count nouns (Gillon, 1999), which "incarnate complication for word embeddings"(Basirat and Tang, 2018a, p. 672).

3.3.5 Conclusion

Our main contributions are as follow: we were able to show that word embeddings combined with neural networks are capable of capturing the information of grammatical gender in Swedish with an accuracy of 92%. From a linguistic approach, we added an analysis with regard to the errors generated by neural networks and scrutinized the effect of word frequency. The error analysis have shown that word embeddings encounter difficulties in cases of polysemy, i.e., a linguistic form may link to different referents which belong to different part of speech categories or different gender. Such phenomenon is explained by linguistic theories of gender assignment, as neuter nouns are generally mass nouns, which undergo conversion between different part of speech categories. Additional tuning of the computational model in that direction is thus expected to improve the performance. Finally, we have demonstrated that our dataset was not affected by word frequency; yet, we strongly recommend this verification for all types of classification tasks since different datasets may behave differently in terms of internal distribution.

4. Concluding discussions

Within this chapter, we explain how the main aims of this thesis were accomplished and list the limitations of our studies along with suggestions for future research.

4.1 Summary

The first aim of this thesis was to provide the description and functional analysis of three languages with gender, classifiers, and both. This aim has been fulfilled in Chapter 2 via the analysis of gender in Marathi, classifiers in Assamese, and gender systems plus classifiers in Nepali. The first two papers provided examples of canonical gender and classifier languages, whereas the third paper demonstrated a co-occurrence of two gender systems plus a sortal classifier system in one language and their complementary distribution of functions. This contributes to research on typology by presenting data from canonical nominal classification systems and a rare and untypical case study. Moreover, the complementary of functions between gender systems and classifiers also relate to the discussion of linguistic complexity, as lexical and discourse functions of the two types of systems obey the principles of economy and distinctiveness in language.

The second aim of this thesis was to address the lack of arbitrariness of nominal classification systems at three different scales: The distribution of classifiers at the worldwide level, the presence of gender within a language family, and gender assignment at the language-internal level. This aim has been realized in Chapter 3 via the use of the different quantitative methods on linguistics data. We have shown that the output of computational methods mostly correlated with the linguistic hypotheses. The distribution of sortal classifiers in languages of the world could be predicted by random forests based on the existence/absence of morphosyntactic plural markers and multiplicative bases. Moreover, measuring phylogenetic inferences of grammatical gender within a language family could provide information on language change. For instance, we were able to identify which features were more likely to be influenced from language contact than others. Finally, the tendencies of grammatical gender affiliation in Swedish were observed in our experiment via word embeddings.

The third aim of this thesis was to introduce new applications of quantitative methods from biology and computer sciences to answer linguistic questions.

This aim was accomplished via the use of three different quantitative methods in Chapter 3. Random forests was used to investigate probabilistic universals with regard to the distribution of sortal classifiers. Phylogenetic inferences were used to investigate the diachronic change of grammatical gender in the Indo-Aryan language family. Word embeddings with neural networks could retrieve semantic and syntactic regularities of grammatical gender in Swedish. The classification errors also matched the linguistic hypotheses on gender assignment in Swedish.

4.2 Future studies

Typical large scale quantitative studies could include comparison between different types of data and quantitative methods. For instance, random forests and neural networks are both computational classifiers. Their respective performance could be compared in our analysis of probabilistic universals and gender assignment. These two classifiers also have different variants that could perform better or worse depending on the the classification task, e.g., we used a single layer feed-forward neural network in our experiments, but other structures of neural networks could be tested too. These possibilities were not investigated in this thesis since our main aims were to introduce these quantitative methods to linguistic data. However, it could be interesting to test the same data on different quantitative methods and assess the converge and divergence of their results. Likewise in terms of data, using different sources of input on the same method could reveal its strengths and weaknesses. By way of illustration, a typical computational study would compare the performance of word embeddings and neural networks on several grammatical gender languages that have different types of gender, e.g., French with biological gender and German with masculine, feminine, and neuter. We did not include these parts in the thesis to maintain focus on the methods; yet, other projects of the author are currently investigating these research questions.

