Predicting the gender of Welsh nouns

Michael Hammond*

Predicting the gender of Welsh nouns

DOI 10.1515/cllt-2015-0001

Abstract: Welsh grammatical gender exhibits several unusual properties. This paper argues that these properties are necessarily connected. The argument is based on a series of corpus investigations using techniques from statistical natural language processing, specifically distinguishing properties that exhibit significant statistical patterns from those which can be used to make useable predictions. Specifically, it's shown that the grammatical properties of Welsh gender are such that its unusual statistical properties follow.

Keywords: Welsh, grammatical gender, learning, mutation

1 Introduction

Welsh gender exhibits three quite interesting properties:

(1) a. There is a numerical asymmetry between masculine and feminine nouns, with masculines greatly outnumbering feminines.
b. There is a fairly high number of nouns with indeterminate gender, or gender that differs across speakers or dialects.
c. The cues for gender are quite indirect, not typically exposed in the morphology of the noun, but more typically in mutation options.

In this paper, I argue that these three properties are connected. Specifically, I show that the indirect nature of gender marking in Welsh (1c) entails the other two properties: (1a) and (1b).

The empirical basis of this claim comes from an examination of a number of statistical models of gender prediction built using data from the CEG corpus (Ellis et al. 2001). While there are a number of interesting and significant differences between gender classes in Welsh in terms of various properties, our model focuses on the more difficult goal of gender category prediction. (See also Cucerzan and Yarowsky 2003.) Thus, while the gender of nouns correlates with a number of phonological and morphological properties, it's

Corresponding author: Michael Hammond, Department of Linguistics, University of Arizona, Tucson, Arizona 85721, USA, E-mail: hammond@u.arizona.edu

shown here that it is the mutation properties that are most probative in terms of how gender can be learned.

More critically, however, the results here show that in terms of predicting gender, properties other than mutation are less successful than simply guessing that every word is masculine gender. This in turn makes very powerful predictions about the fate of individual words and about the fate of the gender system of the language more generally.

From a theoretical perspective, these results suggest a conception of grammar where morphological, syntactic, and phonological properties are interdependent, where the structure of each grammatical module is contingent on that of the others.[1] In addition, the model developed makes predictions about how Welsh might change over time (compare Scottish Gaelic; Dorian 1976), and how it is acquired by children (Gathercole and Thomas 2001).

Finally, from a practical perspective, the results demonstrate that techniques from computational linguistics, from statistical natural language processing specifically, can be of great use in understanding language structure, acquisition, and historical development.

2 Basic facts

In this section, we review the basic facts of Welsh grammatical gender and how gender is expressed. Welsh has two genders: masculine and feminine (King 2003).[2] See Table 1. These are generally arbitrary, though terms for animals and people often bear the expected gender. See Table 2.

Table 1: Examples of masculine and feminine gender.

Masculine			**Feminine**		
pen	[pɛn]	'head'	llaw	[ɬaw]	'hand'
ci	[ki]	'dog'	coes	[kɔjs]	'leg'
ceffyl	[kɛfɨl]	'horse'	cath	[kaθ]	'cat'
gobaith	[gobajθ]	'hope'	ffatri	[fatri]	'factory'
afal	[aval]	'apple'	almon	[almɔn]	'almond'
	...			...	

1 See Daland et al. (2007) for a similar integrative view.

2 Readers familiar with the Welsh grammatical system can safely skim here or skip ahead to Section 3.

Table 2: Gender of animals and people.

Masculine			**Feminine**		
dyn	[dɨːn]	'man'	dynes	[dənɛs]	'woman'
mab	[mab]	'son'	merch	[mɛrx]	'girl'
tad	[tad]	'father'	mam	[mam]	'mother'
tarw	[taru]	'bull'	buwch	[bɨwx]	'cow'
brawd	[brawd]	'brother'	chwaer	[xwajr]	'sister'
ceffyl	[kɛfɨl]	'horse'	caseg	[kasɛg]	'mare'
ceiliog	[kejljɔg]	'rooster'	iâr	[jaːr]	'hen'
	...			...	

There is no overt affix that marks either gender; the gender of a noun is rarely deducible from the phonological or morphological form. Rather, gender is marked in a number of other ways, listed in Table 3. We review each of these below.

