Chapter 3

The feminine anaphoric gender gram, incipient gender marking, maturity, and extracting anaphoric gender markers from parallel texts

Bernhard Wälchli
Stockholm University

The aim of this paper is to carry out a typological study of feminine anaphoric gender grams (such as English *she/her*) in a large world-wide convenience sample of 816 languages based on a strictly procedural definition. The investigation pursues a radically functional approach where the functional equivalence of the forms under study is assured by exploring an identical search space in parallel texts (translations of the New Testament) in all languages of the sample. This is the first large scale typological study of grammatical gender based on parallel texts, and a large part of the paper is devoted to methodological aspects. The study shows that gender has a functional core like any other grammatical category, and that it can at least partly be studied without resort to the notions of noun class, agreement and system. The results show that a large number of languages possess simple forms of gender, often representing incipient gender from a grammaticalization perspective. The paper discusses how simple gender differs from more mature and genealogically more stable forms of anaphoric gender. Finally the feminine anaphoric gram type is considered in its wider context, reconciling it to the traditional global approach focusing on the notions of system, noun class and agreement.

Keywords: feminine gender, anaphora, anaphoric pronouns, grams, grammaticalization, grammatical relations, functional domains, constructional islands, cue validity, maturity, parallel texts.

Bernhard Wälchli. 2019. The feminine anaphoric gender gram, incipient gender marking, maturity, and extracting anaphoric gender markers from parallel texts. In Francesca Di Garbo, Bruno Olsson & Bernhard Wälchli (eds.), *Grammatical gender and linguistic complexity: Volume II: World-wide comparative studies*, 61–131. Berlin: Language Science Press. DOI:10.5281/zenodo.3462780

1 Introduction

The traditional definition of gender ("Genders are classes of nouns reflected in the behavior of associated words"; Hockett 1958: 231) rests on the notions of noun class and agreement. With the exception of classifiers, for which noun classes are crucial as well, these notions do not figure in the definitions of other grammatical categories. This makes gender stand out among grammatical categories as very specific by definition. In this paper it is argued that it is also possible to address gender as any other grammatical category by defining it as "grammatical category expressing meaning X", where X can be feminine, masculine, animate and inanimate, given that the most widespread meanings in gender are animacy and sex (Dahl 2000: 101; Corbett 1991: 68; Luraghi 2011), at least as far as anaphoric gender is concerned.[1] In order to make clear that this paper mainly deals with gender marking in anaphoric contexts, I will use the term "anaphoric gender". The question of how grammatical gender is defined is highly relevant for assessing the complexity of grammatical gender.

Noun classes and agreement are complex phenomena. Accepting the traditional definition of gender as the only option would mean to take for granted that grammatical gender is complex by definition. In order to assess the complexity of grammatical gender empirically it is indispensable to explore the possibility of simpler alternative definitions. Linguists nowadays often understand "gender" and "noun class" as full synonyms. This may be appropriate for the study of gender within noun phrases, but does not do justice to the use of gender in the anaphoric domain, which is the topic of this paper.

The major aim of this paper is to show that gender has a functionally motivated semantic core that can be considered in abstraction from the notions of noun class and agreement. This is done by formulating a procedural definition of feminine anaphoric gender which is so explicit that it can be implemented in a computer program in order to extract certain feminine gender markers from parallel texts (here translations of the New Testament). Feminine is chosen for practical reasons. It is the easiest to address in this particular corpus (see §2.2).

[1] One of the first things I was ever taught in linguistics is that gender and sex are absolutely not the same thing and, since my department found me highly suitable for teaching numerous courses in discourse studies, sociolinguistics, and pragmatics and intercultural communication, I am quite familiar with gender studies and the notion of performative gender. However, the approach pursued in this paper focuses exclusively on the semantic core of feminine and masculine grammatical genders and here the gross simplification that sex is the core meaning of masculine and feminine gender grams has proven to be very useful in practice.

The underlying idea is that grammatical categories can be captured in terms of GRAMS. A gram is a grammatical item in a particular language with specific form and specific meaning and/or function (Bybee & Dahl 1989; Dahl & Wälchli 2016). Grams can be considered in abstraction from the language-specific systems they are part of. For instance, perfect and progressive can be investigated in abstraction from tense and aspect systems. For gender grams this means that the units of research are feminine, masculine, animate and inanimate, rather than gender systems. Virtually all gender systems are sensitive to the meanings sex and/or animacy (whereby different segments of the animacy hierarchy can be affected). It is true that gender in many languages also comprises other meanings, such as size and shape, and these other meanings are very important for the study of gender as systems. With the gram approach, however, it is possible to address the semantic core areas and to study them cross-linguistically, without having to consider the entire gender systems. A strength of the gram approach is its selectivity. Only salient semantic core uses are considered and compared cross-linguistically. A gram necessarily has a semantic core, but not all of its uses need be semantically motivated. The gram approach focuses on the semantic core of grammatical categories and investigates to what extent grams across different languages share their semantic core, put differently, cluster to cross-linguistic gram types. In order to find out whether a language has a gram reflecting a cross-linguistic gram type, it is sufficient to consider the prototypical uses of a gram type.

Focusing on the semantic core means focusing on those uses of a grammatical category where it is most transparent semantically. We know, among other things, from Corbett's (1991, chap. 8) study of the Agreement Hierarchy that gender use tends to be most transparently semantic in third person anaphoric pronouns. According to Audring (2009), all pronominal gender systems (where gender is restricted to pronouns) are semantically organized, which further supports the view that gender is most semantic in anaphoric use.

A feminine gender gram – in its prototypical use – is a grammatical element picking up reference to a female human, such as the English third person singular pronominal forms *she* and *her* exemplified in (1). (1) is one of 74 parallel corpus passages that are used as a search space for feminine anaphoric gender grams in this paper.

(1) English (Indo-European; Matth. 15:26–27): gender marking on free pronouns
But he answered: "…" But ***she*** *said: "…"*

What I have said so far may suggest that this is a paper about gender in personal pronouns such as English *she* (see, e.g., Audring 2009), but the search space is much broader. In many languages the functional equivalent to *she* and *her* in English is an affix on verbs and/or adpositions as in (2) from Garifuna.

(2) Garifuna (Arawakan; Matth. 15:26–27): gender marking on bound pronouns and prepositions
Ába ***l****-aríñagun Jesúsu* ***t****-un:* *"…" Ába* ***t****-aríñagun: "…"*
and 3SG.M-say Jesus 3SG.F-to and 3SG.F-say
'But he answered: "…" But she said: "…" '

Third person pronouns and affixes for third person have in common that they are REDUCED REFERENTIAL DEVICES in terms of Kibrik (2011; ch. 3), who calls them FREE and BOUND PRONOUNS. In (1) from English the gender marking is located in free pronouns, but in (2) from Garifuna and (3) from Ama it is in bound pronouns (pronominal affixes). While Garifuna has bound pronouns indexing subject, Ama has bound pronouns indexing absolutive (S, P and R[ecipient]). Hausa in (4) marks pronominal gender mainly on aspect words, a kind of auxiliary that is preposed to the verb, but also has optional free pronouns.

(3) Ama (Arai/Left May; Matth. 15:26–27): gender marking on bound pronouns (S, P only)
[…] no-na-ni *imo na* *i-****so****-ki,* *Isiso mo. Ulai*
that-FOC-here talk FOC say-O.3SG.F-REM.PST Jesus TOP but
no-na-ni *nukonu* *mo na* ***imo****-ki,* *"…"*
that-FOC-here woman.SPEC TOP FOC say[O3SG.M]-REM.PST
'But he answered ("**to her**"): "…" But she said ("**to him**"): "…" '

(4) Hausa (Afro-Asiatic; Matth. 15:26–27): gender marking on aspect words
Ya *amsa* ***ya*** *ce: "…" Sai* ***ta*** *ce: "…"*
3SG.M answer PST.3SG.M say then PST.3SG.F say
'But he answered: "…" But she said: "…" '

However, even if we consider affixes on verbs to be bound pronouns following Kibrik, the search space is not restricted to pronouns. Many languages have anaphoric forms intermediate between nouns and pronouns, for which I will use the name "GRAMMATICAL ANAPHOR" in want of a better term. Third person pronouns are, of course, also grammatical and anaphors, but since pronoun and third person pronoun are established terms, there is little risk of confusion. A grammatical

anaphor is illustrated in (5) from Kiribati. Kiribati has a personal pronoun not distinguishing gender (*e* 3SG), but there is also the "person demonstrative" (Trussel 1979: 176) *neierei* 'that woman', which is a noun phrase and displays the word order of a full noun phrase (VOS), but is different from the full demonstrative noun phrase *te aine arei* [ART woman DEM.DIST] 'that woman' and does not contain the noun *aine* 'woman'. Kiribati *neierei* (70 times in the N.T.) mostly translates to 'she' and can also pick up reference to *teinaine* 'girl' and *tina-* 'mother' whereas *te aine arei* (13 times) [ART woman DEM.DIST] translates to 'the woman'.

(5) Kiribati (Austronesian, Micronesian; Matth. 15:27): grammatical anaphor
Ao e taku ***neierei*** *…*
and 3SG say **that[DIST].woman**
'But she said: "…" '

Grammatical anaphors, such as Kiribati *neierei* 'that[DIST].woman', are less grammaticalized than pronominal gender markers such as English *she*. Grammatical anaphors tend to be INCIPIENT GENDER MARKERS, nouns on their way to be grammaticalized to pronominal indexes.

One possibility of interpretation is to argue that pronominal gender is more MATURE than non-pronominal gender in anaphors. Mature phenomena imply some sort of non-trivial historical development (Dahl 2004: 2; Trudgill 2011). Pronouns often differ from nouns in being suppletive according to grammatical relation. English *she* (subject) and *her* (object, indirect object, possessor) illustrate this point. Nouns are not entirely precluded from suppletion according to grammatical relation, but such suppletion in nouns is rare. Free and bound pronouns, however, usually display some sort of suppletion and/or neutralization according to grammatical relation. In Ama (3), gender is distinguished in S, P and R, but not in A. Suppletion or neutralization in pronouns can be viewed as a feature of complexity and a feature of maturity.

Another possible interpretation is that gender cumulating with case (grammatical relation), as it often occurs in free and bound pronouns, is a different kind of phenomenon. Wälchli (2019 [this volume]) argue following Nichols (1992: 142) that agreement (and notably agreement in case and number) often triggers noun classification rather than vice versa. Put differently, at least in some instances, gender originates from case, and gender then tends to exhibit particular cumulation patterns with case from its very origin. While, following the second interpretation, cumulation and/or neutralization in certain grammatical relations might be incipient within gender, it is still mature in the sense of grammaticalization, as the development of gender then draws on preexistent grammatical categories (case, number and person).

In this paper I will extract feminine gender grams from translations of the New Testament (N.T.). Translations of the N.T. are parallel texts, and parallel texts allow us to define a semantic core in a very simple manner as a set of aligned passages. The N.T. comes segmented in chunks slightly larger than sentences (so-called verses), which is why no sentence alignment has to be made. The N.T. is translated into many languages and many translations are available electronically. Working with unannotated translations from many languages has the advantage that larger samples than usual can be used and that the dependence on individual grammar writers' reporting or not reporting relevant characteristics is reduced. The most important advantage, however, is that working with automatic extraction forces us to formulate a fully explicit PROCEDURAL DEFINITION of the wanted category, which is then applied in exactly the same way to all languages considered. In particular, the heuristic potential of automatic extraction is invaluable. The automatic device is naive and does not have any preconceived opinions about what kinds of markers should be included or not. In this particular study, this has helped me find various non-mature gender grams which have been overlooked in the gender literature so far, such as Kiribati (5).

The procedural definition of the feminine gender gram will be discussed in more detail in §2. It has essentially two components: (a) finding markers associated with a semantic core in a FUNCTIONAL DOMAIN (Givón 1981) and (b) filtering out markers which are also associated with other semantic cores (notably masculine gender and female light nouns such as 'woman', 'girl' and 'mother'). Despite differences concerning parts of speech (pronouns, verbs, auxiliaries) and grammatical relations (A, S, R, P) exhibiting or not exhibiting feminine gender, all languages exemplified in (1–5) mark feminine gender in the same context in the parallel text corpus. The markers all occur in the same functional domain. Nothing in the procedural definition is in any way related to the notions of noun class and agreement. This means that if the endeavor is successful, it is possible to define feminine anaphoric gender grams in abstraction from the notions of noun class and agreement.

What does all this mean for the understanding of gender? Corbett's Agreement Hierarchy is evidence that there is a semantic pole (anaphors) and a syntactic pole (NP-internal agreement) in gender. Traditional research focusing on noun classes and syntactic agreement considers the syntactic pole to be basic. This culminates in the Canonical Approach to gender, which focuses on gender values of nouns and considers redundant gender marking and local agreement domains to be canonical (Corbett & Fedden 2016). In this paper I argue that a shift of perspective is possible where semantic and referential gender in anaphora is the primary concern of grammatical gender, whereas syntactic, lexical and redundant gender is secondary.

The following sections are structured as follows. §2 motivates and formulates the procedural definition of the feminine anaphoric gender gram and §3 discusses its practical implication in the parallel text corpus and reports the results. §4 elaborates on the distinction between mature and non-mature grams and how it is related to grammatical relations. §5 focuses entirely on those non-mature gender grams that are non-pronominal and arguably incipient anaphoric gender markers. Finally, §6 discusses how the functional approach developed in this paper can be connected to the traditional system perspective on gender and §7 concludes this paper.

2 A procedural definition of the feminine anaphoric gender gram

2.1 Overview

This paper focuses on a domain where gender is most obviously used semantically and which is easiest to address by automatic extraction in the N.T. corpus. In §2.2 I am going to discuss why feminine is easiest to address. I will then discuss why feminine anaphoric can be viewed as a functional domain which can be defined as a set of passages in the parallel text corpus (§2.3). The next step is to discuss what makes markers of feminine gender differ from other markers closely associated with the feminine anaphoric functional domain (§2.4). This will allow us to formulate a procedural definition of the feminine gender gram which is sufficiently elaborate for the purposes of this paper. Finally, based on the notions of cue validity and constructional islands, §2.5 discusses why anaphoric gender grams in most languages are accessible without previous familiarity with the entire language system.

2.2 Why feminine, why singular and why anaphora?

We know from Corbett's Agreement Hierarchy that the semantically most transparent use of gender is found in third person anaphoric pronouns. However, this does not mean that grammatical gender has the function of reference tracking in discourse.[2] Within anaphoric use, the descriptive content of gender is most active in contrastive use in implicit or explicit focus (Bosch 1988: 227; Seifart 2018:

[2] According to Kibrik, gender is used as a deconflicter in reference tracking in an "opportunistic way" Kibrik (2011: 359). Languages rely on referential aids to various extent and some languages without gender such as Navajo (Na-Dene) are more strongly inclined to use reduced referential devices than some languages with gender such as Archi (Nakh-Daghestanian) (Kibrik 2011: 336).

25), and contrastive use ('but she') is represented in the clauses selected for the extraction from the corpus as in (1). Since gender is often neutralized in the plural (even though this is no strict universal, see Plank & Schellinger 1997), the search space is restricted to singular. The most widespread meanings in gender grams are animacy and sex. Sex is easier to identify than animacy, since animacy comes in many different forms in grammatical markers, not only as gender feature, but also as condition on gender (Corbett 2006, chap. 6) and is, among other things, also involved in the choice of case or adposition in differential object marking (Croft 2003: 166). This leaves us with masculine (singular) and feminine (singular) as possible choices. In the N.T. corpus, feminine is the much easier choice. Reference to male beings is strongly overrepresented in this text, which makes it difficult to distinguish between third person masculine and third person in general in automatic extraction. A further complication in this particular text is that the distinction between male and deity is fuzzy, which, in many languages, calls for specific solutions where this distinction is relevant in grammar. Thus, feminine singular in the anaphoric domain is clearly the easiest option to choose.

2.3 Feminine anaphoric as a functional domain

Defining feminine anaphoric gender as a functional domain in parallel texts means identifying a set of passages where this function is expressed recurrently across all translations of the text. Such a passage is Matthew 15:27, which has been illustrated from various languages in Section 1 and which is for convenience repeated here in English in (6).

(6) English (Indo-European; Matth. 15:27)
*But **she** said: "..."*

Saying that (6) reflects the feminine anaphoric functional domain abstracts from the fact that this passage is related to another passage earlier in the text given in (7). In (7), the referent of the anaphor in (6) is introduced in the form of an indefinite noun phrase.

(7) English (Indo-European; Matth. 15:27)
*And behold, **a Canaanitish woman** came out from those borders...*

Another way to put it is that anaphors tend to be coreferent with full noun phrases introduced earlier in the text, which is not strange given that anaphora "is the phenomenon whereby one linguistic element, lacking clear independent

reference, can pick up reference through connection with another linguistic element" (Levinson 1987: 379). However, this does not mean that all anaphora have explicit antecedents with which they are exactly coreferent, as illustrated in (8).

(8) Anaphors without explicit antecedent (Hintikka & Kulas 1985: 98):
A couple was sitting on a bench. ***He*** *stood up and* ***she*** *followed* ***his*** *example.*

Not only pronouns, but even full NPs can be used in anaphoric function, and third person pronouns and full NPs have very similar properties in anaphoric function as shown, in (9). Pronominal anaphors and definite NPs can both be used to make attributions of gender and neither of them requires a syntactically explicit antecedent, but they are both definite expressions.

(9) Pronouns and full NPs in anaphoric function (Hintikka & Kulas 1985: 98):
   a. *The teacher addressed the children.* ***He/The man*** *was stern.*
   b. *A couple was sitting on a bench.* ***He/The man*** *stood up and* ***she/the woman*** *followed* ***his/the man's*** *example*

However, when assembling a set of passages expressing feminine anaphoric in a parallel text corpus, it is possible to abstract from the fact that most anaphors have NP antecedents and that a lexical item in the NP can determine the gender value in a way that goes against the core meaning of gender.

2.4 Filtering out markers of feminine gender grams from the feminine anaphoric functional domain

All languages have some anaphoric expressions in the feminine anaphoric domain, but not all expressions are grammatical expressions and not all grammatical expressions are feminine. The anaphoric expressions in the feminine anaphoric domain can be nouns, such as 'woman' or 'girl', or they can be pronouns not distinguishing gender. This is both illustrated in (10) from Turkish with the noun *kadın* 'woman' and the general third person pronoun *o* 'he/she'.

(10) Turkish (Matth. 15:24–27)
İsa*, «…» diy-e cevap ver-di. Kadın ise yaklaş-ıp, «…»*
Jesus say-CVB answer give-PST3 woman however approach-CVB
diyerek […]. İsa o-na, «…» de-di. Kadın, «…» de-di. «…»
say-CVB Jesus 3SG-DAT say-PST3 woman say-PST3
'But **he** [=Jesus] answered and said, "…" But **she** [=the woman] came […] saying, "…". And **he** [=Jesus] answered (to **her**) and said, "…". So **she** [=the woman] said, "…" '

It is thus not all expressions in the functional domain of picking up reference to female humans that instantiate feminine gender. If we extract the forms which are associated with the feminine anaphoric domain, which can easily be done by means of collocation measures (see §3), the recall will be too large. Put differently, many nouns, such as Turkish *kadın* '(a/the) woman', and general anaphoric pronouns, such as Turkish *o* 'he/she', will be extracted as well. One way to account for this is to define the search domain very narrowly by excluding such contexts where many languages use nouns instead of pronouns. But cross-linguistic and stylistic differences in the use of nouns, pronouns and zero anaphors are so large that a restrictive search domain is not sufficient.

The solution which is chosen here is to filter out expressions such as Turkish *o* 'he/she' and *kadın* '(a/the) woman'. By subtracting forms associated with anaphoric masculine and anaphoric in general, we can make sure that none of the extracted forms is third person masculine or third person general. Expressions for 'woman' have their own functional domain, which only marginally overlaps with the feminine anaphoric. Notably they also contain non-anaphoric uses, such as (7), where languages such as English have an indefinite article. Lexical nouns are not restricted to anaphoric uses, but can occur both in definite and indefinite uses. By subtracting all forms associated with the functional domain '(a/the) woman' from the set of forms associated with the feminine anaphoric we can make sure that none of the extracted forms means '(a/the) woman'. The same procedure can be applied to a few other critical lexical domains, such as 'girl' and 'mother'. Nouns are an open word class. Hence, the number of potential female lexical domains is potentially infinite. However, there is no need to care about rare lexical domains. It is sufficient to address the most frequent ones: 'woman', 'girl', 'mother', and 'daughter'. This is sufficient for the particular parallel corpus used. If in another parallel corpus another female lexical domain would be particularly frequent, it would have to be included in the filter as well. Filters must be adjusted to particular parallel corpora. However, their content can be described in general terms in the procedural definition: "frequent female lexical domains".