References

Abdi, H. (2007). The Kendall rank correlation coefficient. In Salkind, N., editor, *Encyclopedia of measurement and statistics*, pages 508–510. Sage, Thousand Oaks.

Acharya, J. (1991). *A descriptive grammar of Nepali and an analyzed corpus*. Georgetown University Press, Washington, D.C.

Agnihotri, R. K. (2007). *Hindi: an essential grammar*. Routledge essential grammars. Routledge, London ; New York. OCLC: ocm72799377.

Aikhenvald, A. Y. (2000). *Classifiers: A typology of noun categorization devices*. Oxford University Press, Oxford.

Albarracin, L., Aymerich, A., and Gorgorio, N. (2017). An open task to promote students to create statistical concepts through modelling. *Teaching Statistics*, 39(3):100–105.

Andersson, A.-B. (1992). *Second language learners' acquisition of grammatical gender in Swedish*. PhD dissertation, University of Gothenburg, Gothenburg.

Andersson, E. (2000). How many gender categories are there in Swedish? In Unterbeck, B., Rissanen, M., Nevalainen, T., and Saari, M., editors, *Gender in Grammar and Cognition*, pages 545–559. Mouton de Gruyter, Berlin.

Apte, M. L. (1962). *A sketch of Marathi transformational grammar*. PhD dissertation, University of Wisconsin, Madison.

Baart, J. L. (1999). *A sketch of Kalam Kohistani grammar*. Summer Institute of Linguistics, Dallas.

Baeza-Yates, R. and Ribeiro-Neto, B. (2011). *Modern information retrieval: The concepts and technology behind search*. Addison Wesley Longman Limited, Essex.

Baldwin, T., Paul, M., and Joseph, A. (2009). Restoring punctuation and casing in English text. *Proceedings of the 22nd Australian Joint Conference on Artificial Intelligence (AI09)*, pages 547–556.

Barz, R. and Diller, A. (1985). Classifiers and standardization: some South and South-East Asian comparisons. *Papers in Southeast Asian Linguistics*, 9:155–184.

Basirat, A. and Nivre, J. (2017). Real-valued Syntactic Word Vectors (RSV) for Greedy Neural Dependency Parsing. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa*, pages 21–28, Gothenburg. Linkoping University Electronic Press.

Basirat, A. and Tang, M. (2018a). Lexical and morpho-syntactic features in word embeddings: A case study of nouns in Swedish. *Proceedings of the 10th International Conference on Agents and Artificial Intelligence*, 2:663–674.

Basirat, A. and Tang, M. (2018b). Linguistic information in word embeddings. In Goebel, R., Tanaka, Y., and Wahlster, W., editors, *Lecture notes in artifical intelligence*. Springer, Dordrecht.

Beekes, R. S. P. (2011). *Comparative Indo-European linguistics*. John Benjamins, Amsterdam.

Ben-Zvi, D. and Garfield, J. (2004). *The challenge of developing statistical literacy, reasoning, and thinking*. Springer, Dordrecht.

Bisang, W. (1999). Classifiers in East and Southeast Asian languages: Counting and beyond. In Gvozdanovic, J., editor, *Numeral types and changes worldwide*, pages 113–186. Walter de Gruyter, Munchen.

Bisang, W. (2012). Numeral classifiers with plural marking: A challenge to Greenberg. In Xu, D., editor, *Plurality and classifiers across languages of China*, pages 23–42. De Gruyter Mouton, Berlin.

Bisang, W. (2014). Overt and hidden complexity: Two types of complexity and their implications. *Poznan Studies in Contemporary Linguistics*, 50(2):127–143.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer, New York.

Blomberg, S. P., Garland Jr, T., and Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution*, 57:717–745.

Borah, G. K. (2012). Classifiers in Assamese: Their grammar and meaning chains. In Hyslop, G., Morey, S., and Post, M. W., editors, *North East Indian Linguistics Volume 4*, pages 292–314. Cambridge University Press, New Delhi.