Table 3: Different ways of marking gender.

a.	pronouns and possessive adjectives
b.	soft mutation with article and the number 1 for fem. sg.
c.	mutation of adjectives with fem. sg.
d.	form of the numbers 2, 3, and 4
e.	form of certain adjectives
f.	form of demonstratives

Strikingly, there are a fair number of words of indeterminate gender. These are words where gender varies across dialects, speakers, or where the words are generally used in contexts where the gender is not overt. Some examples are given in Table 4. As discussed in Section 3 below, words like these are not a tiny

Table 4: Examples of words with indeterminate gender.

nifer	[nivɛr]	'number'
oed	[ɔjd]	'period'
amser	[amsɛr]	'time'
man	[man]	'place'
golwg	[gɔlʊg]	'sight'
rhyfel	[rəvɛl]	'war'
awdurdod	[awdɨrdɔd]	'authority'
arfer	[arvɛr]	'practice'
munud	[mɨnɨd]	'minute'
ystyr	[əstɨr]	'meaning'
...		

fraction of nouns; in our corpus, they comprise approximately 6% of occurring nouns.

Let's now review how gender is expressed in Welsh generally.

2.1 Pronouns

When a singular noun is referred to with a pronoun or with a possessive adjective, the grammatical gender of the noun is apparent.[3] The form of the third person differs for masculine, feminine, and plural. See Table 5. The pronominal form occurs in subject position and as the direct object of an inflected verb. It may also occur redundantly as the direct object of a verbal noun. The possessive form marks possession and the direct object of a verbal noun. See examples (2, 3, 4).

Table 5: Forms of the third person pronouns.

	Pronoun		Possessive	
he	(f)e/(f)o	[(v)ɛ/(v)o]	ei	[i]
she	hi	[hi]	ei	[i]
they	nhw	[nu]	eu	[i]

(2) Feminine:

a. *Dyma gath*. [dəma gaθ]
'This is a cat(fem.).'

b. *Mae* hi'*n fawr*. [maj hin vawr]
'It is big.'

c. *Mae* ei *phen (*hi) *yn fawr*. [maj i fɛn (hi) ən vawr]
'Its head is big.'

d. *Dw i'n* ei *gweld (*hi). [du in i gwɛld (hi)]
'I see it.'

e. *Gweles i* hi. [gwɛlɛs i hi]
'I saw it.'

3 There are dialect differences that affect many of these variables. We'll generally cite northern forms here, including other variants occasionally. Note that the possessive forms are characterized as adjectives here following King (2003), but it might be more appropriate to think of them as determiners.

(3) Masculine:

a. *Dyma gi*. [dəma gi]
'This is a dog(masc.).'

b. *Mae* o*'n fach*. [maj on vax]
'It is small.'

c. *Mae* ei *ben (o) yn fach*. [maj i bɛn (o) ən vax]
'Its head is small.'

d. *Dw i'n* ei *weld (*o). [du in i wɛld (o)]
'I see it.'

e. *Gweles i* fo. [gwɛlɛs i fo]
'I saw it'

(4) Indeterminate:

a. *Dyma ieir*. [dəma jejr]
'These are chickens.'

b. *Maen* nhw*'n fach*. [majn nun vax]
'They are small.'

c. *Mae* eu *pennau (*nhw*)'n fach*. [maj i pɛnaj (nu)n vax]
'Their heads are small.'

d. *Dw i'n* eu *gweld* nhw. [du in i gwɛld nu]
'I see them.'

e. *Gweles i* nhw. [gwɛlɛs i nu]
'I saw them.'

The possessive forms are phonologically identical in normal speech, but can sometimes be distinguished in terms of the mutations they trigger. Digressing, there are three mutations in Welsh schematized in Table 6. When words

Table 6: The three mutations.

Input	Soft	Nasal	Aspirate
p	b	m̥	f
t	d	n̥	θ
k	g	ŋ̊	x
b	v	m	n/a
d	ð	n	n/a
g	ø	ŋ	n/a
m	v	n/a	n/a
ɬ	l	n/a	n/a
r̥	r	n/a	n/a

beginning with the relevant consonants occur in appropriate morphosyntactic contexts, the changes indicated occur (Morgan 1952; Watkins 1961; King 2003; Hannahs 2013, etc.). Other consonants of the language are unaffected.