Filtering out all forms that might be associated with a lexical domain, we can also make sure that the remaining set of forms consists exclusively of grammatical markers. This does not restrict the set to pronouns. Grammatical anaphors, such as *neierei* in Kiribati (5), will still be included.

What has been said above, results in the procedural definition for feminine anaphoric gender grams given in (11):

(11) Procedural definition of feminine anaphoric gender markers:

a. Extract all markers picking up reference to female humans
b. unless they can also be used to pick up reference to male humans, and
c. unless they express frequent female lexical domains (such as 'woman', 'mother', 'girl', and 'daughter')

The concrete implementation of this definition is discussed in §3.

2.5 Constructional islands and cue validity

The approach implemented in this paper rests on the assumption that markers expressing a grammatical or lexical meaning X can be viewed as constructional islands with high cue validity. I take these terms from the literature on first language acquisition (Tomasello 2003: 113). In general terms, constructional islands can be defined as utterance-meaning pairings, where one part of the utterance, the marker, is constant, such as in the set: ***more*** *milk*, ***more*** *grapes*, ***more*** *juice*. The marker has high cue validity, if it is sufficiently distinct from all other markers in the language and if it can be immediately recognized without any previous analysis of the morphology of a language, simply as a continuous sequence of sounds (a word form or a continuous segmental morph without allomorphs).

The notions of constructional island and cue validity can be directly applied to parallel text corpora, where a constant meaning can be defined as a set of passages in which a meaning is instantiated. In written corpora we have to take continuous sequences of characters instead of phonemes. All word forms and all continuous substrings of words are candidates for markers that are directly accessible without any previous analysis of the language system. Constructional islands with high cue validity can be detected in the corpus without any knowlegde about the structure of a language and without any resort to parts of speech, grammatical or lexical categories, paradigms or systems.

My assumption is that if a language has a feminine anaphoric gender gram, there will usually be at least one marker with high cue validity. Not all markers will have high cue validity, so the extraction will not be complete. But the approach will be sufficient in most cases for finding out whether or not the language has a feminine anaphoric gender gram. For this purpose, it is sufficient to find one marker if there is more than one.

Put succinctly, if there is no gram, no marker is detected, if there is a gram, at least one of its markers is extracted.

There may be languages where the cue validity of anaphoric gender grams is low, where gender is highly integrated in grammatical systems. These may be cases where the marker is short (just a single phoneme within words of a particular word class) and often neutralized or where the marker is zero (as opposed to a non-zero masculine marker). However, my assumption is that in the vast majority of languages, feminine gender grams have high cue validity and can be viewed as constructional islands, at least to some extent.

3 Extracting feminine gender grams from parallel texts

3.1 Sample, data, and procedure

The sample consists of 816 languages (listed in Appendix A and B) and is not stratified. It simply contains the languages for which I happened to have an electronic version of the New Testament available when I started this work, and, as in other work based on Bible translations, some areas, in particular North America and Australia, are strongly underrepresented. The texts are not annotated. Some texts which are not in Roman script have been Latinized, but differences in writing systems have very little impact on the extraction procedure. Where the writing system is relevant, this is discussed below. For a few languages, more than one translation has been used (a total of 858 texts). The differences within languages are not reported, since in most cases the results were largely constant within a language,[3] but this does not change the fact that the translations represent particular varieties (doculects), and in a few cases there may be intra-language variety that has not been detected. In one case, Uduk, feminine anaphors have been deliberately created by missionaries (see §5.1), but language planning is an issue only for few languages of the sample, which is why it is not excessively discussed in this paper.

While the theoretical notion of procedural definition of a category type (11) is very general, there are several practical details in the extraction process that can be adjusted and must be adjusted (see below). As is usually the case in typological investigations, there is no gold standard. It is not known what the result is going to be before the investigation has been carried out. Hence, the automatic extraction must be complemented by an evaluation by means of grammars and other reference material. However, since grammatical gender is known to be genealogically stable in many language families, it was very useful to have a large

[3] There are some minor differences as in German where the form *ihr* [3SG.F.DAT] is not extracted in some texts.

number of languages from a few large families in the sample. I expected feminine anaphoric gender to be lacking in most languages of the following families: Austronesian (134 lgs.), Niger-Congo (127 lgs.), Trans-New Guinea (90 lgs.; except Ok-branch known to have gender), Quechuan (25 lgs.), Sino-Tibetan (24 lgs.), Uto-Aztecan (18 lgs.), Turkic and Uralic (17 lgs.), and to be present in most languages in the following families: Indo-European (50 lgs.; except for some Indo-Iranian languages and Armenian known to lack gender), Arawakan (17 lgs.) and Tucanoan (13 lgs.). This means that for roughly two thirds of the sample there was an expected result and the details of the extraction mechanism (set of verses included in the search space, filters, how to compare a filter with the search space, see below) could be adapted in a process of trial and error until the outcome largely matched the expected result. In practice, the most difficult thing was to avoid extraction of forms in languages without anaphoric gender grams, so it is very important that the sample contains a large number of such languages (Appendix B). This means that only about a third of the languages of the sample had to be checked manually with grammars and other reference material. Hence, due to its genealogical stability, gender is an exceptionally favorable domain for a typological investigation based on parallel texts with many languages.

In the course of investigation it then turned out that in several dozens of languages the results yielded other forms than just the expected third person free and bound pronouns even after the necessary practical adjustments in the algorithm. At closer introspection, it became clear that many of these languages had incipient anaphoric gender; put differently, anaphoric gender that is so simple that it has not figured prominently in the literature on gender so far, which traditionally focuses on complex cases of gender. This made it necessary to devote a large part of this paper to languages with incipient gender (§5) and these languages also turned out to be typical exceptions to the expected genealogical stability of gender. The rest of the unexpected forms could be accounted for as various types of systematic errors due to the naive mechanic nature of the extraction algorithm (§3.3).

3.2 Extract all markers picking up reference to female humans

The starting point for the extraction of feminine gender from the N.T. parallel corpus is the procedural definition in (11).

First, the algorithm extracts markers picking up reference to female humans, based on collocation with a set of contexts where feminine anaphoric gender occurs.

In parallel texts, meaning can be equated with a set of contextually embedded situations where the markers encoding that meaning (which are language-particular form classes) are expected to occur (Wälchli & Cysouw 2012: 672). In order to identify the situations across translations into different languages, the texts must be aligned with each other on a level coming close to sentences (sentence alignment). The N.T. is aligned in verses and verses are often somewhat larger than sentences, but verse alignment comes close to sentence alignment. Extraction is much easier if the texts are also word-aligned, but here I use only verse alignment which is a crude approach.

For the sake of simplicity it is assumed that a marker is either a word form or a morph (a continuous part of a word form; in concrete terms, any continuous sequence of characters in a word form). This makes it possible to explicitly define the set of potential markers as all word forms and all continuous sequences of characters within word forms.

The easiest way to design a search domain is to take one or several SEED GRAMS (Dahl & Wälchli 2016), forms from particular languages where it is known that they more or less accurately instantiate a gram. Such forms are the third person singular feminine personal pronoun forms in English (*she/her*) or in Scandinavian languages (Swedish *hon/henne/hennes*). The English forms *she* and *her* occur together in 292 verses in the N.T. (American Standard translation). An extraction of potential markers is nothing else than a list of the word forms and character sequences (approximating morphs) that collocate best with the search space above a certain threshold with an appropriate collocation measure. If these 292 verses are used as a search space, an extraction of collocating forms will contain many of the wanted markers, but it will also contain many forms that should not be extracted (boldface in Table 1).

A good extraction must meet two conflicting criteria. There should be as many correct extracted forms as possible (high recall), but there should also be as few wrongly extracted forms as possible (high accuracy). Since the majority of languages in the sample lack feminine gender grams, high accuracy is not as trivial as it might seem at first glance.

There are three ways to improve accuracy: (i) We can use a higher threshold, but this is no good solution, since it has devastating effects on the recall. (ii) We can filter out wrongly extracted forms, since they can be grouped according to certain meanings which we can search for as well, such as 'woman' or general third person singular. (iii) We can reduce the search domain, so that the conflicting meanings are removed from it.

After many attempts I have decided to use a combination of (ii) and (iii). Probably it would be possible to work with the 292 verse search space and filtering, but I have not managed to design the filters such that the extraction is optimal.

Table 1: Word forms and morphs best collocating with English *she+her*. Here and elsewhere the notation >x< will be used for morphs and # is used for word boundaries.

Language	Forms	Gloss of forms in boldface
Turkish	**kadın**	'woman'
Swedish	hon, henne, hennes, **kvinna**	'woman'
English	her, she, **woman**	'woman'
Koine Greek	αυτης, αυτη, >σα#<, **γυνη**, **η**	'woman', DEF.NOM.SG.F
Estonian	**naine**, **ta**, **tema**	'woman', 3SG, 3SG.EMPH
Tok Pisin	**meri**, **en**, **maria**	'woman', 3SG, 'Mary'
Indonesian	**perempuan**, >**nya#**<	'woman', POSS.3SG

In the best attempt, there are wrongly extracted forms in 33 more languages and 10 languages are lost in comparison to the extraction reported here. The larger the search space, the more sophisticated the filters have to be. In larger search spaces there are simply more meanings represented and there is more that can go wrong.

In the extraction reported in this paper I have used a subset of 74 clauses as search space. The clauses have been selected manually, but more important than which clauses are selected is the simple fact that the set has about that size. If smaller sets are chosen it is increasingly more difficult to extract short bound morphemes, such a Garifuna >#t-< in (2). Explicit marking of word boundaries by a character makes peripheral morphs more salient and easier to extract.

The following criteria have been used to select the 74 clauses.

(i) Include verses where feminine anaphoric gender is instantiated several times, for instance, as in (12):

(12) Two of 76 verses of the trigger domain (given in the English Lexham translation)
42015009 (=Luke 15:9) And when **she** has found it, **she** calls together **her** friends and neighbors, saying, 'Rejoice with me, because I have found the drachma that I had lost!'
44016015 (=Acts 16:15) And after **she** was baptized, and **her** household, she urged us, saying, "If you consider me to be a believer in the Lord, come to my house and stay." And **she** prevailed upon us.

(ii) exclude long verses (where many other meanings are expressed);

(iii) exclude clauses containing words for 'woman' in most texts;

(iv) exclude most verses where feminine anaphoric gender is contrastive ('but she'), because many texts have nouns for 'woman' there;

(v) exclude verses with 'Mary', so this proper name need not be filtered, and

(vi) exclude (as far as possible) clauses with masculine anaphoric contexts (in fact, this cannot be strictly implemented, because masculine anaphoric contexts are omnipresent in the text).

This results in a set of 74 verses[4] two of which have been illustrated in (12). Choosing the verse (or sentence/clause) as unit of alignment has an important consequences for the extraction of gender. It is not easily possible to distinguish between different grammatical relations, since the same verse often contains the feminine gender gram in various functions. This is notably true of reflexive possessors (as in *she calls together **her** friends*) where even the clause is too large as a unit of alignment. Thus, the extraction applied here is not helpful in deciding which grammatical relation a marker encodes; only that it is some sort of feminine gender marker. The classification of markers according to grammatical relations in Appendix A has therefore been made manually with the help of reference grammars.

Furthermore, it needs to be pointed out that the N.T. is a text where feminine anaphoric gender is strongly underrepresented. Together with the considerable number of verses that have been excluded, this results in a quite small search domain, less than 1% of the text. However, there are enough examples in the text for a mostly correct automatic extraction of frequent feminine gender grams, even if this sometimes means that only some, not all, markers of a feminine anaphoric gender gram are extracted. Extraction works quite well, despite the fact that the algorithm used here is crude. This testifies to the high cue validity of feminine

[4] 40001019 (i.e., 40 1:19 or Matth. 1:19; Matthew is the 40th book in the Bible), 40002018, 40008015, 40009025, 40012042, 40014008, 40014011, 40015023, 40015027, 40026012, 41005042, 41006024, 41006025, 41006028, 41007030, 41010004, 41014005, 41014006, 41014008, 42001029, 42001035, 42001036, 42001057, 42001058, 42001061, 42002006, 42002007, 42002036, 42002037, 42002038, 42007013, 42007035, 42007038, 42007047, 42008054, 42008055, 42008056, 42010040, 42010041, 42011031, 42013012, 42015009, 42018005, 42020031, 43004013, 43004016, 43004026, 43008005, 43011023, 43011033, 43011040, 43012007, 43019027, 43020014, 43020017, 44005008, 44005009, 44005010, 44009037, 44009040, 44012014, 44016015, 44016019, 44019027, 45007003, 45009012, 45016002, 46007028, 54005010, 58011031, 59002025, 66002021, 66002022, 66021011.

gender markers. Put differently, in most languages identifying feminine gender grams is not particularly complex and does not presuppose any knowledge about gender systems.

The algorithm goes through all candidates and checks which of them matches best with the trigger domain according to a collocation measure (here T-score as defined by Fung & Church 1994 is used) above a certain threshold. The threshold is determined empirically so that no or few incorrect forms appear. In order to demonstrate that this can be done in slightly different ways, two different thresholds have been applied: t = 3.4 and t = 3.19. The higher threshold prevents the first entirely wrong form to be extracted (Buglere *chku* [arrive:PFV] 'arrived'). However, with the higher threshold we also lose three languages with a feminine gender gram: Kabyle, Angami Naga and Owa (Owa is actually a borderline case, see 5.4), but there are also a large number of arguable errors among the 44 forms that are not extracted with the higher threshold. Since many errors are very interesting from a methodological point of view, I have chosen not to use only the higher threshold, which would probably have been the most reasonable thing to do for an optimal extraction. Forms only extracted with the lower threshold are given in curly braces in Table 2 and in Appendix A.

Table 2: Selected languages where feminine anaphoric gender markers have been extracted

Language	Extraction	T-value of first form
Akateko (knj)	[ix]1	7.682
Ama (amm)	[isoki]1	4.113
Carapana (cbc)	[cõ]1 [>upo#<]2 [>ñupõ#<]3 [>mo#<]4	7.738
English (eng) [amstd]	[her]1 [she]2	7.2
Garifuna (cab)	[>#t<]1	6.008
Hausa (hau)	[ta]1 [>ta#<]2	5.309
Kaingang (kgp)	[fi]1	7.636
Latvian (lav)	[viņai]1 {[>usi<]2 }	4.152
Owa (stn)	{[kani]1 }	{3.191}
Zapotec, Miahuatlan (zam)	{[xa']1 }	{3.310}

Although Indo-European languages have been the starting point for determining the distribution, it is rather languages from other families that have the best extraction values (the top three are Carapana *cõ*, Kaingang *fi* and Akateko *ix*; see Table 2).

3.3 Filtering out conflicting meanings

While the procedure described in §3.2 yields the correct result for most languages with anaphoric gender, the recall is too large in languages where anaphoric gender is lacking. The kind of forms wrongly extracted fall mainly into two semantic groups:

(i) Indexes for third person singular not distinguishing gender. Forms expressing third person singular in general without making a gender distinction also collocate with feminine gender.

(ii) Words for 'woman', 'girl/daughter', and 'mother'. This is surprising at first glance since most texts in Indo-European languages of Europe do not contain instances of 'woman' in the smaller search domain of 74 verses and too few for 'girl/daughter' and 'mother' to be extracted. These "errors" reflect the fact that many translations into languages without feminine anaphoric gender use words for 'woman' in contexts where languages with feminine gender use forms such as *she* and *her*, as in (10) from Turkish. For determining whether a language has feminine anaphoric gender, the procedure must be refined so that such forms are not extracted.

If forms collocating with the feminine third singular also include some forms for third person singular general and some forms for 'woman' and other general feminine nouns, extraction must take this into account by excluding forms which have a better correlation with third person singular masculine and with 'woman', 'girl' and 'mother'.[5] The best way of doing this would be to define sets of verses for all conflicting meanings as carefully as for feminine anaphoric gender. Here, a cruder approach is used where these conflicting domains are simply represented by some characteristic instances in particular languages (Table 3).

(i) *The masculine filter*: For excluding general third person use, a form is not extracted if it correlates better with at least one of the following sets: (a) English *he*, (b) English *him*, (c) all uses of anaphoric masculine singular in English together (*he*, *him* and *his*), and (d) all uses of *said to him*. These

[5]To identify better correlations is not trivial since T-score values with larger search domains are generally higher than with small domains. Since there happen to be roughly two kinds of sizes of domains (smaller than 164 and larger than 742, see Table 3), it is for practical reasons possible to apply a very crude solution by dividing all values of the larger domains by two before comparison. If this correction is not applied, a considerable number of feminine gender markers, for instance those in Kuot and Paumari, are filtered out.

Table 3: Filters in the extraction of feminine gender grams

Masculine filter (relates to (11b))	English *he* [2347 verses], English *him* [1836 verses], English *he/him/his* [3570 verses], English *said to him* [164 verses]
'Woman' filter (11c)	English *woman* [54 verses], Xaasongaxango *muso* 'woman' [39 verses] Yau (yuw) *owi* 'woman, grandmother' [1953 verses]
'Mother' filter (11c)	English *mother* [76 verses]
'Girl' filter (11c)	Nalca *gelma* 'girl, daughter' [40 verses], Upper Pokomo *mwanamuke* 'girl' [41 verses]
'Child' filter	Tok Pisin *pikinini* [743 verses]

four distributions all serve the same purpose, but conflicting forms can have different extensions, so all four of them are needed. Together they constitute the masculine filter.

(ii) *The 'woman', 'mother' and 'girl' filters*: For the exclusion of lexical feminine meanings, a form is not extracted if it correlates better with at least one of the following sets: (a) the English singular form *woman*, Xaasongaxango *muso* 'woman', and Yau *owi*, which is an instance of a very extensive use of a word for 'woman' occurring also in the co-compound *owi amna* [woman man] 'people' (Sarvasy 2014: 104), (c) English *mother*, (d) Nalca *gelma* 'girl, daughter', (e) Upper Pokomo *mwanamuke* 'girl'. This is to make sure that the basic meaning of an extracted form is not 'woman', 'mother' or 'girl' and only incidentally also occurs in the anaphoric domain. Several forms are needed since the semantic extension of words can vary (in some languages 'daughter' and 'girl' is expressed by the same word, in others by different words).

After this is done, a smaller problem area remains which is presented here directly with the remedy resolving it:

(iii) *The 'child' filter*: In a few languages a word for 'child' is extracted. This is because children, child bearing, giving birth to children happens to collocate

with the search domain in the N.T. This is solved by removing all forms that collocate better with Tok Pisin *pikinini* 'child' than with the search domain. This is a practical complication that is so specific that I have not included it in the more abstract procedural definition in (11).

To paraphrase the whole procedure in a simple way: a feminine singular anaphoric gender marker is any form that collocates with the feminine singular anaphoric gender domain unless it rather means third person singular in general, 'woman', 'mother', 'girl, daughter', or 'child'. Put differently, forms collocating with the feminine anaphoric singular gender must pass the masculine, 'woman', 'girl', 'mother' and 'child' filters before it is likely that they really represent the feminine anaphoric gender gram.

If the larger search space of English *she+her* is used, further filters have to be added, notably 'wife', 'husband' and 'Mary' filters. There are also complex adjustments required for comparing T-score values with search spaces of different magnitudes.

3.4 Unexpected extracted forms and whether they are errors

Since there is no gold standard, extracted forms were checked with grammars and dictionaries. Checking revealed that after markers with conflicting meanings have been removed by filtering, there remain some unexpected extracted forms which could be considered errors. However, almost all "errors" are highly interesting in that they are somehow associated with the meaning of the feminine anaphoric gram. They fall into five types:

(a) anaphoric (demonstrative or definite) forms of a word for 'woman',

(b) demonstrative pronouns,

(c) person name markers (determiners or titles), mostly female person name markers,

(d) gender markers within noun phrases, and

(e) the masculine gender form by female speakers.