Borer, H. (2005). *Structuring Sense, part I*. Oxford University Press, Oxford.

Borin, L., Forsberg, M., and Lonngren, L. (2008). The hunting of the BLARK - SALDO, a freely available lexical database for Swedish language technology. *Studia Linguistica Upsaliensia*, pages 21–32.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. (1984). *Classification and regression trees*. Taylor & Francis, New York.

Broekhuis, H. (2011). Word order typology. In van Craenenbroeck, J., editor, *Linguistic variation yearbook 10*, pages 1–31. John Benjamins, Amsterdam.

Brown, M. (2017). Making students part of the dataset: A model for statistical enquiry in social issues. *Teaching Statistics*, 39(3):79–83.

Buckley, J., Brown, M., Thomson, S., Olsen, W., and Carter, J. (2015). Embedding quantitative skills into the social science curriculum: case studies from Manchester. *International Journal of Social Research Methodology*, 18:495–510.

Carnell, L. J. (2008). The effect of a student-designed data collection project on attitudes toward statistics. *Journal of Statistics Education*, 16(1):1–15.

Carter, D., Kaja, J., Neumeyer, L., Rayner, M., Weng, F., and Wiren, M. (1996). Handling compound nouns in a Swedish speech-understanding system. *Proceedings of the Fourth International Conference on Spoken Language*, 1:26–29.

Chen, X. and Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, 99(6):323–329.

Chierchia, G. (1998). Plurality of mass nouns and the notion of semantic parameter. In Rothstein, S., editor, *Events and grammar*, pages 53–104. Kluwer, Dordrecht.

Clahsen, H. (2016). Contributions of linguistic typology to psycholinguistics. *Linguistic Typology*, 20(3):599–614.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Contini-Morava, E. and Kilarski, M. (2013). Functions of nominal classification. *Language Sciences*, 40:263–299.

Corbett, G. G. (1979). The agreement hierarchy. *Journal of Linguistics*, 15:203–224.

Corbett, G. G. (1991). *Gender*. Cambridge University Press, Cambridge.

Corbett, G. G. (2012). *Features*. Cambridge University Press, Cambridge.

Corbett, G. G. (2013a). Number of Genders. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Corbett, G. G. (2013b). Systems of gender assignment. In Dryer, M. S. and Haspelmath, M., editors, *The world atlas of language structures online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Corbett, G. G. and Fedden, S. (2016). Canonical gender. *Journal of Linguistics*, 52(3):495–531.

Dahl, O. (2000). Elementary gender distinctions. In Unterbeck, B. and Rissanen, M., editors, *Gender in grammar and cognition*, pages 577–593. Mouton de Gruyter, Berlin.

Dhongde, R. V. and Wali, K. (2009). *Marathi*. John Benjamins, Amsterdam.

Di Garbo, F. (2014). *Gender and its interaction with number and evaluative morphology an intra- and intergenealogical typological survey of Africa*. PhD dissertation, Stockholm University, Stockholm.

Dixon, R. M. W. (1982a). Noun classes. In Dixon, R. M. W., editor, *Where have all the adjectives gone? And other essays in semantics and syntax*, pages 157–184. Mouton de Gruyter, Berlin.

Dixon, R. M. W. (1982b). Noun classifiers and noun classes. In Dixon, R. M. W., editor, *Where have all the adjectives gone? And other essays in semantics and syntax*, pages 211–233. Mouton de Gruyter, Berlin.

Dixon, R. M. W. (1986). Noun class and noun classification. In Craig, C., editor, *Noun classes and categorization*, pages 105–112. John Benjamins, Amsterdam.

Doetjes, J. (2012). Count/mass distinctions across languages. In Maienborn, C., Heusinger, K. v., and Portner, P., editors, *Semantics: an international handbook of natural language meaning, part III*, pages 2559–2580. Mouton de Gruyter, Berlin.

Dryer, M. S. (2013). The Order of numeral and noun. In Dryer, M. S. and Haspelmath, M., editors, *The word atlas of language structures online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Dryer, M. S. and Haspelmath, M. (2013). *The world atlas of language structures online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Dunn, M. (2015). Language phylogenies. In *The Routledge Handbook of Historical Linguistics*, pages 190–211. Routledge, New York.