The possessive forms, rightmost column of Table 5, trigger different mutations on a following noun. The third singular masculine form triggers soft mutation, the feminine form triggers aspirate mutation, and the plural form triggers no mutation. The feminine and plural forms also trigger the insertion of an [h] on vowel-initial forms (see Table 7). When a form begins with a consonant that does not mutate, the only way the possessives can be distinguished is either by context or an independent pronoun placed after the noun (5).

Table 7: Mutations triggerd by possessive forms.

	mam 'mother'	**tad 'father'**	**afal 'apple'**
3sgm	ei fam [i vam]	ei dad [i dad]	ei afal [i aval]
3sgf	ei mam [i mam]	ei thad [i θad]	ei hafal [i haval]
3pl	eu mam [i mam]	eu tad [i tad]	eu hafal [i haval]

(5) a. *Gwelodd o* ei *nai*. [gwɛlɔð o i naj]
'He saw his/her nephew.'
b. *Dyna* ei *nai* hi. [dəna i naj hi]
'That's *her* nephew.'
c. *Lle mae* ei *ffôn o?* [ɬɛ maj i foːn o]
'Where is *his* phone?'
d. *Dyma* ei *ffôn*. [dəma i foːn]
'This is his/her phone.'

2.2 Soft mutation with the article

The gender of nouns is also apparent with the definite article. Specifically, if the definite article immediately precedes a feminine singular noun, the noun will undergo the soft mutation, as in (6). Hence the feminine noun *cath* [kaθ] 'cat' shows up as *gath* [gaθ] after the definite article *y* [ə]. The masculine *ci* [ki] 'dog' does not mutate here.

(6) a. *Dw i'n gweld ci*. [du in gwɛld ki]
'I see a dog (masc.).'
b. *Dw i'n gweld cath*. [du in gwɛld kaθ]
'I see a cat (fem.).'

c. *Dw i'n gweld y* ci. [du in gwɛld ə ki]
'I see the dog.'
d. *Dw i'n gweld y* gath. [du in gwɛld ə gaθ]
'I see the cat.'

This does not occur with the plural. Feminine plural nouns do not undergo soft mutation in this environment, as in (7).

(7) a. *Dw i'n gweld cŵn.* [du in gwɛld ku:n]
'I see dogs.'
b. *Dw i'n gweld cathod.* [du in gwɛld kaθɔd]
'I see cats.'
c. *Dw i'n gweld y cŵn.* [du in gwɛld ə ku:n]
'I see the dogs.'
d. *Dw i'n gweld y* cathod. [du in gwɛld ə kaθɔd]
'I see the cats.'

While Table 6 shows that [ɬ] and [r̥] undergo the soft mutation to become [l] and [r] respectively, they do not undergo soft mutation in this environment. Hence we have the contrast in (8, 9, 10); *cath* mutates after *y*, but *llinell* [ɬinɛɬ] 'line' and *rhaff* [r̥af] 'rope' do not.

(8) a. *Mae o'n gweld cath.* [maj ɔn gwɛld kaθ]
'He sees a cat(fem.).'
b. *Dyma hi.* [dəma hi]
'Here it is.'
c. *Dyma ei gath.* [dəma i gaθ]
'Here is his cat.'
d. *Mae o'n gweld y* gath. [maj ɔn gwɛld ə gaθ]
'He sees the cat.'

(9) a. *Mae o'n gweld llinell.* [maj ɔn gwɛld ɬinɛɬ]
'He sees a line(fem.).'
b. *Dyma hi.* [dəma hi]
'Here it is.'
c. *Dyma ei linell.* [dəma i linɛɬ]
'Here is his line.'
d. *Mae o'n gweld y* llinell. [maj ɔn gwɛld ə ɬinɛɬ]
'He sees the line.'

(10) a. *Mae o'n gweld rhaff*. [maj ɔn gwɛld r̥af]
'He sees a rope(fem.).'
b. *Dyma hi*. [dəma hi]
Here it is.'
c. *Dyma ei raff*. [dəma i raf]
'Here is his rope.'
d. *Mae o'n gweld y* rhaff. [maj ɔn gwɛld ə r̥af]
'He sees the rope.'