Finally, four occasional forms for 'woman', third person singular personal pronouns, and an entirely occasional verb form meaning 'arrived' escaped filtering with the lower threshold.

(a) *Anaphoric (demonstrative or definite) forms of a word for 'woman'*: In South Tairora the form *nraakyeva* [*nraakye-va* 'woman-DEM'] is extracted, because the naive algorithm cannot recognize that it contains *nraakye* 'woman' and should therefore be removed by the 'woman' filter. In South Tairora demonstrative NPs are formed by a free demonstrative, *mwi*, *mwa*, or *mwatai* in the N.T. text, followed by a noun with an obligatory *-va* suffix (Vincent 2010: 584). The form *nraakyeva* has the correct distribution since it only occurs in the feminine anaphoric domain; it is not a general form for 'woman' and passes therefore the 'woman' filter. This error thus derives from the fact that the algorithm applied here does not have the capacity to segment word forms into morphemes. Extracted forms with the same kind of error include Sabaot *(:)cheebyoosyaanaa* 'this woman', Endo *cheepyoosoonoonēē*, Ayautla Mazatec *chjunbiu*, Safeyoka (Wojokeso) *a'musi*, Umbu-Ungu *ambomo*, and Rawa *barega* (see Appendix A IV). Several similar forms are slightly below the lower threshold for extraction, such as Low Tarahumara (*muki-ka* 'woman-EMPH') and Auhelawa (*waihi-una-ne* woman-DEM/DEF'). Also Ama *nukonu* [woman.SPEC] (see (3)) sorts here, with an irregular form of the specifier (suffix *-ta* in other nouns; Årsjö 1999: 92); however, this form is not extracted.

Generally, a demonstrative or definite form of 'woman' tends to be extracted whenever the demonstrative or definite marker is synthetic. This kind of error is particularly instructive because it shows us how anaphoric gender markers may emerge. Expressions for 'that woman' may qualify as anaphoric gender markers to the extent that the noun and demonstrative have become opaque. This is exactly what has happened in languages with non-compositional complex NPs such as Japanese (see §5.2). The errors made by the computer derive from the fact that more forms are opaque for the computer than for humans.

(b) *Demonstrative pronouns*: Since complex expressions of 'that/the woman' are common feminine anaphoric expressions, it is not entirely unexpected that demonstratives and articles are occasionally wrongly extracted. This happens in several Trans-New Guinea languages such as Mountain Koiali *ke-u* [that-SUBJECT] (Garland & Garland 1975: 428; in the N.T. in *keate keu* 'woman that', *ma keu* 'girl that'), Folopa *kale* 'the' (Anderson 1989: 85; in *kale so[-né]* 'the woman[-ERG]'), Fore *kana-* 'this mentioned one, the aforementioned' (Scott 1989: 45), and Awa *mi* 'that' (Lowing & Lowing 1975) (Appendix A VI). I have not tried to add a demonstrative filter because

demonstratives are too different in their distribution from each other and there is no point in adding filters that remove just one or two problematic cases.

(c) *(Female) person name markers*: It is not uncommon for anaphoric gender markers to also be used together with person names. In a few languages the form is slightly different, thus Kiribati uses *Nei* as a female person name marker and *neierei* as anaphoric gender form. In North Halmaheran languages of the sample female names contain a form *ngo*, which combines with the general determiner *o*. A few languages in the sample have female person name markers but lack anaphoric gender. If the language at the same time happens to use many person names in the anaphoric domain, the person name marker can be wrongly extracted (Appendix A III). This is the case for Iraya *bayi* (probably a shortening of *babayi* 'woman'), Uab Meto *bi*, Satere-Mawe *mana*, and Huave *müm*.

(d) *Gender markers within noun phrases are special cases of (b) and (c)*: demonstratives or extended person name markers that happen to bear NP-internal gender. In a sense these are not errors, since the forms mark feminine gender, but they mark feminine gender only NP-internally with common nouns and person names. This holds for Abau (*sokwe* [DIST.DEM.F.OBJ]; Lock 2011: 87), where there are also correctly extracted anaphoric forms, and for Kadiweu, Mocoví, and Nalca.

The Guaicuruan languages Kadiweu and Mocoví have so-called local classifiers (standing, sitting, lying, coming, going, absent; Sándalo 1997: 62) in attributive demonstratives, which combine with masculine and feminine gender markers. In both languages only the form with the 'going' classifier is extracted: Kadiweu *naɣ-a-jo* CLOSE-F-going and Mocoví *a-so'-maxare* F-GOING-PRO (Appendix A VI).

Nalca (Mek, Trans-New Guinean) has developed a gender system from person name markers (Wälchli 2018), and the female person marker *ge-* grammaticalized from *gel* 'woman' has extended also to some female kinship terms and the word for 'woman'. The extracted form is the topic form *gera* [F-TOP], which occurs in the search domain 15 times with female person names, 12 times with *gel* 'woman' and twice with two different words for mother (Appendix A III). In the whole N.T. this form is only used once anaphorically, but not within the search domain.

It is not unexpected that some NP-internal non-feminine anaphoric gender forms, as in Abau, Kadiweu, Mocoví, and Nalca, are extracted by the algorithm, because, as far as anaphoric NPs occur in the search domain, they have the right distribution and are not filtered since they are both dedicated to feminine and non-lexical.

Some languages have derivational noun suffixes in female nouns, such as Parecis *-halo*, Esperanto, and Iraqw *o'o* (Mous 1992: 63). The Iraqw form is not extracted, the Esperanto form is eliminated by the 'woman' filter and the Parecis form is eliminated by the 'girl' filter.

(e) *Masculine gender for female speakers and second person feminine*: Kayabi (Tupian) distinguishes both speaker and referent gender (see §4). The verses of the search domain happen to contain a considerable number of quotations from female speakers which are basically useless for the extraction of the feminine gender gram. While the quotations do not do any harm for most languages, for Kayabi they cause with the lower threshold the error that *kĩã* 'M 3SG (female speaker)' is wrongly extracted. Also due to direct speech in the search domain is the extraction of Mwaghavul *yi*, a form for second person feminine reference, even this only with the lower threshold.

Finally, the most problematic wrongly extracted forms are four forms that escaped filtering. But three of them are extracted only with the lower threshold t=3.19. One form for 'woman' Ama *iní* 'woman' escaped filtering (Appendix A V). General third person pronouns in two Zapotecan languages were wrongly extracted (Appendix A VII). In Chichicapan Zapotec *bi* is opposed to third person respect *ba* (Benton 1975) and escapes the masculine filter, probably because Jesus is referred to with the respect form. For Chichicapan Zapotec *bi* even using the higher threshold does not help; the T-value is high (t=5.24). Miahuatlan Zapotec *xa'* is another general anaphoric marker for third person (both masculine and feminine) that happens to have escaped filtering with masculine domains. These cases show that filtering is not always reliable, especially if forms for 'woman' and general third person singular deviate from their expected distribution in the text. Finally, as mentioned in §3.2, Buglere *chku* 'arrived' is the first fully unsystematic kind of error at t=3.39.

3.5 Languages where the automatic extraction fails to detect gender

Languages that have gender but where it is not extracted can be ordered into the following groups:

(a) There is agreement gender or there are noun classifiers reminiscent of agreement gender within the NP, but no or virtually no anaphoric gender: Limbu (van Driem 1987: 21), Baruya, Biangai, and Mopan Maya (Contini-Morava & Danziger 2018) (for Nalca, Kadiweu and Mocoví, see §3.4 above).

(b) Gender is distinguished in pronouns, but only in the second or in the second and first persons: Basque, Paez (Jung 2008: 136, first and second person, but not third person) and Iraqw. However, in Mwaghavul some second person singular form *yi* has been wrongly extracted, since second person with female referent often occur in direct speech in the search domain (see §3.4 (e)).

(c) There is feminine anaphoric gender, but it only covers the domain of girls or young women, the adult women domain is covered by a general human respect gender: Coatzospan Mixtec and Texmelucan Zapotec. These are removed by the 'girl' filter. The 'girl' filter is also responsible for eliminating the reduced nominal anaphor *tạhn* in Teutila Cuicatec. In Tlalcoyalco Popoloca the anaphoric forms generally correspond to specific feminine lexemes and are therefore filtered out (see §5.3) by the 'woman' and 'girl' filters. A more problematic case is Southern Puebla Mixtec, where the gender marker has many allomorphs (*-nè*, *-ne*, *-né*, *-ñá*, *-ña* F), and the only one that is detected happens to be removed by the 'woman' filter.

(d) Gender marking is restricted to a limited part of the S and P domain and the markers do not have high cue validity: Chechen, Hindi, Gujarati, and Eastern Panjabi. These are languages with feminine genders, but the anaphoric function in those languages is marginal or non-existing. In Avar only gender on free pronouns is detected.

(e) The marker is partly zero as opposed to a non-zero masculine marker: This holds for the Arawakan languages Ashéninka Pajonal, Asháninka, Caquinte, Pichis Ashéninka and Nomatsiguenga. The algorithm as implemented here is simply not smart enough for recognizing zero as the marker of the feminine gender gram. The recognition of zero morphemes requests some understanding of systems or at least oppositions.

(f) Gender is too inconsistently marked to be extracted: In Iraqw (Cushitic, Afro-Asiatic; Mous 1992), masculine and feminine are not distinguished in third person free pronouns, and in affixes in verbs and auxiliaries, the markers are manifold both for the expression of subject and object (e.g.,

ó' 'she said' vs. *óo'* 'he said'). It is not possible to detect feminine marking as constructional island without previous analysis of the paradigms. The algorithm fails to detect feminine anaphoric gender in Iraqw.

(g) The dominant marker is orthographically identical with another form: Teutila Cuicatec.

The types (a) and (b) are no real errors since the algorithm only extracts feminine anaphoric gender in third person. The cases in (c) are too weakly grammaticalized or do not have general feminine gender grams, and can therefore not really be counted as errors. The cases in (d) are errors, but these are all languages where anaphoric gender has a very weak functional load. In Chechen only a small proportion of verbs have a feminine prefix *j-*[6] in S and P. In Hindi and some other Indo-Aryan languages, some verbs in some tenses have a feminine singular suffix *-ī*, not restricted to third person. The errors in (e) are due to the unsophisticated design of the algorithm that cannot recognize zero marking as a marker. All errors of the types (d), (e), and (f) concern languages where there are only bound gender markers consisting maximally of two phonemes; in most instances there is even only a single character. These are most difficult to identify.

Finally the failure in (g) is probably an artifact of the orthography not distinguishing tone, but I do not have any description of Teutila Cuicatec available to check whether te occurring 3573 times in the N.T., only a small part of which is the feminine gender marker, is a case of homonymy or undifferentiated orthography. But Cuicatec languages also have a general respect gender that makes extraction more difficult.

Using a larger search domain would be helpful for a few languages. With a search space of 293 verses mainly based on English *she/her* markers are extracted even for Ashéninka Pajonal >#ok, >#op<, Asháninka >#o<, Caquinte >#o<, and more markers in other languages, such as Avar, >#й<, Tachelhit >#t<, Tamasheq >#tě<, >#tă<, Maltese >et<, Machiguenga >#os< are extracted. (Note also that Kabyle >#te< is only extracted with the lower threshold.) However, using a larger search domain comes at the cost of more wrong forms not filtered and nine languages with non-mature feminine gender markers and Yagua not extracted. I have not managed to extract any forms in Hindi, Gujarati, Eastern Panjabi, Chechen, and Iraqw, however the extraction is designed.

Explaining away exceptions is always problematic. However, the discussion shows that there are good reasons why the algorithm misses gender in a few languages.

[6] The Cyrillic alphabet not representing /j/ with a single letter is an additional difficulty, but the extraction does not succeed even when the text is transliterated.

3.6 Cases where the automatic extraction fails to extract particular forms

It is quite astonishing that in most languages anaphoric gender markers can be identified without previous analysis of any other grammatical categories or lexemes. This means that in most languages at least some anaphoric gender markers tend to have very high cue validity and are constructional islands (item-based constructions with a constant element; Tomasello 2003; see §2.5) which can be considered in abstraction from most other aspects of grammar as a form-meaning relationship in the text. The only exception the extraction algorithm has to make is that it must consider the feminine anaphoric singular domain in opposition to the anaphoric masculine singular domain and to the lexical domain 'woman', however, without having acquired the grammar of how feminine and masculine gender interact with other categories. This entails that the algorithm fails to recognize cases of "diagonal" syncretism involving cumulation (Table 4). Diagonal syncretism is similar to neutralization in that a form is used for more than one category. However, the opposition is not neutralized since there is another cumulating category that keeps the values distinct. An example is the Latvian third person feminine nominative singular pronoun *viņa* 'she', which has the same form as the masculine genitive singular form. The algorithm used here excludes it, because this form is also used in the masculine singular anaphoric gender domain. The algorithm fails to recognize that there is cumulation with an entirely different category: case. Another case in point is Afrikaans *sy* 'she' which is also used for possessive masculine 'his'; only *haar* 'feminine oblique' is extracted. "Diagonal" syncretism only occurs in mature gender markers.

Table 4: Cases of "diagonal" syncretism

Latvian			Afrikaans		
	F	M		F	M
NOM.SG	***viņa***	*viņš*	3SG	***sy***	*hy*
GEN.SG	*viņas*	***viņa***	POSS.3SG	*haar*	***sy***

Interestingly, there is no language in the sample where a feminine gender gram is missed due to "diagonal" syncretism. All languages of the sample with "diagonal" syncretism also have another feminine anaphoric gender marker with higher cue validity.

Some forms are not extracted due to other cases of homonymy where the other homonymous form is much more frequent. French *la* '3SG.F.ACC' is not extracted, because this form is primarily used as a definite article outside the anaphoric gender domain.

Affixes, especially short affixes, are more difficult to extract than free forms. This holds especially of affixes restricted to object, absolutive, and/or recipient marking. In some cases the form for 'said to her' is extracted instead of the feminine recipient affix. This holds for some languages of New Guinea and for some languages of South America: Ama *i-**so**-ki* [say-**O3SG.F**-REM.PST], Mian *baa-b-**o**-n-e-a* [say.PFV-BEN:PFV-**IO.3SG.F.PFV**-SS.SEQ-S.3SG.M-MED] (Fedden 2007), Bine ***jo**-ji-ge* [**ABS.3SG.F**-say-ERG.3SG] (as opposed to *je-ji-ge* [ABS.3SG.M-say-ERG.3SG] 'said to him'). In Kamasau the only form extracted is ***w**-uso* [**3SG.F**-go] 'she went' (Sanders & Sanders 1994: 21). This is partly an artifact of the size of the search space. With larger search spaces, short bound morphs are more easily detected.

Due to the statistical nature of the algorithm, rare forms cannot be extracted since it cannot be known whether rare forms only accidentally occur in the search domain. This means in practice that forms occurring in less than eleven verses (or 15% of the search domain) cannot be extracted. This affects, for instance, contrastive subject forms, such as Welsh *hithau*, possessive forms with gender agreement, such as German *ihr-e/en/es/er* [3SG.F-AGR], demonstratives used for referents of relatively low activation (Kibrik 2011: 327), such as Latvian *t-ā* [DEM.DIST-NOM.SG.F] and Latin *hæc*, and the Latin relative pronoun *quæ* [REL.NOM.F.SG] in non-relative use marking text coherence. Since there can be many feminine anaphoric gender markers, especially when markers are mature, there is a considerable amount of forms missed in languages with mature gender.

Gender markers for special groups of female beings, such as young women or female deities, as they frequently occur in Mesoamerican languages, are not extracted by the algorithm. Forms for young women are mostly filtered by the lexical 'girl' filter. Other groups, such as female deities, are not represented with sufficiently high frequency in the text.

3.7 Conclusions

As can be seen in more detail in Appendix A and B, there are 629 languages in the sample lacking a feminine anaphoric gender gram and 187 languages where such a gram is attested. Furthermore, it can be seen in Appendix A that the automatic extraction fails to detect feminine gender in 18 languages (3 Indo-Aryan, 1 Nakh-Daghestanian, 1 Cushitic, 5 Tucanoan, 1 Mayan, and 7 Otomanguean). Wherever extraction fails, there is a good reason for it (anaphoric function for animate

nouns highly restricted, very short bound or different bound affixes on verbs, zero exponence, or low degree of grammaticalization of the gram).

With one exception the wrongly extracted forms are all closely related semantically to feminine anaphoric gender and include feminine person name markers (5 lgs.), forms of a noun for 'woman' with a demonstratrive or definite affix (9 lgs.), other forms of 'woman' (1 lg.), demonstratives and definite articles (5 lgs., 2 of them distinguishing gender within the NP), and general third person pronouns (2 lgs). With the higher threshold, a feminine anaphoric gender gram is missed in 21 languages and a marker is wrongly extracted in 15 languages (all with some semantic resemblance to feminine anaphoric gender).

We can therefore conclude that almost all errors are systematic errors. Some are due to the crude nature of the algorithm that cannot segment word forms into morphemes. Some are due to the fact that some other grammatical phenomena are very closely related to anaphoric gender. Some failures are due to the fact that anaphoric gender has low cue validity in some languages. Rare forms are not detected. Throughout this section we have also seen that errors are sometimes even more valuable than correct results as they reveal where gender is particularly complex in certain ways. The procedure is highly useful as a heuristic device to check whether there are feminine anaphoric singular gender markers in a language.

4 Cumulation with grammatical relations and maturity of anaphoric gender

Once feminine anaphoric gender grams have been extracted for the languages of the sample, we can arrange the forms as they are distributed over various grammatical relations. This has been done by means of manual analysis and Table 5 illustrates the results for a few languages of the sample where there is some suppletion and/or neutralization for some grammatical relations. The languages listed in Table 5 represent different patterns of suppletion and/or neutralization and are discussed in more detail later in this section. The grammatical relations listed are A (transitive subject), S (intransitive subject), P (monotransitive object), R (recipient, indirect object), Poss1 (non-reflexive possessor or alienable possessor) and Poss2 (reflexive possessor or inalienable possessor; i.e., any less independent kind of possessor). Bound forms are indicated as affixes to the verb (-)V(-) or noun (-)N(-). See Appendix A for the whole sample. The examples in Table 5 are discussed in more detail below.

Table 5: Feminine gender grams (third person singular) in selected languages

	A	S	P	R	Poss1	Poss2
English	*she*	*she*	*her*	*her*	*her*	*her*
Belize Kriol	*shee*	*shee*	–	–	–	–
German	*sie*	*sie*	*ihr*	*ihr*	*ihr*-AGR	*ihr*-AGR
Welsh	*hi*	*hi*	*hi*	*wrthi*	*ei*+ASPIR	*ei*+ASPIR
Latin	*illa, quæ, hæc*	*illa, quæ, hæc*	*eam, illam*	–	–	–
Latvian	*viņa,* (V-*usi*)	*viņa, viņa,* (V-*usi*)	–	*viņai*	*viņas*	–
Northern Kurdish	*wê*	–	–	*wê*	*wê*	–
Hindi	–	V-*ī*	– (V-*ī*)	–	–	–
Ama	–	V-*mo*-	V-*mo*-	V-*mo*-	–	–
Au	*hɨre* / *w*-V	*hɨre* / *w*-V	V-*p*	V-*we*	AGR-*ɨre*	AGR-*ɨre*

As argued in §1, feminine anaphoric gender grams as those listed in Table 5 are mature. The markers have the function of noun phrases, but suppletion and neutralization is not characteristic of nouns. While mature anaphoric gender markers are often shorter phonologically than non-mature markers, a more reliable token of maturity is higher complexity in the sense of formal variability. The incipient anaphoric gender markers discussed in §5 are typically invariant across grammatical relations and not systematically absent from any grammatical relations (except sometimes reflexive possessor). This makes them differ from most pronominal anaphoric gender markers which exhibit cumulation and/or neutralization. English *she* (subject) and *her* (object, indirect object, and possessor) illustrate this point. Nouns are not entirely precluded from suppletion according to grammatical relation, but such suppletion in nouns is rare. Vafaeian (2013) shows that suppletion in nouns is common according to number, possession, and vocative case. In her sample of 63 languages there is only one language, Archi (Nakh-Daghestanian) with suppletion according to grammatical relation (absolutive/ergative in two nouns). Pronouns, however, and especially if bound

pronouns are included, usually display some sort of suppletion and/or neutralization according to grammatical relation. In Turkish third person, for instance, the free pronoun has the stem *o* and the possessive suffix is *-i/ı/u/ü*. Pronouns can lack suppletion or neutralization according to grammatical relation, such as Mandarin Chinese *ta*[1] 's/he', but in pronouns this is the less frequent option cross-linguistically.