Dunn, M., Greenhill, S. J., Levinson, S. C., and Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82.

Dunn, M., Kruspe, N., and Burenhult, N. (2013). Time and place in the prehistory of the Aslian languages. *Human biology*, 85(1):383–400.

Emeneau, M. B. (1956). India as a Lingustic Area. *Language*, 32(1):3–16.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Fedden, S. and Corbett, G. G. (2017). Gender and classifiers in concurrent systems: Refining the typology of nominal classification. *Glossa: a journal of general linguistics*, 2(1):1–47.

Fraurud, K. (2000). Proper names and gender in Swedish. In Unterbeck, B., Rissanen, M., Nevalainen, T., and Saari, M., editors, *Gender in Grammar and Cognition*, pages 167–220. Mouton de Gruyter, Berlin.

Freckleton, R. P., Harvey, P. H., and Pagel, M. (2002). Phylogenetic analysis and comparative data: a test and review of evidence. *American Naturalist*, 160:712–726.

Fromkin, V., Rodman, R., and Hyams, N. (2011). *An introduction to language*. Wadsworth, Boston.

Gair, J. W. (2007). Sinhala. In Cardona, G. and Jain, D., editors, *The Indo-Aryan languages*, pages 766–817. Routledge, New York.

Gerner, M. (2006). Noun classifiers in Kam and Chinese Kam-Tai languages: Their morphosyntax, semantics and history. *Journal of Chinese Linguistics*, 34(2):237–305.

Ghomeshi, J. (2003). Plural marking, indefiniteness, and the noun phrase. *Studia Linguistica*, 57(2):47–74.

Gil, D. (2013). Numeral classifiers. In Dryer, M. S. and Haspelmath, M., editors, *The world atlas of language structures online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Gillon, B. S. (1999). The lexical semantics of English count and mass nouns. In Viegas, E., editor, *Breadth and depth of semantic lexicons*, pages 19–37. Springer, Dordrecht.

Goswami, G. C. and Tamuli, J. (2003). Asamiya. In Jain, D. and Cardona, G., editors, *The Indo-Aryan Languages*, pages 391–443. Routledge, New York.

Gould, R. (2010). Statistics and the modern students. *International Statistical Review*, 78:297–315.

Gowri-Shankar, V. and Rattray, M. (2007). A reversible jump method for bayesian phylogenetic inference with a nonhomogeneous substitution model. *Molecular Biology and Evolution*, 24(6):1286–1299.

Green, P. J. (1995). Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732.

Greenberg, J. H. (1990a). Generalizations about numeral systems. In Denning, K. and Kemmer, S., editors, *On language: Selected writings of Joseph H. Greenberg*, pages 271–309. Stanford University Press, Stanford. [Originally published 1978 in Universals of Human Language, ed by Joseph H. Greenberg, Charles A. Fergson, & Edith A. Moravcsik, Vol 3, 249-295. Stanford; Stanford University Press.].

Greenberg, J. H. (1990b). Numeral classifiers and substantival number: Problems in the genesis of a linguistic type. In Denning, K. and Kemmer, S., editors, *On language: Selected writings of Joseph H. Greenberg*, pages 166–193. Stanford University Press, Stanford. [First published 1972 in Working Papers on Language Universals 9. 1-39. Stanford, CA: Department of Linguistics, Stanford University.].

Grinevald, C. (2000). A morphosyntactic typology of classifiers. In Senft, G., editor, *Systems of nominal classification*, pages 50–92. Cambridge University Press, Cambridge.

Grinevald, C. (2015). Linguistics of classifiers. In Wright, J. D., editor, *International Encyclopedia of the Social & Behavioral Sciences*, pages 811–818. Elsevier, Oxford.

Grinevald, C. and Seifart, F. (2004). Noun classes in African and Amazonian languages: Towards a comparison. *Linguistic Typology*, 8(2):243–285.