2.3 Soft mutation with adjectives

Adjectives normally follow the noun and will undergo the soft mutation when the noun is feminine singular.

(11) Masculine:
a. *ci mawr* [ki mawr]
'big dog'
b. *ci du* [ki diː]
'black dog'
c. *ci tenau* [ki tɛnaj]
'thin dog'

(12) Feminine:
a. *cath* fawr [kaθ vawr]
'big cat'
b. *cath* ddu [kaθ ðiː]
'black cat'
c. *cath* denau [kaθ dɛnaj]
'thin cat'

Feminine plurals do not trigger the soft mutation here. Thus we get *cathod mawr*, not *cathod fawr*.

(13) Feminine plural:
a. *cathod mawr* [kaθɔd mawr]
'big cats'
b. *cathod du* [kaθɔd diː]
'black cats'
c. *cathod tenau* [kaθɔd tɛnaj]
'thin cats'

Note that adjectives that begin with [ɬ] or [r̥] do undergo soft mutation after feminine singular nouns:

(14) a. *lloerig* [ɬɔjrɪg]
'crazy'
cath loerig [kaθ lɔjrɪg]
'crazy cat'
b. *llygadog* [ɬəgadɔg]
'sharp-eyed'
cath lygadog [kaθ ləgadɔg]
'sharp-eyed cat'
c. *rhudd* [r̥ɨð]
'red'
cath rudd [kaθ rɨð]
'red cat'
d. *rhesymol* [r̥ɛsəmɔl]
'reasonable'
cath resymol [kaθ rɛsəmɔl]
'reasonable cat'

2.4 Form of 2, 3, and 4

The numbers 2, 3, and 4 take different forms with feminine nouns, as in Table 8. This is true of combinations that end in those digits as well, e.g. *dau ddeg dwy* '22 (fem)' [daj ðeg duj] vs. *dau ddeg dau* '22 (masc)' [daj ðeg daj], *pedwar deg tair* '43 (fem)' [pɛdwar deg tajr] vs. *pedwar deg tri* '43 (masc)' [pɛdwar deg tri].

Table 8: Forms of the numbers 2, 3, 4.

	Masculine		**Feminine**	
2	dau	[daj]	dwy	[duj]
3	tri	[tri]	tair	[tajr]
4	pedwar	[pɛdwar]	pedair	[pɛdajr]

The number 2 triggers soft mutation on a following noun regardless of gender. The number 3 can trigger aspirate mutation on a following masculine noun.[4]

4 With low numbers, the noun is singular; with higher numbers plural. The dividing line is typically above 9, but varies greatly.

(15) a. *dau gi* [daj gi]
‘2 dogs’
b. *dwy gath* [duj gaθ]
‘2 cats’
c. *tri chi* [tri xi]
‘3 dogs’
d. *tair cath* [tajr kaθ]
‘3 cats’
e. *pedwar ci* [pɛdwar ki]
‘4 dogs’
f. *pedair cath* [pɛdajr kaθ]
‘4 cats’

2.5 Form of certain adjectives

A few adjectives have a feminine form, as in Table 9. This can be used attributive position with feminine singular nouns. The feminine form of the adjectives will undergo soft mutation in this environment as well, as already described above, as exemplified in (16).

Table 9: Certain adjectives have feminine forms.

Masculine	Feminine	
gwyn	gwen	‘white’
melyn	melen	‘yellow’
bychan	bechan	‘small’
cryf	cref	‘strong’
trwm	trom	‘heavy’
byr	ber	‘short’
llym	llem	‘strict’
tlws	tlos	‘pretty’

(16) a. *ci gwyn* [ki gwɪn]
‘white dog’
b. *cath* wen [*ka*θ wɛn]
‘white cat’
c. *dyn bychan* [diːn bəxan]
‘small man’

d. *dynes fechan* [dənɛs vɛxan]
'small woman'
e. *afal trwm* [aval trʊm]
'heavy apple'
f. *almon* drom [almɔn drɔm]
'heavy almond'

The masculine form is used in predicative position regardless of gender or number, or attributively with plurals regardless of gender.[5]

(17) a. *Mae'r gath yn wyn.* [majr gaθ ən wɪn]
'The cat is white.'
b. *Mae'r gath yn drwm.* [majr gaθ ən drʊm]
'The cat is heavy.'
c. *cathod gwyn* [kaθɔd gwɪn]
'white cats'
d. *cathod trwm* [kaθɔd trʊm]
'heavy cats'

2.6 Form of demonstratives

Finally, demonstratives agree in gender whether used attributively or independently, as in Table 10. The attributive use is rather formal, but occurs fairly often in writing. In speech, one hears the forms built on *yma* 'here' [əma] and *yna*

Table 10: Gender marking with demonstratives.