Anaphoric gender grams exhibiting suppletion or neutralization must have undergone some kind of grammaticalization process. They presuppose earlier stages with simpler gender grams which are more similar to nouns or have developed from markers of other grammatical categories (such as case or number). How anaphoric gender grams can develop from nouns and noun phrases will be discussed in §5 based on the languages of the sample lacking suppletion and neutralization according to grammatical roles. Suppletion and/or neutralization are not necessary properties of gender grams with a long prehistory, but since most grams extracted here with long prehistories of gender exhibit these properties, I will refer to grams lacking suppletion and neutralization as "non-mature". Figure 1 shows the distribution of mature and non-mature feminine anaphoric gender grams in the languages of the sample.

Let us now discuss the languages listed in Table 5 one-by-one:

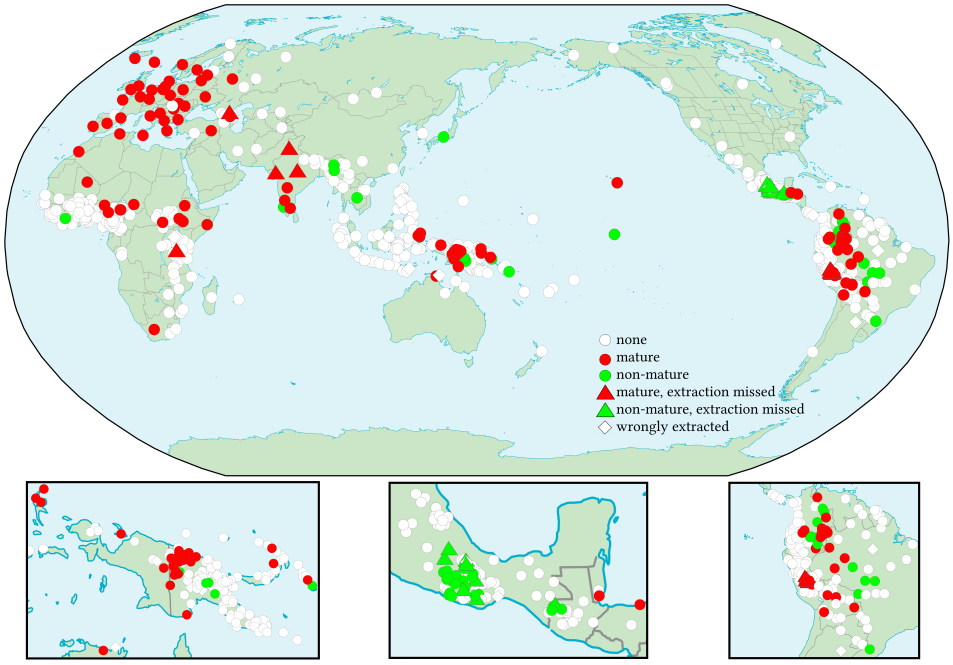
While English has a feminine marker for all relations – *she* for subject and *her* for all other ones – Belize Kriol English (at least the N.T. version) distinguishes feminine *shee* only for S and A (subject); object *ahn* and the possessor *ih* do not distinguish gender. Even though there is only one form, there is different behavior across grammatical relations since the single feminine form does not occur as non-subject, where gender is neutralized in Belize Kriol English.

Agreement of possessors with head nouns is indicated by AGR in Table 5 and illustrated in (13) for German and (14) for Au. These examples show that gender indexation (boldface) and NP-internal gender agreement (arrow) can be expressed on the same word form.

(13) German (Indo-European; Mk. 3:31 ; Matth. 14:8)

a. ***sein**-e* *Mutter*
POSS.3SG.**M**-NOM.SG.F← Mutter(F)[NOM]
'**his** mother'

b. *von* ***ihr**-er* *Mutter*
from POSS.3SG.**F**-DAT.SG.F← mother(F)[DAT]
'by **her** mother'



Map designed with the WALS Interactive Reference Tool by Hans-Jörg Bibiko.

Figure 1: Languages of the sample with mature and non-mature feminine gender grams

(14) Au (Torricelli; Mk. 3:31, Mk. 9:21, Matth. 14:11)

a. *miye* *p-**irak***
mother(F) →SG.F-POSS.3SG.**M**
'**his** mother'

b. *haai* *k-**irak***
father(M) →SG.M-POSS.3SG.**M**
'**his** father'

c. *miye* *p-**ire***
mother(F) →SG.F-POSS.3SG.**F**
'**her** mother'

Welsh (15) represents a special case in that anaphoric gender in possessive pronouns is marked only as a sandhi phenomenon spread to the following head noun. The third person singular masculine form *ei* causes soft mutation (among

other things *m->f-*); the third person singular feminine form *ei*, however, causes aspirate mutation (no change for *m-*). This looks as if there was agreement into the wrong direction, but is simply a rather intricate case of anaphoric gender marking.

(15) Welsh (Indo-European; Matth. 14:8, Matth. 12:46)

a. ***ei*** *fam*
POSS.3SG POSS.3SG.**M**:mother(F)
'**his** mother'

b. ***ei*** ***m****am*
POSS.3SG POSS.3SG.**F**:mother(F)
'**her** mother'

Latin lacks gender distinctions in the dative and in the possessor (both non-reflexive *eius* and reflexive *su-*AGR). Latvian lacks a gender distinction for the direct object (*viņu* ACC.SG.M/F) and for reflexive possessors (*sav-*AGR RPOSS.M/F). In the subject, gender in Latvian is indexed not only by the free pronoun, but sometimes also in participles (*-usi* PTCP.PST.ACT.NOM.SG.F). Northern Kurdish distinguishes gender in the oblique (*wî* M, *wê* F), which covers A (ergative), R and non-reflexive possessor, but not in the absolutive (*ew* M/F) S and P relations. Hindi lacks gender in free pronouns, and in the perfective past, which I take here as the most representative form since it is used in narrative function, gender is marked on the verb (*-ī* F) only in intransitive verbs and in some transitive verbs for the object. Ama (see also (3)) marks gender on the verb, but only for the absolutive, which, however, also covers the primary object (P and R): *ko-so-ki* [see-O.3SG.F-REM.PST] 's/he saw her' vs. *ki-Ø-ki* [see-O.3SG.M-REM.PST] 's/he saw him', *i-so-ki* [say-O.3SG.F-REM.PST] 's/he said to her' vs. *i-mo-ki* [say-O.3SG.M[7]-REM.PST] (Årsjö 1999).[8]

In all languages listed in Table 5, anaphoric gender is well entrenched, which can be seen from the fact that its marker interacts in some way with grammatical relations, either by means of cumulation or neutralization. This situation is characteristic of mature gender grams where anaphoric gender has a long history. This is opposed to incipient gender marking where the gender gram is less

[7]The masculine form is zero except for a few relics with *-mo-* as in the verb 'say'.

[8]Some predicates are especially salient in terms of frequency in the corpus with animate participants, these are notably 'go/come/arrive' for S, 'see' for P, and 'say' for R. However, the indexes listed in Table 5 are not always equivalent in translation; for instance, not in all languages 'see' is transitive.

complex and usually only has a single form irrespective of grammatical relation and where the use of the gram tends to be optional.

All examples in Table 5 have in common that anaphoric gender marking is pronominal (whether free or bound) and has variable formal expression across grammatical relations as opposed to the invariant anaphoric gender markers of nominal origin or supposedly nominal origin discussed in §5. However, not all invariant anaphoric gender markers can be proven to have nominal origin. There are, for instance, two Tupian languages in the sample with invariant markers. Kayabi *ẽẽ* F (male speaker) M, *kyna* F (female speaker), M *'ga* (male speaker), and *kĩã* M (female speaker), distinguishing both speaker and referent gender. These markers also follow person names and animate nouns in anaphoric use. Tenharim has *hẽa* F and *'ga* M (singular and plural), which also occur as suffixes on referring person names and animate nouns. Like in other Tupian languages the pronominal prefixes on nouns and verbs do not distinguish gender in Kayabi (Dobson 2005: 27) and Tenharim (Betts 1981: 17). The lack of gender markers in most Tupian languages might suggest that anaphoric gender in Kayabi and Tenharim are innovations.

However, invariant marking does not always testify to recent origin of gender. Malayalam (Dravidian) has the constant pronominal stem *avaḷ(-)* 3SG.F and no bound pronouns. But Old Malayalam still had subject indexes on the verb (*-āḷ* 3SG.F) (Andronov 1996: 120). Anaphoric gender marking was thus not invariant in Old Malayalam. While all Indo-European languages and all Creole languages with anaphoric gender in the sample have variant anaphoric gender marking, the artificial language Esperanto has invariant marking with the constant markers *sxi(-)* F and *li(-)* M.

Anaphoric gender can occasionally have quite unexpected sources. In Yagua, women who have borne children are referred to by dual forms (Payne 1985: 42) – 3DU *naada-* (often realized as *naan-*), *naadá*, 2DU *sáána-*, *saadá*. Men, however, are referred to with singular bound pronouns: 3SG *sa-* [I], *-níí* [II], 2SG *jiy-* [I], *jíy* [II].[9] In the N.T. dual forms are used as a default for adult women for whom it is not specified in the text whether they have given birth to children. Even if this is lack of gender from the point of view of the system – and Payne (1985: 42) says explicitly that Yagua lacks gender – this is an anaphoric gender marking opposition from the point of view of language use. Anaphoric gender in Yagua hijacks another highly grammaticalized category, number. This is why the markers are mature even if they are presumably young as gender markers. Yagua is thus an example of a very specific origin of an anaphoric gender opposition which has a

[9]Set II forms are used for direct objects and some intransitive subjects.

mature marker from the very beginning. However, since the origin of gender is often associated with case or number (Wälchli & Di Garbo 2019 [this volume]), the example of Yagua is perhaps less parochial than it seems at first glance.

To summarize: Even though there are a few exceptions, cumulation and/or neutralization testify to mature anaphoric gender marking whereas lack of cumulation and/or neutralization typically goes hand in hand with incipient gender marking. Since cumulation and neutralization can be considered to reflect an increase in complexity, this is evidence that complexity in anaphoric gender increases over time.

5 Grammatical anaphors and incipient anaphoric gender markers

5.1 Introduction

Third person pronouns (*he/she*) and full NPs have very similar properties in anaphoric function. Notably, there is very little semantic difference between a gender marked anaphoric pronouns (*he/she*) and a full definite NPs containing a light noun (a noun with a very general meaning, such as 'man', 'woman', 'thing'). This contrasts with their very different form – pronoun vs. noun – which assigns them entirely different roles in the typology of referential devices. As mentioned above, Kibrik (2011) makes a distinction between full referential devices (common nouns with or without modifiers, and person names) and reduced referential devices (pronouns and zero forms) and claims that it is universal: "The only truly universal opposition is that between full and reduced referential devices" (Kibrik 2011: 42). Grammatical anaphors are intermediate referential devices in the sense that they are neither lexical nouns nor third person pronouns. However, the distinction is still clear-cut in the sense that grammatical anaphors are grammatical in the same way as personal pronouns and hence to be included when discussing gender grams. Kibrik (2011: 123–136) discusses several of the grammatical anaphors considered here, such as Jacaltec classifiers and Japanese *kare* 'he' and *kanojo* 'she', under the heading "functional analogues" of personal pronouns.

Describing grammatical anaphors is essentially a synchronic aim. However, since grammaticalization tends to be unidirectional (Haspelmath 1999) and intermediate forms do not seem to evolve from more grammaticalized pronominal anaphoric gender markers, there is automatically also a diachronic dimension. Put differently, forms intermediate between nouns and indexes also tend to be INCIPIENT GENDER MARKERS. Intermediate forms (grammatical anaphors) keep

from their lexical origin the property of distinguishing the basically lexical meanings 'woman' and 'man', but they are decategorialized from the lexical category of nouns. However, since the diachrony of grammatical anaphors often remains opaque, this is in some cases only a hypothesis. It is important to point out that incipient gender markers do not necessarily further grammaticalize to mature gender markers. It is very well possible that incipient gender markers can be lost or remain incipient. As discussed in §4, mature gender markers can develop from other grammatical categories, such as number, case or person, and need not necessarily develop from incipient anaphoric gender markers.

Grammatical anaphors have both pronominal and nominal properties. Three different subtypes are discussed in this section as illustrated in Table 6.

Table 6: Three subtypes of grammatical anaphors

Subtype	Example	Subsection
Non-compositional complex NP	Japanese *kano(-)jo* *'that(-)woman'	§5.2
Reduced nominal anaphor	Chalcatongo Mixtec *-ña* (*ñã'ã* 'woman')	§5.3
General noun	Northern Khmer *niang* 'girl; she'	§5.4

Non-compositional complex NPs differ from the other types in that they are diachronically complex (more than one morpheme). Reduced nominal anaphors differ from the other two simplex types in that they diachronically reflect reduced nouns. General nouns have the form of a non-reduced noun, but they are so extended in use that they are semantically difficult to distinguish from pronouns. What makes them pronoun-like is not their form or word class, but the fact that their use is broader than in their lexical nominal use. Put differently, general nouns have specific meaning when used as nouns and more general meaning when used as grammatical anaphors.

Two further issues need to be specified. The first one is that not all instances of incipient anaphoric gender markers reflect genuine grammaticalization developments since linguistic gender categories can be subject to deliberate language planning. As there are sometimes attempts to eliminate anaphoric gender by language planning (for instance, in Swedish, a gender neutral form *hen* has been suggested to replace *han* 'he' and *hon* 'she' and is now partly gaining ground especially in generic use; see Milles 2011: 27), there have been attempts to im-

plement gender distinctions in pronouns where there a none. A case in point is Uduk where the N.T. uses the noun *(a)yim* [CLASS2] 'female friend' for 'she' even though this noun does not have any anaphoric use in spoken Uduk (Don Killian, p.c.). Thus, Bible translation Uduk has a special pronominal noun whereas there are no indications of a grammaticalization of an anaphoric gender gram in spoken Uduk (for more information on gender in Uduk, see Killian 2019 [in Volume I]).

The second one is that the presence of a masculine grammatical anaphor does not entail the presence of a feminine form.[10] As other Mek languages, Yale (13) has a masculine, but no feminine grammatical anaphor. Yale does not distinguish gender in third person pronouns (*el* 3SG), but has a special form *bone* glossed 'this.man' by Heeschen (1992), which does not contain the noun *nimi* 'man', but rather looks like a demonstrative pronoun as it cumulates the expression of spatial deixis with its nominal meaning (*ane* 'this here', *ani* 'that up there', *anu* 'that down there', *bini* 'that man up there', *bunu* 'that man down there'; Heeschen 1992: 15). All three devices, demonstrative NP, grammatical anaphor and personal pronoun, occur in example (16) and are summarized in Table 7.

(16) Yale (Mek, Trans-New Guinea phylum; Heeschen 1992: 29)
Nimi ane *dinge,* ***bone*** *dinge* *dane,* ***el****-di* *kwaneng*
man this property, this.man property DEM:PL 3SG-GEN sweet.potato
wa-m-la=ba, *na* *do-do* *de-n.*
be-PRF-PRS.3SG=CONNECT 1SG take-CVB eat[PFV].PRS.1SG
'I have taken and eaten (earlier today) this man's sweet potatoes.'

While the etymology of *bone* 'this.man' is opaque, there is a second grammatical anaphor in Yale which obviously derives from a full NP: *mene* 'this.child' (*mini* 'that child up there', *munu* 'that child down there' < *me ane/ani/anu*).

This section does not discuss all languages in the sample where gender has emerged recently. Due to genealogical considerations, in some languages feminine gender must have emerged recently (all related languages lack feminine; this holds, e.g., for Northern Wè within Niger-Congo; Paradis 1983), but it is not possible to trace a non-pronominal origin of gender markers.

It should be also stressed that automatic extraction of anaphoric gender (§3) has been the dominant heuristic in identifying the relevant set of languages. Many languages discussed here are not traditionally considered gender languages

[10] I do not know of any case of the contrary, a feminine grammatical anaphor without a corresponding masculine form.

Table 7: Yale third person pronouns, grammatical anaphors and demonstrative NPs

3SG	Grammatical anaphors	N DEM
	bone 'this.man'	*nimi ane/ene* 'this man'
el 'she/he'	—	*kel ane/ene* 'this woman'
	mene 'this.child'	*me ane/ene* 'this child'

and when I obtained forms in the automatic extraction I first thought that there must be some mistake in the algorithm.[11]

Some of the forms to be discussed in this section figure prominently in the literature on classifiers, especially NOUN CLASSIFIERS. This is no surprise since anaphoric use is a well-recognized function of noun classifiers in some languages. According to Aikhenvald (2000: 87) "noun classifiers are typically used with anaphoric function". Aikhenvald discusses especially Mayan languages of the Kanjobalan branch (Jacaltec and Akateko) and some Australian languages (notably Yidiny). It is thus not unexpected that some noun classifier languages are found to exhibit anaphoric gender which does not presuppose agreement as definitional property.

The literature on noun classifiers has in common with the literature on gender that it considers anaphoric use to be secondary. Noun classifiers as grammatical markers co-occurring with nouns in the same NP are not the topic of this paper, and in the same way as anaphoric gender can be considered without making reference to the notion of agreement, it can also be considered without making reference to the notion of noun classifiers.

5.2 Non-compositional complex NPs

Non-compositional complex NPs have similar uses as expressions for 'that man/woman', and sometimes they are entirely opaque, as the example from Kiribati illustrated in §1. However, non-compositional complex NPs are not usually condensed forms of 'that woman/man'; rather they contain other nouns that have been generalized to general meanings of feminine or masculine, such as 'mother' or 'elder sister' or 'body' or they contain obsolete or irregular forms of demonstrative pronouns.

[11] Since many languages also have third person singular forms not distinguishing gender they are not usually captured in Siewierska's (2005) typology (except Japanese where the third person singular pronoun is zero anaphor).

English has no anaphoric non-compositional NPs, but a related phenomenon is indefinite pronouns originating from NPs, such as *somebody*. *Somebody* contains the noun *body*, but does not have the meaning that the noun *body* has. For a typology of indefinite pronouns, see Haspelmath (1997). In the languages of the sample, non-compositional complex NPs are attested in Kiribati (Austronesian), Japanese (isolate), Kannada (Dravidian), Zome (Sino-Tibetan), Golin and Chuave (Chimbu, Trans-New Guinea phylum). Anaphoric gender markers in some South American languages with noun classifiers, notably Nambikuara and in Guahiban and Witotoan languages, are highly reminiscent of non-compositional complex NPs and can perhaps be interpreted as more advanced stages of grammaticalization. Table 8 summarizes the forms of the languages discussed in this section.