Gupte, S. M. (1975). *Relative constructions in Marathi*. PhD dissertation, Michigan State University, Ann Arbor.

Hammarstrom, H., Bank, S., Forkel, R., and Haspelmath, M. (2018). *Glottolog 3.2*. Max Planck Institute for the Science of Human History, Jena.

Harrell, F. (2001). *Regression modeling strategies*. Springer, New York.

Harrell, F. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, Dordrecht.

Hawkins, J. (2004). *Efficiency and complexity in grammars*. Oxford University Press, Oxford.

Her, O.-S. (2012a). Distinguishing classifiers and measure words: A mathematical perspective and implications. *Lingua*, 122(14):1668–1691.

Her, O.-S. (2012b). Structure of classifiers and measure words: A lexical functional account. *Language and Linguistics*, 13:1211–1251.

Her, O.-S. (2017). Deriving classifier word order typology, or Greenberg's Universal 20a and Universal 20. *Linguistics*, 55(2):265–303.

Her, O.-S. and Lai, W.-J. (2012). Classifiers: The many ways to profile one, a case study of Taiwan Mandarin. *International Journal of Computer Processing of Oriental Languages*, 24(1):79–94.

Her, O.-S., Tang, M., and Li, B.-T. (2018). Word order of numeral classifiers and numeral bases: Harmonization by multiplication. *Language Typology and Universals*, (to appear).

Holmes, P. and Hinchliffe, I. (2013). *Swedish: a comprehensive grammar*. Routledge, New York.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–574.

Huelsenbeck, J. P. and Ronquist, F. (2001). MrBayes: Bayesian inference of phylogeny. *Bioinformatics*, 17:754–755.

Ives, A. R., Midford, P. E., and Garland Jr, T. (2007). Within-species variation and measurement error in phylogenetic comparative biology. *Systematic Biology*, 56:252–270.

Jeffreys, H. (1961). *Theory of probability (3rd ed.)*. Oxford University Press, Oxford.

Joshi, S. (1993). *Selections of grammatical and logical functions in Marathi*. PhD dissertation, University of California, Stanford.

Kalita, J. C. (2003). *Nouns and nominalisations in Assamese: A microlinguistic study*. PhD dissertation, Gauhati University, Guwahati.

Katz, G. and Zamparelli, R. (2012). Quantifying count/mass elasticity. *Proceedings of the 29th West Coast Conference on Formal Linguistics*, pages 371–379.

Kemmerer, D. (2017a). Categories of object concepts across languages and brains: the relevance of nominal classification systems to cognitive neuroscience. *Language, Cognition and Neuroscience*, 32(4):401–424.

Kemmerer, D. (2017b). Some issues involving the relevance of nominal classification systems to cognitive neuroscience: response to commentators. *Language, Cognition and Neuroscience*, 32(4):447–456.

Kibort, A. and Corbett, G. G. (2008). Number: Grammatical features.

Kilarski, M. (2014). The Place of Classifiers in the History of Linguistics. *Historiographia Linguistica*, 41(1):33–79.

Kiryu, K. (2009). On the rise of the classifier system in Newar. *Senri Ethonological Studies*, 75:51–69.

Kolipakam, V., Jordan, F. M., Dunn, M., Greenhill, S. J., Bouckaert, R., Gray, R. D., and Verkerk, A. (2018). A Bayesian phylogenetic study of the Dravidian language family. *Royal Society Open Science*, 5(3):1–17.

Lakoff, G. and Johnson, M. (2003). *Metaphors we live by*. The University of Chicago Press, London.

Lee, S.-W. (2011). Eastern Tamang grammar sketch. Master's thesis, Graduate Institute of Applied Linguistics, Dallas and SIL Nepal.

Legendre, P., Fortin, M.-J., and Borcard, D. (2015). Should the Mantel test be used in spatial analysis? *Methods in Ecology and Evolution*, 6:1239–1247.

Levinson, S. C., Greenhill, S. J., Gray, R. D., and Dunn, M. (2011). Universal typological dependencies should be detectable in the history of language families. *Linguistic Typology*, 15(2).