	This	that
Masc.	hwn [hʊn]	hwnnw [hʊnʊ]
Fem.	hon [hɔn]	honno [hɔnɔ]
Pl.	hyn [hɨn]	hynny [hənɨ]

5 Some adjectives also have a plural form, e.g. *trymion* in *cathod trymion* 'heavy cats' [kaθɔd trəmjɔn], but these tend to be literary or poetic and are not relevant to gender, so we set them aside.

'there' [əna] more often. These are typically reduced to [ma] and [na], as in (18) and (19).

(18) a. *y ci hwn* [ə ki hʊn]
'this dog'
b. *y gath hon* [ə gaθ hɔn]
'this cat'
c. *y cathod hyn* [ə kaθɔd hɨn]
'these cats'
d. *y ci 'ma* [ə ki ma]
'this dog'
e. *y gath 'ma* [ə gaθ ma]
'this cat'
f. *y cathod 'ma* [ə kaθɔd ma]
'these cats'

(19) a. *y ci hwnnw* [ə ki hʊnʊ]
'that dog'
b. *y gath honno* [ə gaθ hɔnɔ]
'that cat'
c. *y cathod hynny* [ə kaθɔd hənɨ]
'those cats'
d. *y ci 'na* [ə ki na]
'that dog'
e. *y gath 'na* [ə gaθ na]
'that cat'
f. *y cathod 'na* [ə kaθɔd na]
'those cats'

The demonstratives agree in gender when they are used independently, if the gender of the referent is known.

(20) a. *Dyma gi.* [dəma gi]
'Here is a dog.'
b. *Wyt ti eisiau hwn?* [ujt ti iʃɔ hʊn]
'Do you want this?'
c. *Dyma gath.* [dəma gaθ]
'Here is a cat.'
d. *Wyt ti eisiau hon?* [ujt ti iʃɔ hɔn]
'Do you want this?'

The plural form is used when the gender is unknown:

(21) a. *Beth ydy hynny?* [bɛθ ədɨ hənɨ]
'What is that?'
b. *Mae hyn yn ddiddorol.* [maj hɨn ən ðiðɔrɔl]
'This is interesting.'

2.7 Interim summary

Summarizing to this point, Welsh makes a gender distinction among nouns. Aside from cases where grammatical gender lines up with natural gender, the gender of a noun is not apparent from its form. Rather, one only sees gender through indirect mechanisms. These include the form of pronouns, possessive adjectives, and demonstratives, the form of certain numbers and adjectives, and the applicability of the soft mutation to feminine singular nouns that begin with appropriate consonants.

Some of these options are quite rare, as we'll see below. For example, feminine forms of adjectives are only relevant for specific adjectives and are not required. Attributive use of demonstratives is formal and not used in colloquial speech.

We next consider the basic statistical generalizations governing gender and then various models for predicting gender.

3 Basic statistical regularities

We can calculate the basic statistical regularities of Welsh using the tagged CEG corpus (Ellis et al. 2001). The corpus contains 1,223,649 word tokens tagged for part of speech, lemma, and mutation. The corpus is a written one, composed of 500 samples from newspaper articles of various sorts, fiction, nonfiction, and official documents. In this corpus, we have the breakdown of nouns by gender in Table 11. The first column gives the total number of tokens of each type while

Table 11: Distribution of gender categories in the CEG corpus.

	Tokens		Types	
Masc.	120,646	64%	5,302	69%
Fem.	57,178	30%	2,037	27%
Indet.	11,598	6%	303	4%

the second column gives the total number of distinct words in each category. Note that the type and token distributions do not differ a great deal, from which we can conclude that the overall distribution of words of each gender class is not distinct.