Table 8: Languages with non-compositional complex NPs for female reference

	Index (3sg general)	Grammatical anaphor	NP 'that woman'	'woman'
Japanese	zero anaphor	*kanojo*	*sono onna*	*onna*
Kannada	*avaḷu* (F), V-*aḷu* (F)	*āke* (honorif.)	*ā strīyu*	*strīyu*
Zome	*amah*	*tuanu*	*tua numei*	*numei*
Kiribati	*ngaia, e*	*neierei*	*te aiine aarei*	*aiine*
Golin	V-*m*, V-*ngw*	*abalini*	*abal i*	*abal*
Chuave	V-*m*, V-*ngu*	*oparomi*	*opai,*	*opai*
S. Nambikuara	*te²na²*, zero, V-*la¹*	*ta¹ka³lx(ai²n)a²*	*txu¹h(a³ka³lx)ai²na²*	*txu¹ha²*
Cuiba	–	*barapowa*	*barapo petsiriwa, yabʉyo*	*yabʉyo, petsiriwa*
Guayabero	–	*-ow, hapow*	*ampow pawis*	*pawis*
Huitoto Murui	*ie*	*naiñaiño*	*naie riñño*	*riñño*
Huitoto Minica	*ie*	*afengo*	*afe riŋgo*	*riŋgo*
Bora	*(i-)*	*diílle, -lle*	*áalle*	*walle*

Japanese *kanojo* 'she' means originally 'that woman', but it is not a reduced form of *sono onna* [that woman] 'that woman'. *Kano* is originally the attributive form of a distal demonstrative (free form *kare*) that has come out of use except in a few fixed archaic expressions such as *kare kore* 'that and this'. *Jo* is the Sino-Japanese expression for 'woman' (Ishiyama 2008: 141). *Kanojo* and its masculine counterpart *kare* 'he' (originally 'that') were established in the Meiji period (1868–1912) in the literary movement *genbun-itchi* (unification of written and spoken language) where translations from European languages played an important role (Ishiyama 2008: 139). There is some element of deliberate manipulation in this grammaticalization process and there is no reduction or erosion contributing to the grammaticalization of *kanojo* 'she'. The reason why *kanojo*

cannot be analyzed as a compositional NP anymore is that the demonstrative *kano* has disappeared. Although *kanojo* usually is translated with 'she' it could also still be translated as 'that woman'. In the N.T. *kanojo* competes in the anaphoric domain with *onna* 'woman' and *sono onna* 'that woman' (*suruto onna ha itsut-ta* [and woman TOP say-PST] 'So she said'; Matth. 15:27). *Kanojo* and *kare* cannot be compared to *she* and *he* in terms of text frequency (Ishiyama 2008: 36). Japanese prefers zero anaphor as reduced referential device (Kibrik 2011: 44). *Kanojo* also has some rather nominal uses: *kanojo wa?* [she TOP] 'Do you have a girlfriend?' (Ishiyama 2008: 232). It can also be used as a term for address (Ishiyama 2008: 232) which further shows that it is not a canonical third person pronoun.

Kannada (Dravidian) has so called honorific pronouns *āke* 'that woman, she', *īke* 'this woman', which have developed from the demonstratives *ā* 'that', *ī* 'this' and *akka* 'elder sister'. The second component in *ātanu* 'that man, he', *ītanu* 'this man' (honorific) is of Sanskrit origin: *dēha-* 'person, body'. Similar forms are found in Telugu (Andronov 2003: 171). Kannada and Telugu are the only languages I am aware of which have both gender-distinguishing third person pronouns (Kannada *avaḷu* 'she', *avanu* 'he') and grammatical anaphors.

Zome (Sino-Tibetan) *nu* and *pa* mean 'mother' and 'father' when possessed (*ka/na/a nu* [1SG/2SG/3SG mother]), but with the demonstratives *tua* 'that' and *hih* 'this' they are non-compositional complex NPs: *tuanu* 'that woman, she', *hih nu* 'this woman, she'. The corresponding nouns are *numei* 'woman' and *mi* 'man'. Rather than just pronouns and NPs there are three sets of forms in Zome: *ama(h)* 'he/she', *tuanu* 'she, that woman', and *tua numei* 'that woman'. It might be argued that *tuanu* 'that woman, she' is not sufficiently opaque to qualify as a non-compositional complex NP and is not much different from cases such as South Tairora *nraakyeva* [*nraakye-va* 'woman-DEM'] that have been removed as errors (see §3.4(a)). Indeed, no form is extracted for Zome if the form is spaced *tua nu*, where *nu* 'mother' is removed by the 'mother' filter. However, Zome is different from South Tairora in that the demonstrative is written without space only in few forms where it is semantically non-compositional, it is not generally an affix. Looking more closely for non-univerbated collocations of 'that mother' in the search domain in other Sino-Tibetan languages did not yield any further cases like Zome *hih nu* 'this woman, she', which suggests that Zome is different from other Sino-Tibetan languages in the sample.

In the variety of Golin (Trans New Guinea, Chimbu; documented by Bunn 1974: 55) which is the same as in the N.T., the pronouns for third person plural

abalíni 'she' < *abál inín* [woman REFL] and *yalíni* 'he' < *yál inín* [man REFL],[12] are not reflexive although they seem to contain reflexive markers. The variety documented by Evans et al. (2005) does not seem to have the same forms, but even this variety uses almost consistently NPs containing *abal* 'woman' or *gi* 'girl' and *yal* 'man' wherever the English translation has 'she' or 'he' as in (17) while in few cases where the reference is repeated within the same sentence there is only a bound affix for third person which does not distinguish gender.

(17) Golin (Lee 2005: 35)
abal i takal no-m
woman TOP what eat-3
'What did she eat?'

In the closely related language Chuave *opai* 'woman' and *yai* 'man' are opposed to *opa-rom-i* 'woman-?-DIST and *ya-rom-i/day* 'man-?-DIST/PROX' (Thurman 1987) where the element *-rom-*, misleadingly glossed 'this' by Thurman, only occurs in these two non-compositional anaphoric forms.

In Southern Nambikuara *txu¹ha²* 'woman' is opposed to *ta¹ka³lxai²na²* 'the woman, she' (*in³txa²* 'man' vs. *jah¹lai²na²* 'the man, he'). Lowe (1999: 283) lists *ta¹ka³lxai²na²* and *jah¹lai²na²* as third singular feminine free pronouns although they contain the demonstrative nominal ending *-ai²na²* and the base can take many other nominal endings including demonstrative emphatic *-ai²li²* and indefinite *-su²* (*ta¹ka³lxu²su²* once in the N.T. for 'a woman').[13] Kroeker (2001: 71) gives instead the forms with definite suffix (*-a²*) as third person forms (*ta¹ka³lxa²* and *jah¹la²*). There is also a third person form *te²na²* not distinguishing gender, which is used mostly in generic contexts where gender is not specified. Nambikuara has a large set of noun classifiers including *-a³ka³lx(i³)* feminine and *-(j)ah¹lo²* masculine which are always followed by nominal endings. These classifiers are placed at the end of NPs following adjectives and relative clauses. Thus, example (18) is one noun phrase. I interpret *Ta¹ka³lx(ai²n)a²* and *jah¹l(ai²n)a²* as non-compositional complex NPs.

(18) Southern Nambikuara (Rev. 17:18)
txu¹ha² ta¹ka³lx-a² ĩ²-in¹-ta³ka³lx-ai²na²
woman woman[ANA]-DEF see-2SG-F-DEM
'the woman whom thou sawest'

[12] The N.T. also has a few occurrences of *ibalini* (*ibal* 'people').
The documentation of Golin by Evans et al. (2005) has *yal (i) inin* 'he' [man (TOP) REFL] only twice and in both cases *inin* can be interpreted reflexively.

[13] Note, however, that even the free forms for first and second person have the demonstrative and emphatic noun suffixes *txai²na²/txai²li²* 'I', *wxãi²na²/wxãi²li²* 'you', but they do not take the definite and the indefinite endings.

In Guahiban and Witotoan languages feminine anaphoric and masculine forms consist of demonstratives with classifier suffixes which can perhaps be considered opaque grammaticalized forms of non-compositional complex NPs.

Guahiban languages use demonstratives with classifier suffixes as special anaphoric forms. Guayabero differs from Cuiba and Guahibo in that the forms have become bound indexes on verbs, which suggests a higher degree of grammaticalization. Cuiba (Guahiban) has the demonstratives *ba(ra)po-wa*, *po-wa* [this-F, that-F] and *ba(ra)po-n*, *po-n* [this-M, that-M]. Machal (2000: 237) lists the proximal <*bajapowa/bajaponü*> as personal pronouns, Merchán (2000: 589) the distal *powa/pon*; neither source mentions the forms *barapowa/barapon*. In the N.T. mainly the forms *ba(ra)powa/ba(ra)pon* are used anaphorically – both longer and shorter forms very much in similar contexts – often also preposed to person names in anaphoric use. *Powa/pon* are mostly used NP-internally as a relative clause introducer. The suffixes *-wa* F and *-n* M make part of a larger set of classifier suffixes. Merchán (2000: 589) lists eight other inanimate suffixes, which do not seem to occur with demonstrative stems, however. Attributive demonstratives usually lack classifier markers. For the closely related language Guahibo, de Kondo (1985, 1: 15) gives *pówa̲* F and *pónë̲* M as personal pronouns (which are, however, used only in relative function in the N.T. and rare) and the forms with proximal circumfix *ma̲-je* and distal prefix *baja-* as demonstratives (de Kondo 1985: 2: 49). In the N.T. *barapova* is the dominant feminine anaphoric form; *mapovaje* is mainly used for 'this woman', a combination of demonstrative and *petiriva* woman (*bajarapova petiriva*) is attested only once; for definite uses of 'woman' the demonstrative with the feminine classifier suffix is preferred in proximal or distal form. Guayabero, a third Guahiban language, is different in that F *-(p)ow* and M *-(p)on* are used as bound indexes on verbs if there is no NP subject (they are two of at least nine third person markers, including various diminutive and neuter forms, Keels 1985: 79, 86) and have become the major anaphor in the subject relation rather than the demonstratives *japow* and *japon*. According to Keels (1985: 79), subject and object indices can be combined on the same verb, but in the N.T. the object is usually expressed by the full pronoun *japow/japon*. The tendency to reduce subject markers more often than object markers can be seen as a first trait of maturity in Guayabero anaphoric indexes.

The special anaphoric form in Huitoto Minica (Witotoan) *afengo* 'she, that woman' (masculine *afemɨe*) consists of the demonstrative *afe* 'that' and the feminine noun classifier *-ngo* (masculine *-mie*) and is opposed to the noun *rɨngo* 'woman' (*ɨima* 'man') (Minor et al. 1982). The demonstrative can also combine with the noun: *afe rɨngo* 'that woman', *bie ringo* 'this woman'. The numeral for

'one' can combine both with the noun *daa ringo* 'a woman' and the classifier *daa-ngo* 'a woman' (rare). There is also a third person singular pronoun *ie* not distinguishing gender which is predominantly used in possessive function. Huitoto Murui is structurally very similar, except that the feminine classifier has various forms (*-ño*, *-ñaìño*) and is freer in combining with different pronominal stems (*naì-ñaìño* DEM.DIST-F, *baì-ñaìño* DEM.VIS-F, *bi-ñaìño* DEM.PROX-F, *i-ñaìño* 3SG-F). However *i-ñaìño* 3SG-F is rare and never used as a pronominal form (it is rather a free form of the classifier suffix). The most dominant anaphoric form is the distal *naìñaìño* 'she; that/the woman'. It is a matter of debate how closely related Bora and Huitoto are, but as far as the domain discussed here is concerned, the structural parallels are very strong. The major difference is that the Bora classifiers are not restricted to nouns and nominalizations but have extended to indexation on verbs, which is why Bora *-lle* 'F' and Muinane *-go* 'F' are much more frequent than Huitoto Minica *-ngo*. A special property of the Bora text is that the noun for 'woman', *walle*, is very rare in the N.T.; it is used almost exclusively in generic contexts. Almost the whole range of the nominal domain is covered by the classifier suffix *-lle*. With numerals, the classifier is used: *tsáápille* 'one/a woman'. The possessive prefix for third person *i-* does not distinguish gender.

Non-compositional complex NPs tend not to be genealogically pervasive. They pop up occasionally in most different language families, except in Guahiban and Witotoan where we also encounter the most mature exemplars. It can be assumed that non-compositional complex NPs originate from transparent complex NPs when one of their parts becomes opaque or as they acquire a non-compositional meaning. However, the nominal origin is a hypothesis as far as Kiribati and the South American languages are concerned, where the etymology of the forms cannot be traced.

5.3 Reduced nominal anaphors

While the non-compositional complex NPs discussed in §5.2 are found in a wide range of language families, the reduced nominal anaphors in the sample all come from Mesoamerica and almost exclusively from one family, Otomanguean. Table 9 lists examples from six Otomanguean examples, where reduced nominal anaphors occur in subject and reflexive possessor roles.

Reduced nominal anaphors in Otomanguean are both more grammaticalized and less grammaticalized than non-compositional complex NPs discussed in §5.2. They are rather highly grammaticalized in that they quickly increase in token frequency as they extend to all grammatical relations including reflexive possessors. However, they tend to remain more restricted in use semantically. There can be

Table 9: V-subject and N-reflexive possessor in 'and she (=the girl) brought it to her mother' (Matth. 14:11) in selected Otomanguean languages with anaphoric gender

Tlalcoyalco Popoloca	*co jehe xan joanjo* ***xan*** *ngain janné* ***xan***
	and 3 child gave **child[ANA]** give mother **child[ANA]**
San Miguel Mixtec	*te máá-**i**, nį janchą́ka-**i** nuų náą́-**i***
	and self-YOUNG COMPL gave-YOUNG to mother-YOUNG
Tepeuxila Cuicatec	*ní* ***tá^{n}'ā*** *mii^{n} ní ca'a* ***tá*** *chɛɛcu* ***tá***
	and **woman.F** there/DEF ? COMPL:give:3 **F** mother **F**
San Martín Itunyoso Triqui	*nį naga'ui'* ***ún'*** *ra'a nni* ***ún'***
	and gave **F** to mother **F**
Chiquihuitlan Mazatec	*ca-sua* ***na*** *naa rë* ***na***
	COMPL-give **F** mother POSS **F**
Amatlan Zapotec	*nu lee* ***me*** *m-zaaya lo xnaa* ***me***
	and FOC **F** COMPL-give to mother **F**

separate forms for young humans, as in San Miguel Mixtec, and often there are separate forms for human respect and for deities.

In some languages the nominal origin of the reduced forms can clearly be traced. This is most obvious in Tlalcoyalco Popoloca (Stark 2011). Although Tlalcoyalco Popoloca has a third person pronoun *je'e* not distinguishing gender there is a large number of short forms of nouns with anaphoric use (termed "short pronouns" in Stark 2011: 3). The most common include *xii* 'man[SG]' (anaphoric *xa*) and *nchrii* 'woman[SG]' (anaphoric *nchra*). Example (19) illustrates a plain noun *janna'a* 'mother' and its corresponding anaphoric form *jan*:

(19) Tlalcoyalco Popoloca (Stark 2011: 4)
Naa ***janna'a*** *jian anseen* ***jan*** *ixin rinao* ***jan*** *kain*
one mother fine heart mother[ANA] because loves mother[ANA] all
xe'en ***jan***.
children mother[ANA]
'A mother has a good heart because she loves all her children.'

Some condensed anaphoric NPs are reminiscent of noun classifiers ("pronouns that echo a prefix"; Stark 2011: 4) and some uses are compatible with a noun class with agreement interpretation as when animals take the pronoun *ba*. However,

anaphoric noun formation is productive and applies even to Spanish loanwords (*guitaarra*, anaphoric *guitarra*).

Tlalcoyalco Popoloca *nchra* 'woman[ANA]' is so specific in its meaning that it can hardly be considered a grammaticalized feminine gram. It has the distribution of a word for 'woman', and other female nouns have other anaphoric forms.

All Mixtec languages have clitic anaphoric gender markers usually following their head (following a verb for subject and object, following a relational noun for oblique and following a noun for possessor) which mostly have the phonological structure CV (see Macri 1983 for a survey of several Mixtec languages) and are much more strongly grammaticalized than Popoloca anaphoric nouns. Unlike first and second person, there are no full free forms for third person clitics, or rather the corresponding full free forms are nouns. Chalcatongo Mixtec (Macaulay 1996: 139) has the following six sets (in parentheses the nouns corresponding to the reduced nominal anaphors): masculine *-ðe* (*čàà* 'man'), feminine *-ña* (*ñã'ã* 'woman'), polite, older *-to* (*to'ò* 'older person'), supernatural *-ža* (*í'a*, *íža* 'god'), *-tɨ* animal, and *-ži* (no related noun, *žii* is 'male'). The clitics are usually not tenacious (i.e., they are dropped if there is an explicit NP), unless the NP preposed to the verb is a topic (Macaulay 1996: 140). A way to supplete the missing full forms needed for contrastive purposes is to add the clitic to the emphatic form *máá* 'self' (Macaulay 1996: 106, see also Table 4 above). The meaning of Mixtec genders is much more general than those of Tlalcoyalco Popoloca genders. But 'girl, young women' is often covered by the child gender in many Mixtec languages (see Table 9 for an example from San Miguel Mixtec). In Coatzospan Mixtec, feminine gender is of limited use since there is a general adult respect human gender *ña* that does not distinguish men and women. "[T]he use of a specifically masculine or feminine noun or pronoun to refer to an adult is usually considered disrespectful" (Small 1990: 406).

Reduced nominal anaphors or forms reminiscent of reduced nominal anaphors can also be found in Cuicatec (Bradley 1991), in Chiquihuitlan Mazatec (Capen 1996; but not in three other Mazatec languages included in the sample), and in Triqui (see Table 9).

Most Zapotec languages have some forms that are intermediate between nouns and third person pronouns. Feminine is not always a salient category though, because many Zapotec languages have a special respectful form used for both genders, especially for women by men speaking. In Texmelucan Zapotec respect is used for deity, respect human in women's speech and respect feminine in men's speech (Speck 1972: 290). Texmelucan Zapotec has masculine (*yu*, *-y*), feminine (*fiñ*, *ñi*, *-ñ*), respect (*mi*, *-m*), animal (*ma*, *bañ*) and neuter (*ñi*, *-ñ*), which occur

both in fuller and more reduced forms. As shown in (20), masculine and feminine can be modified by adjectives, numerals, and demonstratives, which makes them look rather like nouns, but they can even be reduced subject indexes on verbs.

(20) Texmelucan Zapotec (Speck 1972: 32)
Benu sac ***fiñ feñ*** *nu gusht ni* ***yu*** *feñ ze´ lugaar ze´ nu*
if not.be **3F** young COMP please PP **3M** young that place that COMP
cyiiñ ***yu***, ***yu ze´*** *neñ* ***yu*** *nu zu* ***tub ñi*** *ca zi´l na tub*
POT:live **3M** **3M** that hear **3M** COMP POT:stand one **3F** where only be one
ranch nu zet, ze´ a´ yu´ lo nap ***yu-ñ***, *orze´ uz* ***yu***
ranch COMP far but NEG PROG.be.in face good **3M-3F** then father **3M**
gzuu nez ***yu*** *i´ñ* ***yu*** ***yu*** *feñ ze´ nu cha-****y*** *cha*
POT:CAUS:stand trail **3M** child **3M** **3M** young that COMP POT:go-**3M** POT:go
*gwii-****y*** ***fiñ*** *mñaa ze´ ben a gyet lagy* ***yu-ñ***.
POT:see-**3M** **3F** woman that if Q POT:descend liver **3M-3F**
'If there are no young women who appeal to the young man at the place where he lives, but if he hears that there is one at some ranch or another that is far away, but if he doesn't know her well, his father will send his child, the young man, to go see if he likes her or not.'

For Mixtepec Zapotec, Hunn et al. (n.d.: 11) list fourteen categories of third person pronouns, twelve of which refer to persons and only one of which is a reduced form (C-*á*, V-*w* inanimate). Their use depends on the speaker as is quite common across Zapotec: e.g., *nîip*, *nîib* is used by men for a young man and by women for a man of their age or younger. Several categories refer to men and women of lesser respect. *Zhó* <zho> 'person of minor respect, group of people' is used, for instance, in the N.T. for the Samaritan (Lk. 10:33). Shifting use depending on speaker attitude is not easily understandable in terms of noun classes, but well in-line with the idea of anaphoric gender.

Gender is more strongly grammaticalized in Southern Rincón Zapotec, where the familiar forms lack a gender opposition and respectful forms distinguish masculine and feminine gender (Earl & de Earl 2006: 363). While the feminine consistently has the form *-nu* (free form *lë-nu*), the masculine form varies (free form *lë-'*): *blé'i-****në'****=nu* [COMP.saw-3SG.**M**.RESP=3SG.F.RESP] 'he saw her', *blé'i-nu=****në'*** [COMPL.see-3SG.F.RESP=3SG.**M**.RESP] 'she sees him', *cati' blé'i-****në' lë'*** [when COMPL. see-3SG.**M**.RESP 3SG.**M**.RESP] 'when he saw him', *rë-'-nu* [CONT.say-3SG.**M**.RESP-3SG.F.RESP] 'he said to her'. The allomorphs cannot be clearly ascribed to different

grammatical relations, however: *-(ë)'*, *-në'*, and *-lë'* all occur in direct object function. Aside from familiar (*-bi'*), feminine respect, and masculine respect, there are also forms for animal (*-ba*) and neuter in the third person singular.