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. John Benjamins, Amsterdam.

Lewis, P., Simons, G. F., and Fennig, C. D. (2009). *Ethnologue: Languages of the world*. SIL International, Dallas.

Li, X. and Bisang, W. (2012). Classifiers in Sinitic languages: From individuation to definiteness-marking. *Lingua*, 122:335–355.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3):18–22.

Lopez, A. (2008). Statistical machine translation. *ACM Computing Surveys*, 40(3):1–49.

Luraghi, S. (2011). The origin of the Proto-Indo-European gender system: Typological considerations. *Folia Linguistica*, 45(2):435–464.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

MacDonell, A. A. (1999). *A Vedic grammar for students*. Printworld, New Delhi.

Mace, R. and Holden, C. J. (2005). A phylogenetic approach to cultural evolution. *Trends in Ecology & Evolution*, 20(3):116–121.

MacInnes, J. (2009). *Proposals to support and improve the teaching of quantitative research methods at undergraduate level in the UK*. ESRC, Swindon.

Manders, C. J. (2007). *A foundation in Nepali grammar*. AuthorHouse, Bloomington.

Marnita, R. (1996). Classifiers in Minangkabau. Master's thesis, Australian National University, Canberra.

Master, A. (1964). *A grammar of Old Marathi*. Clarendon Press, Oxford.

Matthews, D. (1998). *A course in Nepali*. Curzon SOAS books. Curzon, Richmond. OCLC: 833322821.

Melamud, O., McClosky, D., Patwardhan, S., and Bansal, M. (2016). The role of context types and dimensionality in learning word embeddings. *arXiv*, pages 1–11.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Weiling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in neural information processing systems*, pages 3111–3119. Curran Associates, New York.

Moral, D. (1997). North East India as a Linguistic Area. *Mon-Khmer Studies*, 27:43–53.

Nasa, C. and Suman (2012). Evaluation of different classification techniques for web data. *International Journal of Computer Applications*, 52(9):34–40.

Nastase, V. and Popescu, M. (2009). What's in a name?: in some languages, grammatical gender. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 3:1368–1377.

Nayudu, A. (2008). *Issues in the syntax of Marathi: A minimalist approach*. PhD dissertation, Durham University, Durham.

Noonan, M. (2003). Recent language contact in the Nepal Himalaya. In Bradley, D., LaPolla, R. J., Michailovsky, B., and Thurgood, G., editors, *Language variation: Papers on variation and change in the Sinosphere and in the Indosphere in honour of James A. Matisoff*, pages 65–88. Pacific Linguistics, Canberra.

Ostling, R. and Wiren, M. (2013). Compounding in a Swedish blog corpus. *Acta Universitatis Stockholmiensis*, pages 45–63.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, 401:877–884.

Paluszynska, A. (2017). Structure mining and knowledge extraction from random forest with applications to the cancer genome atlas project. Master's thesis, University of Warsaw, Warsaw.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Peterson, J. (2017). Fitting the pieces together: Towards a linguistic prehistory of eastern-central South Asia (and beyond). *Journal of South Asian Languages and Linguistics*, 4(2):211–257.

Peyraube, A. (1998). On the history of classifiers in Archaic and Medieval Chinese. In T'sou, B. K., editor, *Studia linguistica serica*, pages 131–145. City University of Hong Kong, Hong Kong.

Phillips, M. P. (2012). *Dialect continuum in the Bhil tribal belt: Grammatical aspects*. PhD dissertation, School of Oriental and African Studies, University of London, London.

Pokharel, M. (1997). Gender system in Nepali. *Nepalese Linguistics*, 14:40–70.

Pokharel, M. P. (2010). Noun class agreement in Nepali. *Kobe papers in linguistics*, 7:40–59.

Popescu, A. A., Huber, K. T., and Paradis, E. (2012). Ape 3.0: New tools for distance based phylogenetics and evolutionary analysis in R. *Bioinformatics*, 28:1536–1537.

Poudel, K. P. (2010). Gender system in Nepali and Tamang. *Circle of English Teachers Journal*, 2(2):7–17.