The only non-Otomanguean language to be discussed in this section is Todos Santos Mam (Mayan). Mam has a set of twelve human classifiers which are reduced forms of nouns, non-compositional forms, or pronominal nouns (common noun *txiin* 'young woman' CL *txin*; *xu7j* 'woman' CL *xu7j* 'woman', CL *xuj* 'old woman (respectfully); *yaab'aj* 'grandmother' CL *xhyaa7* 'old woman'; England 1983: 158). While their use in Northern and Central Mam is mostly restricted to one occurrence per clause, Todos Santos Mam has extended them even to reflexive possessors as in (21).

(21) Todos Santos Mam (40014011)
*[…] bix e xi' t-k'o-'n-tl-**txin** t-e*
and ? go/DIR ERG.3-SG-give-DIR-again-CL.girl POSS.3SG-to
*t-txu-**txin**.*
POSS.3SG-mother-CL.girl
'and she (=the girl) brought it to her mother'

Note that both the ergative subject (A) and the reflexive possessor are indexed twice in (21), by the suffixed anaphoric gender marker and by the general third person singular prefix *t-*.

5.4 General nouns

General nouns have the form of a non-reduced noun, such as 'woman', 'girl' or 'wife', but because of their extension in use they are difficult to distinguish from pronouns. In the sample general nouns occur in four Mayan languages: Jacaltec, Akateko, Ixil Nebaj and Chuj, in Northern Khmer, and perhaps in the Austronesian language Owa.

It may seem strange at first glance that general nouns can be extracted by the algorithm since they have the same form as lexical nouns whose domains of use are applied as filters in the algorithm. The reason they can be extracted is that their use as general nouns is so pervasive that it is quite different from what the use of a lexical noun would be if everything is taken together.

The same Jacaltec form *ix* 'woman' is used all the way from the nominal low activation domain up to the top pronominal domain. *Naj* 'he (non-respected, non-kin)' is a reduced nominal anaphor (*winaj* 'man'). *Ix* 'woman; she (non-respected, non-kin)' and *naj* 'he' belong to the set of noun classifiers and are notably used

with thematically salient NPs in referential anaphoric function (Craig 1986: 267; Aikhenvald 2000: 323). There are no free third person pronominal forms except classifiers. Example (22) illustrates the non-respect feminine classifier *ix* 'woman' in non-reflexive possessor and subject function and the non-respect masculine classifier *naj* 'man' as a noun classifier in the anaphoric NP with a person name:

(22) Jacaltec (Matth. 14:8)
Y-al-ni *is-mi'* *ix* *t-et* *tato*
ERG.3-say-DETRANS POSS.3-mother CL.woman/F 3-to COMPL
ch-is-k'an *ix* *is-wi'* *naj* *Juan.*
INCOMPL-ERG.3-ask CL.woman/F POSS.3-head CL.man/M John
'Her mother said that she should ask for John's head.'

However, it is not the entire top activation domain that is covered by the general nouns. Reflexive possessors lack general nouns. Grinevald Craig (1977: 159), who describes the phenomenon in detail, calls this "noun classifier deletion under identity of reference". Diachronically classifiers are not deleted from reflexive possessor function; rather they have never been expanded to that domain. Note that reflexive possessor even includes co-reference with object as shown in (23) ("no constraint on the controller NP", Grinevald Craig 1977: 152).

(23) Jacaltec (Lk. 7:15)
y-a-ni-co *Comam* *naj* *t-et* *is-mi'.*
ERG.3-give-DETRANS-DIR CL.male.deity CL.man/M 3-to POSS.3-mother
'and he gave him$_j$ to his$_j$ mother'

The wider extension of possessive prefixes even to obligatory use with prepositions (*t-et* 3SG-to) testifies to their higher degree of maturity. Not all noun classifiers in Jacaltec (Day 1973: 125) are general or reduced nouns.

For Akateko, which is closely related to Jacaltec, see Zavala (1992). In Nebaj Ixil, which is only distantly related to Jacaltec and Akateko within Mayan, the nominal and general uses of *ixoj(e)* 'woman' and *naj* 'man' differ in that the former have a preposed determiner *u.* Thus, from the point of view of the whole NP the general forms could also be considered to be reduced forms. Chuj *'ix* 'woman' also arguably sorts here, although it is not as easily extracted as the forms in the other three Mayan languages.

In Northern Khmer (Austroasiatic) the noun used prominently in the high activation domain is เนียง *niang* 'girl' rather than ซแร็ย *srej* 'woman'. เนียง *niang* 'girl' also occurs as a term of address and it has probably become a special pronominal form by extension from deictic second person use to anaphoric third person

use. Special pronominal nouns are a feature of Southeast Asia. Vietnamese has a general human special pronominal noun *người* for adult human beings, which is also used as a noun classifier, but Vietnamese lacks a general feminine anaphoric noun.

Owa (Austronesian) *kani* 'she, the woman, that woman; wife', which is just above the lower threshold for extraction, is difficult to classify. One possibility is to interpret it as a general noun with the specific meaning 'wife', but it is not clear to me whether the nominal meaning 'wife', restricted to use with following possessor, is the original one. Mellow's (2013: 273) dictionary analyzes *kani* as "pronoun", but the form is not listed in the grammar's pronoun section, where just the general third person singular form *ngaia* is given (Mellow 2013: 7). As elaborated below, there is some evidence that *kani* might contain the female person name article *ka-*, but personal pronouns can also have articles, although most pronouns are in the i-class. Owa distinguishes five different genders in noun-phrase-initial articles listed here in their cumulative forms with coordination/comitative *mi*, where there are most clearly marked and distinguished: male person names *m-o*, female person names *mi-ka*, some nouns beginning in *e-* (mostly kinship terms, phonologically assigned) *m-e*, location nouns, some pronouns and the word *kare* 'child' *m(-)i*, and default *mi-na* (see also Mellow 2013: 26). The male and female person name classes are extended to some common nouns, especially kinship terms, but not 'father' and 'mother', which are e-class, and to the pronoun 'who' (mostly in the male form *mo o-tai* 'and who?'). The male counterpart of *kani* 'she, the woman, that woman; wife' usually co-occurs with the male person name article *o* as *o wani* 'he, the man, that man, husband', which suggests that *kani* is a condensation of **ka-wani* (compare also *o goana* 'brother' vs. *ka goana* 'sister', *na goana* 'friend(s), sibling(s)'), especially also because all traditional Owa names have the female person article fused as a prefix *ka-* (Mellow 2013: 20). In the N.T. *kani* is i-class in some instances (object *ki kani*; *mi kani* could also be interpreted as lack of article following *mi* 'and'), perhaps in phonological analogy to *kare* 'child' or in functional analogy to pronouns. In the automatic extraction *kani* is only extracted because there is no 'wife' filter. Whatever the origin of *kani*, it is a grammatical anaphor, but it remains unclear whether of the subtype general noun or the subtype non-compositional NP, which suggests that these two subtypes are not neatly different.

There are no examples with 'mother' as a general noun, but Zome, discussed in §5.2 comes close to it.

5.5 Conclusions

The three subtypes of grammatical anaphors discussed above reflect different parameters of grammaticalization that tend to behave differently in different non-mature anaphoric gender grams as summarized in Table 10. The definitional properties, marked with asterisks in Table 10, relate to different parameters. Hence, the types are not strictly opposed to each other, so that some forms, such as Zome *tuanu* (§5.2) and Owa *kani* (§5.4) can have properties characteristic of various subtypes. In reduced nominal anaphors (§5.3) the grammaticalization of form (reduction) is most advanced, which goes together with a high text frequency, whereas generalization can be almost absent as in Tlalcoyalco Popoloca. In general nouns (§5.4), generalization is the relevant factor of grammaticalization whereas formal reduction is absent. Non-compositional complex NPs (§5.2) can have low text frequency, as Japanese *kanojo*, unlike reduced nominal anaphors. The degree of decategorialization from nouns varies greatly. In most cases, grammatical anaphors retain at least some properties of nouns.

Table 10: Different properties of the subtypes of grammatical anaphors

Subtype	Complex	Opaque	Reduced	Frequent	General
Non-compositional complex NP	+*	+	–/+	–	+
Reduced nominal anaphor	–	+	+*	+	–
General noun	–	–	–	+/–	+*

The grammaticalization of grammatical anaphors is gradual for general nouns, while there is a more categorial border for reduced nominal anaphors and for non-compositional complex NPs (for the latter to the extent they are opaque). General nouns are not distinct in form from lexical nouns and generalization must have gone a long way before the markers escape filtering by the lexical noun their form instantiates.

6 Reconciling the gram approach with the system perspective

In the previous sections I have shown that it makes perfect sense to consider feminine anaphoric singular markers as a gram type (dedicated markers with a particular grammatical meaning, prototypically instantiated in a particular functional domain), and a typology of feminine singular anaphoric gender grams in

a sample of 816 languages has been presented, which abstracts away from viewing gender as a system phenomenon resting on the notions of noun class and agreement. However, it is undeniable that gender values form systems and that – even if not always canonical noun classes and canonical agreement – at least some kind of noun-class-like and agreement-like phenomena are crucial for the understanding of gender. The question thus arises as to what the gram approach can contribute to a better understanding of gender systems and of noun-class-like and of agreement-like phenomena in gender.

All gram types are alike in that they are markers instantiating a grammatical meaning X. However, beyond this common ground, different gram types may have different properties, and this is how they may become engaged in complex grammatical structures of particular kinds.

Feminine singular anaphoric gender grams are special in that they almost always are engaged in an opposition to another gram type, masculine singular anaphoric gender grams. This is no strict universal though. In §5.1 we have seen that Yale and some other Mek languages only have masculine anaphoric grams without parallel feminine anaphoric grams. However, Yale and other Mek languages are quite exotic in this respect. Oppositions are nothing strange for gram types. Most tense and aspect grams have some kind of oppositions. Perfect grams, for instance, are opposed to narrative (Dahl & Wälchli 2016: 327), but this does not make every perfect gram to be opposed to a narrative gram. Within the realm of aspect it is certainly perfective and imperfective that are most inclined to engage in a pair of oppositions and, not unexpectedly, perfective and imperfective grams are usually the core of aspect systems.

In the extraction of feminine anaphoric gender grams, I have made practical use of the opposition to anaphoric masculine by using the anaphoric masculine as a filter. I have not been able to design an implementable procedural definition of feminine gender grams that can dispense with filters. Filters are kinds of oppositions and oppositions are the building blocks of systems. Here it is important to point out that the filters that have been used are semantic domains rather than language specific structural elements. Put differently, semantics predestines the feminine anaphoric gender gram type for structural oppositions. However, feminine anaphoric gender grams are not only engaged in one kind of opposition, they are generally and necessarily engaged in two kinds of oppositions: (i) to masculine and (ii) to nominal lexical domains for the designation of female referents, the most important ones being ‘woman’, ‘girl’, ‘mother’, and ‘daughter’, and these are also indispensable as filters in the procedural definition.

What makes feminine anaphoric and masculine grams grammatical from a semantic point of view is their virtual restriction to anaphoric use. Nouns, even nouns that are typically used to designate individual items, such as *mother*, *sun*, and *god*, can be used non-anaphorically: *a mother*, *a sun*, *a god*. Unlike lexical nouns, anaphoric grams are not only dedicated to anaphoric use, they also tend to be more general than lexical nouns. They are almost always in a hyperonymic relation to lexical nouns (see also Seifart 2018). This can also hold when an anaphoric gram is not syntactically a pronoun as in Kiribati where *neierei* 'this woman' picks up reference to a range of female nouns. The least general feminine anaphoric grams we have encountered in Otomanguean languages (§5.3), most markedly in the extreme case of Tlalcoyalco Popoloca, where "short pronouns" are an open set.

As soon as anaphoric grams are "hyperonymic", they are noun-class-like, since they collocate with a set of hyponymic nouns. The Tlalcoyalco Popoloca "short pronoun" for animals is already reminiscent of a noun class, whereas the "short pronouns" for 'woman', 'mother', and 'girl' mainly correspond to particular lexical domains (this is why Tlalcoyalco Popoloca is filtered out in the automatic extraction). Here it is important to emphasize the difference between "noun class" and "noun-class-like". English, *she/her*, for instance, is noun-class-like. In practice, *she* and *her* tend to pick up reference to such nouns as *woman*, *wife*, *girl*, and *mother* etc., but that does not make feminine gender strictly lexical in English.

At the same time, the anaphoric character of "picking up reference" makes anaphoric grams agreement-like, which does not mean that anaphoric gender is agreement. It is important to emphasize the difference between "agreement" and "agreement-like". The agreement-like character of anaphoric grams derives from their semantic properties, it is not a syntactic process. However, due to the similarity of agreement and agreement-like anaphors, anaphoric gender grams are highly compatible with agreement phenomena and can be integrated in agreement systems, even though anaphors are essentially semantic, as they can pick up reference from the context without syntactic antecedents.

Furthermore, anaphoric grams are special in that they tend to form chains (multiple occurrences of the same gram, often in different grammatical relations and in free or bound encoding).

In the previous sections we have seen that feminine gender grams entertain close relationships to other grammatical and lexical categories. Considering the closer neighborhood of the feminine anaphoric gender gram type we may speculate about what might be possible next steps for expanding the gram approach to

gender and related phenomena. Aside of masculine singular and both feminine and masculine plural and dual forms, the most promising candidates for gram types are female and male person name markers and feminine and masculine NP-markers. These have been occasionally extracted as errors in the present investigation, so it might be possible to formulate procedural definitions that focus on these phenomena specifically and view them as gram types.

7 Conclusions

Grammatical gender is usually considered to be highly complex and it is traditionally defined in terms of agreement and noun classes, which are both complex phenomena. Thus, one way to explore whether gender might be simpler than commonly believed is to try to approach it without reference to the notions of agreement and noun classes. In this paper feminine anaphoric gender has been approached by way of a procedural definition which, when implemented in a computer program, extracts feminine gender markers from a parallel text corpus. This procedural definition does very well without any reference to agreement or noun classes suggesting that these notions are entirely dispensable at least for one important core domain of gender. It was also found that many anaphoric gender markers have high cue validity which suggests that they are not particularly complex. The notions the procedural definition relies on are those of functional domain and gram type which have proven to be useful for many other grammatical category types, suggesting that gender may be less puzzling among grammatical categories than commonly believed.

While there is a long research tradition of investigating particularly complex gender phenomena, less effort has been devoted to uncover simple gender. Thus, it has gone largely unnoticed in typology that there are many languages with non-pronominal anaphoric gender markers which are intermediate between full noun phrases and pronouns (grammatical anaphors). Non-pronominal anaphoric gender is less stable diachronically than pronominal anaphoric gender and can sometimes be proven to be very young. Gender in grammatical anaphors is therefore important for understanding how gender can develop diachronically. However, the low complexity of anaphoric gender also invites for deliberate manipulation as in the case of the Uduk New Testament where a feminine gender was created by missionaries.

Unlike non-pronominal anaphoric gender, pronominal gender is usually highly mature. This is reflected in the widespread suppletion and neutralization according to grammatical relations in pronominal gender, which are features of com-

plexity synchronically even in languages such as English and Belize Kriol English where gender is commonly believed to be simple.

Finally, this paper has shown that parallel texts are highly useful for the study of grammatical gender. They help shift the focus of attention to the most functional aspects of gender and away from more idiosyncratic properties. Parallel texts also show that gender is not an isolated phenomenon, but has often very similar functions as, for instance, light nouns. Hence, to uncover the functions of grammatical gender it may be useful to consider it together with other linguistic categories, including non-grammaticalized ones, which have similar functions. Grammatical anaphors which are often not recognized as gender markers in the descriptive literature can effectively be recognized as incipient gender markers in parallel texts.

Acknowledgments

I would like to thank Francesca Di Garbo, Östen Dahl, Andrej Kibrik, Robert Östling, Martin Haspelmath, Bruno Olsson and Annemarie Verkerk for many useful suggestions. While writing this paper I was partly funded by the Swedish Research Council (*Vetenskapsrådet*, 421–2011–1444).

Special abbreviations

The following abbreviations are not found in the Leipzig Glossing Rules:

A	transitive subject	MED	medial
ACT	active	O	(direct) object
AGR	agreement	N	noun
ANA	anaphoric	P	monotransitive object
CL	classifier	Poss1	inalienable or non-reflexive possessor
COMP	complementizer		
COMPL	completive aspect	Poss2	alienable or reflexive
CONT	continuative aspect	POT	potential aspect
CONNECT	connective	PP	preposition
DETRANS	detransitive	PRO	pronominal
DIR	directional	R	recipient/indirect object
EMPH	emphatic	RESP	respect
INCOMPL	incompletive aspect	RPOSS	reflexive possessive
IO	indirect object	S/S	intransitive subject

SEQ	sequential	V	verb/vowel
SPEC	specific noun	YOUNG	gender for children or young people
SS	same subject		

References

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

Anderson, Neil. 1989. Folopa existential verbs. In Karl Franklin (ed.), *Studies in componential analysis*, 83–105. Ukarumpa: Summer Institute of Linguistics.

Andronov, Michail S. 1996. *A grammar of the Malayalam language in historical treatment* (Beiträge zur Kenntnis Südasiatischer Sprachen und Literaturen 1). Wiesbaden: Harrassowitz.

Andronov, Michail S. 2003. *A comparative grammar of the Dravidian languages*. Munich: Lincom.

Årsjö, Britten. 1999. *Words in Ama*. Uppsala: Uppsala University. (MA thesis).

Audring, Jenny. 2009. Gender assignment and gender agreement: Evidence from pronominal gender languages. *Morphology* 18(2). 93–116.

Benton, Joseph. B. 1975. *Glossed text in Chichicapan Zapotec*. SIL Language & Culture Archives. oai:sil.org:59688.

Betts, La Vera. 1981. *Dicionário parintintín-português português-parintintín*. Brasília: Summer Institute of Linguistics.

Bosch, Peter. 1988. Representing and accessing focussed referents. *Language and Cognitive Processes* 3(3). 207–232.

Bradley, David P. 1991. A preliminary syntactic sketch of Concepción Pápalo Cuicatec. In C. Henry Bradley & Barbara E. Hollenbach (eds.), *Studies in the syntax of Mixtecan languages*, vol. 3, 409–506. Dallas: Summer Institute of Linguistics.

Bunn, Gordon. 1974. *Golin grammar* (Workpapers in Papua New Guinea 5). Ukarumpa: Summer Institute of Linguistics.

Bybee, Joan & Östen Dahl. 1989. The creation of tense and aspect systems in the languages of the world. *Studies in Language* 13(1). 51–103.

Capen, Carole Jamieson. 1996. *Diccionario mazateco de Chiquihuitlán Oaxaca*. Tucson: Instituto Lingüístico de Verano.

Contini-Morava, Ellen & Eve Danziger. 2018. Non-canonical gender in Mopan Maya. In Sebastian Fedden, Jenny Audring & Greville Corbett (eds.), *Non-canonical gender systems*, 129–146. Oxford: Oxford University Press.

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

Corbett, Greville G. 2006. *Agreement*. Cambridge: Cambridge University Press.
Corbett, Greville G. & Sebastian Fedden. 2016. Canonical gender. *Journal of Linguistics* 52(3). 495–531.
Craig, Colette. 1986. Jacaltec noun classifiers: A study in language and culture. In Colette Craig (ed.), *Noun classes and categorization*, 263–293. Amsterdam: John Benjamins.
Croft, William. 2003. *Typology and universals*. 2nd edn. Cambridge: Cambridge University Press.
Dahl, Östen. 2000. Animacy and the notion of semantic gender. In Barbara Unterbeck (ed.), *Gender in grammar and cognition*. Vol. 1: *Animacy and the notion of semantic gender: Approaches to gender*, 99–115. Berlin: Mouton de Gruyter.
Dahl, Östen. 2004. *The growth and maintenance of linguistic complexity*. Amsterdam: John Benjamins.
Dahl, Östen & Bernhard Wälchli. 2016. Perfects and iamitives: Two gram types in one grammatical space. *Letras de Hoje* 51. 325–348.
Day, Christopher. 1973. *The Jacaltec language* (Language Science Monographs 12). Bloomington: Indiana University Press.
de Kondo, Riena W. 1985. *El guahibo hablado: Gramática pedagógica del guahibo, lengua de la orinoquía colombiana*. Vol. 1–2. Lomalinda: Instituto Lingüístico de Verano.
Dobson, Rose M. 2005. *Aspectos da língua Kayabí*. 2nd edn. Brasília: Summer Institute of Linguistics.
Earl, Roberto & Catalina de Earl. 2006. *Diccionario zapoteco del Rincón*. SIL Language & Culture Archives. http://www.sil.org/resources/archives/58129.
England, Nora C. 1983. *A grammar of Mam, a Mayan language* (Texas Linguistics Series). Austin: University of Texas Press.
Evans, Nicholas, Jutta Besold, Hywel Stoakes & Alan Lee (eds.). 2005. *Materials on Golin: Grammar, texts and dictionary*. Melbourne: Department of Linguistics & Applied Linguistics, University of Melbourne.
Fedden, Sebastian. 2007. *A grammar of Mian: A Papuan language of New Guinea*. University of Melbourne. (Doctoral dissertation).
Fung, Pascale & Kenneth Ward Church. 1994. K-vec: A new approach for aligning parallel texts. In *Proceedings of the 15th conference on computational linguistics*, vol. 2, 1096–1102. Kyoto.
Garland, Roger & Susan Garland. 1975. A grammar sketch of Mountain Koiali. In Tom E. Dutton (ed.), *Studies in languages of Central and South-East Papua* (Pacific Linguistics C 29), 413–470. Canberra: Australian National University.

Givón, Talmy. 1981. Typology and functional domains. *Studies in Language* 5(2). 163–193.
Grinevald Craig, Colette. 1977. *The structure of Jacaltec*. Austin: University of Texas Press.
Haspelmath, Martin. 1997. *Indefinite pronouns*. Oxford: Oxford University Press.
Haspelmath, Martin. 1999. Why is grammaticalization irreversible? *Linguistics* 37(6). 1040–1068.
Heeschen, Volker. 1992. *A dictionary of the Yale (Kosarek) language (with sketch grammar and English index)*. Berlin: Reimer.
Hintikka, Jaakko & Jack Kulas. 1985. *Anaphora and definite descriptions: Two applications of game-theoretical semantics*. Dordrecht: Reidel.
Hockett, Charles F. 1958. *A course in modern linguistics*. New York: Macmillan.
Hunn, Eugene S., Akesha Baron & Roger Reeck. n.d. Un esbozo de la gramática del zapoteco de los pueblos Mixtepec, Oaxaca, México.
Ishiyama, Osamu. 2008. *Diachronic perspectives on personal pronouns in Japanese*. State University of New York at Buffalo. (Doctoral dissertation).
Jung, Ingrid. 2008. *Gramática del páez o nasa yuwe: Descripción de una lengua indígena de Colombia* (Languages of the World: Materials 469). Munich: LINCOM.
Keels, Jack. 1985. Guayabero: Phonology and morphophonemics. In Ruth M. Brend (ed.), *From phonology to discourse: Studies in six Colombian languages* (Language Data, Amerindian Series 9), 57–87. Dallas: Summer institute of linguistics.
Kibrik, Andrej A. 2011. *Reference in discourse*. Oxford: Oxford University Press.
Killian, Don. 2019. Gender in Uduk. In Francesca Di Garbo, Bruno Olsson & Bernhard Wälchli (eds.), *Grammatical gender and linguistic complexity: Volume I: General issues and specific studies*, 147–168. Berlin: Language Science Press. DOI:10.5281/zenodo.3462764
Kroeker, Menno H. 2001. *Gramática descritiva da língua nambikuara*. Cuiabá: Sociedade Internacional de Lingüística.
Lee, Alan. 2005. Verb morphology. In Nicholas Evans, Jutta Besold, Hywel Stoakes & Alan Lee (eds.), *Materials on Golin: Grammar, texts and dictionary*, 31–53. Melbourne: Department of Linguistics & Applied Linguistics, the University of Melbourne.
Levinson, Stephen C. 1987. Pragmatics and the grammar of anaphora. *Journal of Linguistics* 23(2). 379–434.
Lock, Arjen. 2011. *Abau grammar* (Data Papers on Papua New Guinea Languages 57). Ukarumpa: SIL-PNG Academic Publications.

Lowe, Ivan. 1999. Nambiquara. In Robert M. W. Dixon & Alexandra Y. Aikhenvald (eds.), *The Amazonian languages* (Cambridge Language Surveys), 269–292. Cambridge: Cambridge University Press.

Lowing, Richard & Aretta Lowing. 1975. *Awa dictionary* (Pacific Linguistics C 30). Canberra: The Australian National University.

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

Macaulay, Monica. 1996. *A grammar of Chalcatongo Mixtec* (University of California Publications in Linguistics 127). Berkeley: University of California Press.

Machal, Marcelo. 2000. Cuiba (Jiwi). In Esteban Emilio Mosonyi & Jorge Carlos Mosonyi (eds.), *Manual de lenguas indígenas de Venezuela* (Serie Origenes), 224–265. Caracas: Fundación Bigott.

Macri, Martha J. 1983. The noun class systems in Mixtec. In Alice Schlichter, Wallace L. Chafe & Leanne Hinton (eds.), *Studies in Mesoamerican linguistics*, vol. 4, 291–306. Survey of California & Other Indian Languages.

Mellow, Greg. 2013. *A dictionary of Owa: A language of the Solomon Islands*. Berlin: De Gruyter Mouton.

Merchán, Hernanda Ana Joaquina. 2000. Breve presentación de lengua cuiba (variante maibén). In María Stella González de Pérez & María Luisa Rodríguez de Montes (eds.), *Lenguas indígenas de Colombia: Una visión descriptiva*, 585–598. Santafé de Bogotá: Instituto Caro y Cuervo.

Milles, Karin. 2011. Feminist language planning in Sweden. *Current Issues in Language Planning* 12(1). 21–33.

Minor, Eugene E., Dorothy A. Minor & Stephen H. Levinsohn. 1982. *Gramática pedagógica huitoto*. Bogotá: Ministerio de Gobierno.

Mous, Maarten. 1992. *A grammar of Iraqw*. Rijksuniversiteit Leiden. (Doctoral dissertation).

Nichols, Johanna. 1992. *Linguistic diversity in space and time*. Chicago: University of Chicago Press.

Paradis, Carole. 1983. *Description phonologique du guéré*. Abidjan: Université d'Abidjan, Institut de linguistique appliquée.

Payne, Doris L. 1985. *Aspects of the grammar of Yagua: A typological perspective*. University of California, Los Angeles. (Doctoral dissertation).

Plank, Frans & Wolfgang Schellinger. 1997. The uneven distribution of genders over numbers: Greenberg Nos. 37 and 45. *Linguistic Typology* 1(1). 53–101.

Sándalo, Filomena. 1997. *A grammar of Kadiwéu with special reference to the polysynthesis parameter* (MIT Occasional Papers in Linguistics 11). Cambridge: Massachusetts Institute of Technology.

Sanders, Arden G. & Joy Sanders. 1994. Kamasau (Wand Tuan) grammar: Morpheme to discourse. Unpublished document.

Sarvasy, Hannah Sacha. 2014. *A grammar of Nungon: A Papuan language of the Morobe Province, Papua New Guinea*. Cairns: James Cook University. (Doctoral dissertation).

Scott, Graham. 1989. *Fore dictionary* (Pacific Linguistics C 62). Canberra: The Australian National University.

Seifart, Frank. 2018. The semantic reduction of the noun universe and the diachrony of nominal classification. In William B. McGregor & Søren Wichmann (eds.), *The diachrony of classification systems*, 9–32. Amsterdam: John Benjamins.

Siewierska, Anna. 2005. Alignment of verbal person marking. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *The world atlas of language structures*, 406–409. Oxford: Oxford University Press.

Small, Priscilla C. 1990. A syntactic sketch of Coatzospan Mixtec. In C. Henry Bradley & Barbara E. Hollenbach (eds.), *Studies in the syntax of Mixtecan languages 2* (Summer Institute of Linguistics and the University of Texas at Arlington Publications in Linguistics 90), 261–479. Dallas: Summer Institute of Linguistics & the University of Texas at Arlington.

Speck, Charles H. 1972. *The study of Zapotec language and culture*. SIL Language & Culture Archives. oai:sil.org:59283.

Stark, Sharon L. 2011. *Ngigua (Popoloca) pronouns* (SIL-Mexico Branch Electronic Working Papers 12). SIL.

Thurman, Robert C. 1987. *The form and function of Chuave clauses*. SIL Language & Culture Archives.

Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge: Harvard University Press.

Trudgill, Peter. 2011. *Sociolinguistic typology: Social determinants of linguistic complexity*. New York: Oxford University Press.

Trussel, Stephen. 1979. *Kiribati (Gilbertese): Grammar handbook* (Peace Corps: Language Handbook Series). Brattleboro VY School for International Training. http://www.trussel.com/f_kir.htm#Gil.

Vafaeian, Ghazaleh. 2013. Typology of nominal and adjectival suppletion. *Sprachtypologie und Universalienforschung STUF* 66(2). 112–140.

van Driem, Georg. 1987. *A grammar of Limbu*. Berlin: Mouton de Gruyter.

Vincent, Lois E. 2010. *Tairora-English dictionary*. Wycliffe Papua New Guinea Branch.

Wälchli, Bernhard. 2018. The rise of gender in Nalca (Mek, Tanah Papua): The drift towards the canonical gender attractor. In Sebastian Fedden, Jenny Audring & Greville G. Corbett (eds.), *Non-canonical gender systems*, 68–99. Oxford: Oxford University Press.

Wälchli, Bernhard. 2019. The feminine anaphoric gender gram, incipient gender marking, maturity, and extracting anaphoric gender markers from parallel texts. In Francesca Di Garbo, Bruno Olsson & Bernhard Wälchli (eds.), *Grammatical gender and linguistic complexity: Volume II: World-wide comparative studies*, 61–131. Berlin: Language Science Press. DOI:10.5281/zenodo.3462780

Wälchli, Bernhard & Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics* 50(3). 671–710.

Wälchli, Bernhard & Francesca Di Garbo. 2019. The dynamics of gender complexity. In Francesca Di Garbo, Bruno Olsson & Bernhard Wälchli (eds.), *Grammatical gender and linguistic complexity: Volume II: World-wide comparative studies*, 201–364. Berlin: Language Science Press. DOI:10.5281/zenodo.3462784

Zavala, Roberto. 1992. *El kanjobal de San Miguel Acatán* (Colección Lingüística Indígena 6). Mexico City: Universidad Nacional Autonoma de México, Instituto de Investigaciones Filológicas, Seminario de Lenguas Indígenas.

Appendix A: Languages in the sample with anaphoric gender and the automatic extraction from parallel texts

>x<: morphemes, #: word boundary

I. Languages with a non-mature feminine anaphoric gender gram [59 languages]

Table 11: Languages with a mature feminine anaphoric gender gram

Language	Extracted form	Remarks
Esperanto (epo)	sxi, >#sxi<	*sxi-n* ACC, *sxi-a*-AGR POSS
Malayalam (mal)	avall, avallodu, avalle	*avall* NOM, *avall-e* ACC, *avall-odu* INST
Japanese (jpn)	kanojo	*kano-jo* PROX-woman
Wè Northern (wob)	ʋ, ʋa’	*ʋ-a(‘)* POSS, object *ʋ(‘)*, -‘ intransitivizer, also after object pronouns (Paradis 1983)
Uduk (udu) [artificial variety of Bible translation]	yim, ayim	*yim* ‘female friend’ (noun)
Zome (zom)	tuanu	*tua-nu* DIST-mother, *hih nu* PROX mother
Naga, Angami (njm)	{süpfü}	*sü-pfü* DEM-F
Khmer, Northern (kxm)	เนีย	*niang* young female person
Kiribati (gil)	neierei, nei	*neierei* F.DIST, *Nei* female person name marker
Owa (stn)	{kani}	*kani* ‘that woman; wife’
Naasioi (nas)	teni, tenie	*teni-e* ERG
Ankave (aak)	i'	*i’* F
Chuave (cjv)	oparomi	*opa-rom-i* woman-?-DIST
Golin (gvf)	abalini	*abal-ini* woman-REFL
Oksapmin (opm)	uh, uhnong, uhe, {urhe}	*uh* F, *oh* M, *uh-nong* ACC, *uh-e* GEN, *urhe* REFL.GEN (M *orhe*)
Chuj (cac)		*'ix* woman, noun classifier for woman
Jacaltec (jac)	ix	*ix* woman, noun classifier for woman
Akateko (knj)	ix	*ix* woman, noun classifier for woman
Ixil, Nebaj (ixi)	ixoj	*ixoj(e)* woman
Mam, Todos Santos (mvj)	>xuj#<	*xuj* ‘old woman’, *txin* young woman, *te-* to
Cuicatec, Teutila (cut)		*tɑ̰hn* full form, *te* reduced form

Language	Extracted form	Remarks
Cuicatec, Tepeuxila (cux)	tá, tá^{n}'ã, ta	*tá^{n}'ã* full form, *tá/ta* reduced form
Mixtec, Atatlahuca (mib)	ña	*ña* F
Mixtec, Ocotepec (mie)	ña	*ña* F
Mixtec, San Miguel (mig)	>-ñ<	*-ña* F
Mixtec, Peñoles (mil)	>-a^{n}#<	*-a^{n}* F
Mixtec, Pinotepa Nacional (mio)	ña	*ña* F
Mixtec, Southern Puebla (mit)		*-nè*, *-ne*, *-né*, *-ñá*, *-ña* F
Mixtec, Coatzospan (miz)		*tún* F (girls), adult respect *ña*
Mixtec, San Juan Colorado (mjc)	ña	*ña* F
Mixtec, Silacayoapan (mks)	ñá	*ñá* F
Mixtec, Yosondúa (mpm)	ña	*ña* F
Mixtec, Tezoatlan (mxb)	>án#<	*án*, *-án* F
Mixtec, Jamiltepec (mxt)	ña	*ña* F
Mixtec, Diuxi-Tilantongo (xtd)	>-ña<	*-ña*, F *nuu* 'to'
Triqui, Copala (trc)	no'	*no'* F
Triqui, San Martin Itunyoso (trq)	ún'	*ún'* F
Popoloca, San Marcos Tlalcoyalco (pls)		*nëha* 'woman[ANA]', *xan* 'child, child[ANA]'
Mazatec, Chiquihuitlan (maq)	na	*na* F
Zapotec, Ozoltepec (zao)		*nzaa* girl
Zapotec, Quioquitani Quieri (ztq)	me	*me* F
Zapotec, Rincon (zar)	>nu<	*-nu* F

Language	Extracted form	Remarks
Zapotec, Southern Rincon (zsr)	>nu<	*-nu* F
Zapotec, Santo Domingo Albarradas (zas)		*-m* F
Zapotec, Lachixio (zpl)	>nchu#<	*-nchu* F
Zapotec, Amatlan (zpo)	me	*me* F, *xaa* HONOR
Zapotec, Texmelucan (zpz)		*fiñ*, *ñi*, *-ñ* F, *mi*, *-m* RESPECT
Cuiba (cui)	barapowa	*barapowa*, *bapowa*
Guahibo (guh)	bajarapova	*bajarapova*, *barapova*
Guayabero (guo)	>ow#<	V-*ow*, N-*ow*, free form *japow*
Kaingang (kgp)	fi	*fi*
Rikbaktsa (rkb)	atatsa, >tatsa#<	*atatsa* 3SG.F, *-tatsa* F
Nambikuara, Southern (nab)	ta¹ka³lxai²na²	*ta¹ka³lx-ai²na²* F-DEM, *ta¹ka³lx-a²* F-DEF
Kayabi (kyz)	ẽẽ, {kĩã}	*ẽẽ* F (M speaker) M, *kyna* F (F speaker), M *ʻga* (M speaker), and *kĩã* M (F speaker)
Tenharim (pah)	hẽa	*hẽa* F
Muinane (bmr)	diigoco, >go<	*-go* F
Bora (boa)	>lle<	*-lle* 'F'
Huitoto, Minica (hto)	afengo, {aféngona}	*afe-ngo* DIST-F
Huitoto Murui (huu)	>ñaiñ<	*nai-ñaiño* DIST-F, *bi-ñaiño* PROX-F

II. Languages with a mature feminine anaphoric gender gram [128 languages]

See Table 12.

Table 12: Languages with a mature feminine anaphoric gender gram

Language	Extracted form	A	S	P	R	Poss1	Poss2
Kannada (kan)	>aḷu#<, āke, ākege	V-aḷu, avaḷu, āke(yu)	”	avaḷannu, ākeyannu	avaḷige, ākege	avaḷa, ākeya	-
Tamil (tam)	>ாள்#<, அவள், {-அவளது}	avaḷ, V-āḷ	”	avaḷ-ai	avaḷ-iṭum/ukku	avaḷ-atu	-
Albanian, Gheg (aln)	ajo, saj	ajo	”	-	-	saj	-
Latvian (lav)	viņai, {-usi}	viņa, tā, (V-usi, V-dama)	”	-	viņai	viņas	-
Lithuanian (lit)	ji, jai, ją, {>usi#<}	ji, (V-usi, V-dama)	”	ją	jai	jos	-
Breton (bre)	he, dezhi	hi	”	anezhi	dezhi	he	”
Welsh (cym)	hi, iddi, {wrthi}	hi, hithau, iddi	”	hi, hithau	wrthi, iddi	ei+ASP	”
Norwegian, Bokmål (nor)	hun, henne	hon	”	henne	”	hennes	-
Danish (dan)	hun, hende	hon	”	hende	”	hendes	-
Swedish (swe)	hon, henne	hon	”	henne	”	hennes	-
Faroese (fao)	hon, hana, henni	hon	”	henni	hana	hennara	-
Icelandic (isl)	hún, hana, hennar, {henni}	hún	”	henni	hana	hennar	-
English (eng)	her, she	she	”	her	”	”	”
English, Middle (enm)	hir, sche	sche	”	hir	”	”	”
German, Standard (deu)	sie, ihr	sie	”	”	ihr	ihr-AGR	”
Alemannic (swg)	sie	sie	”	”	ihr	ihr-AGR	”
Afrikaans (afr)	haar	sy	”	haar	”	”	”
Dutch (nld)	haar, zij	zij	”	haar	”	”	”
Saxon, Low (nds)	ää, see, äare	see	”	ää	”	äar-AGR	”
Greek (ell)	της, {αυτήν}	εκείνη	”	την	της	της	”
Greek, Koine (grc)	αυτης, αυτη, αυτην	η, εκεινη, V-ουσα	”	αυτην	αυτη	αυτης	”
Gujarati (guj)		-	V-ī	V-ī	-	-	-
Panjabi, Eastern (pan)		-	V-ī	V-ī	-	-	-
Romani, Sinte (rmo)	joi, li, late, lat, {lakro}	joi, koi	”	lat	late	lakr-AGR	-
Romani, Vlax (rmy)	lake, woi, la	woi	woi, V-i	la	lake	lak-AGR	-
Hindi (hin)		-	V-ī	(V-ī)	-	-	-
Marathi (mar)	तिला, ती, तिने	tine, tī /V-ī / -	tī /V-ī	ti-lā / V-ī / -	ti-cyā	-	
Kurdish, Northern (kmr)	wê	wê	-	-	wê	”	-
Latin (lat)	eam	illa, quæ, hæc	”	eam, illam	-	-	-
Romanian (rmo)	ea	ea	”	(o)	ei	”	-
Italian (ita)	ella	ella, essa	”	la	le	-	-
French (fre)	elle	elle	”	la	-	-	-
Catalan-Valencian-Balear (cat)	ella	ella	”	la	-	-	-
Spanish (spa)	ella	ella	”	la	-	-	-
Portuguese (por)	ela	ela	”	a	-	-	-
Russian (rus)	>ла#<, она, её, ей	она, V-ла	”	её	ей	её	-
Ukrainian (ukr)	вона, >ла#<, її, неї, {їй}	вона, V-ла	”	її	їй	її	-
Bulgarian (bul)	й, тя, я	тя	”	я	й	-	-
Slavonic, Old Church (chu)	ей, ю, ея	она	”	ю	ей	ея	-
Croatian (hrv)	joj, ona, {>la#<}	ona, V-la	”	je	joj	-	-
Czech (ces)	>ła#<, jí, {ji, její}	ona, V-la	”	ji	jí	její	-
Polish (pol)	>ła#<, jej, ją, {ona}	ona, V-ła	”	jej	”	její	-
Avar (ava)	гьей, гьелда, гьелъ	гьелъ	гьей / й-V / -	”	гьелда	гьелъул	-
Chechen (che)		-	-	- / j-V	-	-	