Priestly, T. (1983). On drift in Indo-European gender systems. *Journal of Indo-European Studies*, 11:339–363.

R-Core-Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*, page syy032.

Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850.

Revell, L. J. (2012). Phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.*, 3:217–223.

Riccardi, T. (2003). Nepali. In Cardona, G. and Jain, D., editors, *The Indo-Aryan Languages*, pages 538–580. Routledge, London.

Rijkhoff, J. (2000). When can a language have adjectives? An implicational universal. In Vogel, P. M. and Comrie, B., editors, *Approaches to the typology of word classes*, pages 217–257. De Gruyter Mouton, Berlin.

Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19:1572–1574.

Sanches, M. and Slobin, L. (1973). Numeral classifiers and plural marking: An implicational universal. *Working Papers in Language Universals*, 11:1–22.

Schliep, K. P. (2011). Phangorn: Phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593.

Simons, G. F. and Fennig, C. D. (2018). *Ethnologue: Languages of the world (21st edition)*. SIL International, Dallas.

Sinnemaki, K. (2018). On the distribution and complexity of gender and numeral classifiers. In Di Garbo, F. and Walchli, B., editors, *Grammatical gender and linguistic complexity*. Language Science Press, Berlin. (to appear).

Tagliamonte, S. A. and Baayen, H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*, 24:135–178.

Tang, M. (2017). Explaining the acquisition order of classifiers and measure words via their mathematical complexity. *Journal of Child Language Acquisition and Development*, 5(1):31–52.

Tang, M., Her, O.-S., and Chen, Y.-R. (2018). Insights on the Greenberg-Sanches-Slobin Generalization: Quantitative typological data on classifiers and plural markers. *Folia Linguistica*, (to appear).

Teleman, U., Hellberg, S., and Andersson, E. (1999). *Svenska Akademiens grammatik. Vol. 2: Ord. [The Swedish Academy Grammar, Part 2: Words]*. Norstedts, Stockholm.

Ting, K. M. (2010). Precision and Recall. In Sammut, C. and Webb, G. I., editors, *Encyclopedia of Machine Learning*, pages 781–781. Springer US, Boston, MA.

T'sou, B. K. (1976). The structure of nominal classifier systems. *Oceanic Linguistics Special Publications*, 13:1215–1247.

Ullman, E. and Nivre, J. (2014). Paraphrasing Swedish Compound Nouns in Machine Translation. In *MWE@ EACL*, pages 99–103.

Upadhyay, S. R. (2009). The sociolinguistic variation of grammatical gender agreement in Nepali. *Journal of Pragmatics*, 41:564–585.

Verma, M. K. (2007). Bhojpuri. In Cardona, G. and Jain, D., editors, *The Indo-Aryan languages*, pages 515–537. Routledge, London.

Vogel, P. M. and Comrie, B., editors (2000). *Approaches to the typology of word classes*. Number 23 in Empirical approaches to language typology. Mouton de Gruyter, Berlin ; New York.

Wali, K. (1989). *Marathi syntax: A study of reflexives*. Indian Institute of Language Studies, Patiala.

Wali, K. (2006). *Marathi: A study of comparative South Asian languages*. Indian Institute of Language Studies, Delhi.

Weidert, A. K. (1984). The classifier construction of Newari and its historical Southeast Asian background. *Kailash: A Journal of Himalayan Studies*, 11(3-4):185–210.

Wild, C. and Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, 67:223–265.

Wu, F., Feng, S., and Huang, C. T. J. (2006). Hanyu shu+lianhg+ming geshi de laiyuan [On the origin of the construction of numeral+classifier+noun in Chinese]. *Zhongguo Yuwen [Studies of the Chinese Language]*, 5:387–400.

Yadav, R. (1996). *A reference grammar of Maithili*. Mouton de Gruyter, Berlin.

Zhang, N. N. (2013). *Classifier structures in Mandarin Chinese*. Mouton de Gruyter, Berlin.

Zipf, G. K. (1935). *The psycho-biology of language*. MIT Press, Cambridge.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Hafner, New York.