Language	Extracted form	A	S	P	R	Poss1	Poss2
Tachelhit (shi)	nttat	t-V, nttat	”	=tt	-	-	-
Kabyle (kab)	{>#te<}	te-V	”	=ṭ	-	-	-
Tamasheq (taq)	>tăt<	tĕ-V	”	=tăt	-	-	-
Bana (bcw)	ngəta, nzə	ghənzə / -	V-ta, ghənzə	”	ngə-ta	N-ta, N nzə	N-ta
Gude (gde)	ki, kya	kya	”	ka ki	”	N-tə	-
Dangaléat (daa)	>iit<, >ti#<, ta	ta / te-V			V-t	(not mother)	
Hausa (hau)	ta, >ta#<	ta V, ita	”	V ta	mata	N-ta	”
Mwaghavul (sur)	wura, fira, nwura, {yi}	wura	”	wura	nwura/wura	fira	”
Somali (som)	>eed<, {>say#<}	t-V, V-tay/say, iyada	”	-	-	N-eed	”
Iraqw (irk)		V-eer, V-VC, V-VVn	”	various	-	-	-
Dawro (dwr)	>aaddu#<, izo, izi, iza	iza, V-aaddu	”	izo	izo		
Gamo (gmv)	>adus#<, izis, {izo}	iza, V-us	”	izo	izos, izo	izi	-
Gofa (gof)	>u#<, {iyo}	iya, V-asu	”	iyo	iikko		-
Wolaytta (wal)	>su#<, o	a, V-aasu	”	o	iyyo, o	i	-
Kafa (kbr)	>an#<, >qqa#<	V-an	”	- ?	- ?	- ?	-
Maltese (mlt)	>ha#<	V-et	”	? -h/-tu	V-ha	N-ha	”
Amharic (amh)	>äčə<, >əwamə#<, >atə#<	V-äč	”	V-at	”	N-wa	”
Jur Modo (bex)	låko, ’bënï	låko	”	nï	zï-nï	bënï	N-nï
Belize Kriol English (bex)	shee	shee	”	-	-	-	-
Hawaiian Pidgin (hwc)	her, she	she	”	her	”	”	”
Burarra (bvr)	achila, >#ji<, >ny-<	-	jiny(u)-V	jiny(u)-V	achila	acha	”
Galela (gbi)	muna, ami, >#mo<, >mi<, munaka	mo-V, muna	”	-mi-V	munaka	ami	-
Tabaru (tby)	>#mo<, muna, gumuna, ’ami, mi, ngo	mo-V, (gu)muna ?	”	-mi-V	munaka	ami	-
Tobelo (tbl)	münanga, >#mo<, >#ami<, ngo	mo-V, münanga	”	-mi-V	munangika	ami	-
Rotokas (roo)	oira, >aev<, oirare	V-o-, (oira)	”	oira	oira-re	oira	”
Qaqet (byx)	qia, qi, ara, ki, kia	qia	”	qi	”		
Kuot (kto)	>ieng#<, iang	i-	”	V-ieng	o-	ieng	”
Yawa (yva)	mo	m-V / mo	m-V / mo / r-V	r-V	r- /rai	ama	”
Ama (amm)	isoki	-	V-so-	”	V-so-	-	-
Ambulas (abt)	léku, lat, >lé<	lé	”	lérét	?	léku	”
Iatmul (ian)	>li#<, lila	li, V-li	”	li	li		
Kwoma (kmo)	siina, sii, siiti, {siita}	sii	”	siina	”	siiti	”
Kwanga (kwj)	tini	ti	”	tini	”	ti	”
Mende (sim)	si, simu, sirin	si(mu)	”	sirin	”	sihi	”
Yessan-Mayo (yss)	te, tene, teri	te	”	tene	”	”	”
Abau (aau)	hoko, hoke, sokwe	hok(we)	”	hoke/ke	”	hoko	”
Sepik Iwam (iws)	saeya, saiir	saeya	”	saiir	”	”	”
Mufian (aoj)	>kw<, ako’w, >’w<, >ko<	kw(a)-V / ako’w	”	V-’w	-akw	N -kw/’w	”
Bumbita Arapesh (aon)	okwok, kwape, nakripok, >#k<	kw(a)-V, okwok	”	V-k	okwudok	okwokwik	”
Bukiyip (ape)	>ok<, >#kw<, >#ku<	kw(a)-V, okwok	”	V-k	-p-ok	okwokwik	”
Kamasau (kms)	wuso	w-V, <q>	”	stem inflection	-w	wung	-
Au (avt)	hire, >iwe#<, >#we<, {>iye#<}	hire / w-V	”	V-p	-we	AGR-ire	”
Olo (ong)	ne, >ene<	ne / n-V	”	V-(e)ne	”	pene	”
Yonggom (yon)	yu, >uun<, >een#<	V–een / yu(-mbed)	V–een / yu	V-end- / yu	”	yu	”
Bimin (bhl)	>u#<, >koum<, >ui#<, ulo, um, {wangei}	V–((e)l)u	”	wa-/w-/we-/um-/wam-V	”	um-	-
Faiwal (fai)	uka, >mam<, >#wak<, nadule, {ulum, um}	?	?	u-V	”	ulum	”
Mian (mpt)	o, baabonea	V—o	”	wa-V	V-bo	o	”
Ngalum (szb)	u, ua, >du<, >ukhe<, uede	?	?	?	?	u	”

Language	Extracted form	A	S	P	R	Poss1	Poss2
Telefol (tlf)	>lu#<, tal	V-nulu	”	u-V		umi	”
Bine (bon)	jojige	-	Co-	”	”	-	-
Paumari (pad)	>’ihi#<	-	V-’i-hi	”	”	-	-
Garifuna (cab)	>#t<	t-V	”	-, tugía	t-un	t-N	”
Wayuu (guc)	shia, sümüin	s/sh-V	”	shia	sü-müin	sü-N	”
Piapoco (pio)	>#u<, úa	u-V	”	úa	u-lí	ú-N	”
Yucuna (ycn)	>#ru<, >#ro<	ru-V, V-yo ?		rucá	ro-jló	ru-N	”
Ignaciano (ign)	>#su<	-, su-V, esu	”	esu	”	su-N	”
Trinitario (trn)	esu, >#s<	-, mue-V	”	esu	?	sa-N	”
Ashéninka Pajonal (cjo)		o-V	”	V-ro	V-ro	Ø-N	”
Asháninka (cni)		o-/Ø-V	”	V-ro	V-ro	Ø-N	”
Caquinte (cot)		o-/Ø-V	”	V-ro	V-ro	o-/Ø-N	”
Ashéninka, Pichis (cpu)		o-/Ø-V	”	V-ro	V-ro	o-/Ø-N	”
Machiguenga (mcb)	irorori	o-/Ø-V, iroro(ri)	”	V-ro	V-ro	o-/Ø-N	”
Nomatsiguenga (not)		o-/po-/Ø-V	”	V-ro	V-ro	o-/Ø-N	”
Apurinã (apu)	oa, >#o<, >aro<	o-/Ø-V	”	V-ro	V-ro	o-/Ø-N	”
Yine (pib)	>#t<, wala, {chinro}	t-V, wala	”	V-Lo	V-Lo	t-N	”
Chiquitano (cax)	ɨmo	V-ti	”	ɨmo	ɨmo	ni-N-x-Ø/ IRREG	”
Cacua (cbv)	mi, mit, caántdih, miíh	mit, caántboó	mi	mi, caántdih	caántdih	mi	”
Yagua (yad)	>#nanu<	nan- / -	nan-/ naada / -	naada / -	nan- / -	-	-
Ticuna (tca)	>#ngi <, >#iya<	?	?	?	?	?	?
Tsimane (cas)	mọ’, {je’}	mọ	”	V-’ / mọ’	V-’	mọ’	”
Cubeo (cub)	ő́re, >jaco<, õ, õi	V-(a)co /V- / õ	”	ő́-re	”	jí-N	”
Waimaha (bao)	cõ, cõre, >upo#<, >go<, >mo<, {>rico<}	V-Co / cõ	”	cõ-re	”	cõ	”
Tuyuca (tue)	coo, >go#<, coore	V-Co / coo	”	coo-re	”	coo	”
Desano (des)	igo, igore, >go<, >mo#<, >po#<,	V-gõ/mo / igo	”	igo-re	”	igo	”
Siriano (sri)	igo, igore, >yupo#<, >mo#<, >deo#<, {igoya}	V-gõ/mo / igo	”	igo-re	”	igo	”
Barasana Eduria (bsn)	so, sore, >mo#<	V-Co / so	”	so-re	”	so	”
Macuna (myy)	iso, isore, >yijo#<	V-Co / iso	”	isore	”	iso	”
Carapana (cbc)	cõ, >upo#<, >ñupõ#<, >mo#<	V-Co / V- / cõ	”	cõ	”	cõ	”
Tatuyo (tav)	co, >upo#<, cõre	V-wõ / co	”	(cõre) co V	co	co	”
Piratapuyo (pir)	>icoro<	ticoro / V-?	”	ticoro-re	”	ticoro	”
Tucano (tuo)	koô, koôre, >ko#<, >go<, nïïwõ, >mo#<	V-Co / koô	”	koô-re	”	koô	”
Koreguaje (coe)	>mo#<, repao, repao’te, >si’ko#<, {chikona}	V-mo / repao	”	repao-‘te	”	repao	-
Siona (snn)	>go<, {>si’co<}	V-Co / bago	”	bago-ni	”	bago	”
Chipaya (cap)	na, >inćha#<, näża, {näkiś, näki}	V-inćha / V- / näki	na	”	näkiś	näża , ź-N	”

III. Languages with feminine person name markers, wrongly extracted [6 languages]

Language	Extracted form	Remarks
Uab Meto (aoz)	{bi}	*bi* N, with feminine person names
Iraya (iry)	bayi	*bayi* N, with feminine person names
Huave (huv)	{müm}	*müm* 'mother' used with feminine person names
Satere-Mawe (mav)	mana	*mana* N, with feminine person names
Nalca (nlc)	gera	*ge-ra* F-TOP, also with feminine person names

IV. Languages with wrongly extracted demonstrative/definite forms for 'woman' [9 languages]

Language	Extracted form	Remarks
Sabaot (spy)	:cheebyoosyaanaa	*cheebyoosya* 'woman'
Endo (enb)	cheepyoosoonoonēē	*cheepyooso* 'woman'
Mazatec, Ayautla (vmy)	chjunbiu	*chjun* 'woman'
Djambarrpuyngu (djr)	{miyalknhany}	*miyalk* 'woman'
Safeyoka/Wojokeso (apz)	a'musi	*a'mu* 'girl'
Fasu (faa)	{hinamoamo}	*hinamo* 'woman', *-amo* "referent subject"
Umbu-Ungu (ubu)	ambomo	*ambo* 'woman', *-mo* 'the'
South Tairora (omw)	nraakyeva	*nraakye* 'woman', *-ve* DEM
Rawa (rwo)	barega	*bare* 'woman', *-ga* DEF.SG

V. Wrongly extracted forms for 'woman' [1 language]

Language	Extracted form	Remarks
Awa (awb)	{iní, mi}	*iní* 'woman[ABS]'; *mi* 'that'

VI. Wrongly extracted demonstratives and articles (without or with gender) [5 languages]

Language	Extracted form	Remarks
Mountain Koiali (kpx)	{keu}	*ke-u* [that-SUBJECT]
Folopa (ppo)	kale	'the'
Fore (for)	kana	*kana-* 'this mentioned one, the aforementioned'
Kadiweu (kbc)	nagajo	
Mocoví (moc)	aso'maxare	*a-so'-maxare* F-GOING-PRO

VII. Wrongly extracted general third person forms [2 languages]

Language	Extracted form	Remarks
Zapotec, Miahuatlan (zam)	{xa'}	*xa'* 3 M/F, *mza'* girl
Zapotec, Chichicapan (zpv)	bi	*bi* 3 M/F, *ba* 3.RESPECT

VIII. Entirely wrongly extracted forms

Language	Extracted form	Remarks
Buglere (sab)	{chku}	*chku* arrive.PFV

Appendix B: Languages in the sample without any feminine anaphoric gender gram [629 languages]

Phyla or families and ISO 639-3 codes Languages with only wrongly extracted forms (Appendix A III-VII) are included and underlined.

Creoles and artificial languages

Creoles (12/14): acf, bis, djk, hat, kri, mbf, mfe, pis, rop, srm, srn, tpi

Artificial languages (0/1)

Eurasia

Altaic (10/10): aze, bxr, kaa, kaz, krc, kum, tat, tur, uzb, xal

Basque (1/1): eus

Dravidian (0/3)

Indo-European (10/50): awa, hif, hns, hye, mai, ory, oss, pes, prs, tgk

North Caucasian (1/3): tab

Korean (1/1): kor

Japanese (0/1)

Uralic (7/7): est, fin, hun, kpv, mhr, myv, sme

Africa

Afro-Asiatic (7/24): gnd, hig, meq, mfh, mfi, mif, pbi

Niger-Congo (126/127): acd, adj, ann, anv, atg, bam, bav, bba, bfd, bib, bim, biv, blh, box, bqc, bss, bud, bwq, bwu, cce, cko, cme, csk, cwt, dgi, dnj, dop, dts, dug, dyo, dyu, ewe, fal, fub, fuv, gbo, gej, gkn, gng, gog, gur, gux, guz, hag, hay, heh, izr, jbu, kao, kbp, ken, kez, kik, kin, kki, kkj, kma, kng, knk, kno, kub, kus, las, ldi, lee, lef, lem, lia, maw, mcu, mda, men, mfq, mnf, mnk, mos, muh, myk, mzk, mzm, mzw, ncu, ndz, neb, nfr, nhu, nim, nko, nnw, nso, ntr, nya, nyf, nyy, old, ozm, pkb, rim, sbd, sig, sil, sld, sna, soy, spp, sus, swh, swk, tbz, tem, thk, tik, toh, tum, vag, wmw, wol, vun, xho, xon, xrb, xsm, yal, yam, yor, zul

Nilo-Saharan (14/15): avu, bjv, dik, dip, dje, enb, kyq, lwo, mfz, mur, nus, shk, spy, udu

East Asia and Southeast Asia

Austroasiatic (2/3): bru, vie

Austronesian (132/134): aai, ace, adz, agn, aia, akb, alp, aoz, ban, bbc, bcx, bdd, bhp, bku, blz, bnp, bpr, bps, btd, bth, bto, bts, btx, bug, buk, bzh, ceb, cha, dad, dob, dww, fij, gfk, gor, haw, hil, hla, hnn, hot, hvn, iba, ifb, ifk, ifu, ilo, ind, iry, itv, jav, jvn, kbm, khz, kne, krj, kud, kwf, kzf, lcl, lcm, leu, lew, lid, ljp, mad, mah, mak, mbb, mbt, mee, mek, mhy, min, mlg, mmo, mmx, mna, mnb, mog, mox, mpx, mqj, mri, msm, mta, mva, mwc, mvp, mwv, nak, nia, nij, npy, nsn, pag, pam, plt, pmf, ppk, prf, pse, ptp, ptu, pwg, rai, rro, sas, sbl, sda, sgb, sgz, smk, sml, smo, snc, sps, sso, sun, swp, sxn, tbc, tbo, tgl, tpz, tte, twu, urk, uvl, war, wuv, xkl, xsi, zlm

Hmong-Mien (1/1): mww

Sino-Tibetan (22/24): acn, ahk, bgr, cfm, cmn, cnh, cnk, cnw, csy, ctd, czt, grt, hlt, kac, kyu, lhu, lif, mhx, mwq, nan, pww, taj

New Guinea and Australia

Australian (3/4): djr, gvn, wim

East Bird's Head (3/3): mej, mnx, mtj

East Papuan (2/6): sua, yle

Geelvink Bay (1/2): bvz

Karkar-Yuri (1/1): yuj

Arai (Left May) (0/1)

Sepik-Ramu (2/10): msy, sny

Torricelli (0/6)

Trans-New Guinea (79/90): aey, agd, agg, amn, aom, apz, aso, auy, awb, bbr, bef, big, bjr, bmh, bmu, boj, byr, dah, ded, dgz, faa, for, gaw, gdn, ghs, hui, imo, iou, ipi, kgf, kjs, kmh, knv, kpf, kpr, kpw, kpx, ksr, kue, kyc, kyg, mcq, med, mhl, mlh, mlp, mps, mux, naf, nca, nii, nlc, nop, nou, nvm, okv, omv, ppo, rwo, sll, snp, soq, ssd, ssx, sue, tim, ubu, waj, wer, wiu, wnc, wnu, wsk, xla, yby, yli, yut, yuw, zia,

West Papuan (0/3)

North and Mesoamerica

Algic (1/1): ojs

Eskimo-Aleut (2/2): esk, kal

Hokan (1/1): chd

Huavean (1/1): huv

Iroquoian (1/1): chr

Mayan (23/28): acc, acr, agu, caa, cak, chf, cke, ctu, hus, hva, ixl, kek, lac, mam, mop, mvc, poh, toj, ttc, tzj, tzo, tzt, usp

Mixe-Zoque (8/8): mco, mir, mto, mxp, mxq, mzl, poi, zos

Na-Dene (3/3): caf, crx, gwi

Otomanguean (36/63): amu, azg, cco, chq, chz, cle, cnl, cnt, cpa, cso, ctp, cuc, cya, maa, mau, maz, ote, otm, otn, otq, vmy, zab, zac, zad, zai, zam, zat, zav, zaw, zpc, zpi, zpm, zpq, zpu, zpv, zty

Totonacan (5/5): tku, toc, too, top, tos

Uto-Aztecan (18/18): azz, crn, hch, ncj, ncl, ngu, nhe, nhg, nhi, nhw, npl, ntp, ood, pao, stp, tac, tar, yaq

South America

Arauan (0/1)

Araucanian (1/1): arn

Arawakan (3/17): ame, pab, ter

Aymaran (1/1): ayr

Barbacoan (4/4): cbi, cof, gum, kwi

Cahuapanan (1/1): cbt

Camsa (1/1): kbh

Candoshi-Shapra (1/1): cbu

Carib (7/7): ake, apy, bkq, car, hix, pbc, way

Chibchan (8/8): bzd, cjp, con, gym, kvn, sab, tfr, tuf

Choco (3/3): emp, noa, sja

Guahiban (0/3)

Harakmbet (1/1): amr

Jivaroan (3/3): acu, hub, jiv

Macro-Ge (4/7): apn, mbl, txu, xav

Maku (1/2): mbj

Mataco-Guaicuru (2/3): kbc, mzh,

Nambiquaran (0/1)

Paez (1/1): pbb

Panoan (7/7): cao, cbr, cbs, kaq, mcd, shp, yaa

Peba-Yaguan (0/1)

Quechuan (25/25): inb, qub, quf, qug, quh, qul, qup, quw, quy, quz, qvc, qve, qvh, qwh, qvi, qvm, qvn, qvo, qvs, qvw, qvz, qxh, qxn, qxo, qxr

Tacanan (3/3): cav, ese, tna

Ticuna (0/1)

Tol (1/1): jic

Tsimane (0/1)

Tucanoan (0/13)

Tupi (10/12): gnw, gug, gui, gun, gyr, kgk, mav, myu, srq, urb

Urarina (1/1): ura,

Uru-Chipaya (0/1)

Waorani (1/1): auc

Witotoan (0/4)

Yanomam (1/1): wca,

Yuracare (1/1): yuz

Zaparoan (1/1): arl