In terms of predicting the gender of novel words, this overall distribution provides a benchmark strategy for determining the gender of a noun: *guess masculine*. That is, since words of masculine gender constitute the majority, we'd be right more often than not if we guessed the word was masculine. Specifically, we'd be right 64% of the time in the CEG corpus with respect to noun tokens.

This is not a general fact about gender systems. In other systems, genders are not necessarily distributed so asymmetrically. Table 12 gives a few rough counts for Spanish, French, German, Dutch, and Russian.[6] We set aside the neuter gender which isn't relevant in the case of Welsh. Confining our attention to masculine and feminine, none of these languages exhibits a skew as extreme as Welsh. It's also not the case that the skew is always in the same direction; in some of these languages there are more masculines and in some more feminines.

Table 12: Distribution of gender categories in other languages.

Language	Corpus	Masculine	Feminine	Neuter
Spanish	IULA	69,901	72,088	NA
		49%	51%	NA
French	Lexique380	26,744	18,925	NA
		59%	41%	NA
German	CELEX	10,786	13,688	6,005
		35%	45%	20%
Dutch	CELEX	27,925	24,819	21,795
		37%	33%	29%
Russian	Russ. Nat. Corp.	37,737,473	27,962,098	14,214,372
		47%	35%	18%

4 Morphology

Let's now consider whether there might be more successful strategies for determining the gender of a noun other that simply guessing masculine. One way of

6 This table leaves out various sorts of "common" genders in the different languages.

determining the gender of at least some nouns is morphology. Certain affixes are strongly or uniquely associated with different genders. For example, the suffixes in Table 13 are relatively frequent and show up with masculine words. Table 14 gives some frequent suffixes that show up with feminine nouns.

Table 13: Frequent masculine suffixes.

-deb	absenoldeb	'absence'	[absɛnɔldɛb]
	cytundeb	'agreement'	[kətɨndɛb]
	diddordeb	'interest'	[diðɔrdɛb]
-iant	methiant	'failure'	[mɛθjant]
	moliant	'praise'	[mɔljant]
	peiriant	'engine'	[pejrjant]
-yn	(a)deryn	'bird'	[(a)dɛrɨn]
	bathodyn	'badge'	[baθɔdɨn]
	blodyn	'flower'	[blɔdɨn]
-iad	adolygiad	'review'	[adɔləgjad]
	benthyciad	'borrowing'	[bɛnθəkjad]
	canlyniad	'consequence'	[kanlənjad]
-wr	adarwr	'bird-catcher'	[adarur]
	Albanwr	'Scot'	[albanur]
	arbenigwr	'specialist'	[arbɛnigur]
-ydd	ieithydd	'linguist'	[jejθɨð]
	anarchydd	'anarchist'	[anarxɨð]
	darllenydd	'reader'	[darɬɛnɨð]
-wch	anialwch	'desert'	[anjalux]
	ariangarwch	'avarice'	[arjangarux]
	harddwch	'beauty'	[harðux]
-ter/der	balchder	'pride'	[balxdɛr]
	anhoffter	'dislike'	[anhoftɛr]
	dyfnder	'depth'	[dəvndɛr]
-rwydd	anghofrwydd	'forgetfulness'	[aŋhɔvrujð]
	cwrteisrwydd	'courtesy'	[kʊrtejsrujð]
	hapusrwydd	'happiness'	[hapɨsrujð]

Using just these affixes, we can correctly assign gender to 37,165 word tokens (20%) and incorrectly to 8,997 word tokens (5%) in the CEG corpus. We can also combine this with our *guess masculine* strategy. Specifically, if we can identify a suffix, our guess for gender is based on that; if no affix can be identified, we guess masculine. Using this combined strategy, we get 126,694 word tokens correct (67%) and 62,728 word tokens incorrect (33%). As we saw above,

Table 14: Frequent feminine suffixes.

-aeth	absenoliaeth	'absence'	[absɛnɔljajθ]
	llofruddiaeth	'murder'	[ɬɔvrɨðjajθ]
	cystadleuaeth	'competition'	[kəstadlejajθ]
-en	afallen	'apple tree'	[avaɬɛn]
	cangen	'branch'	[caŋɛn]
	deilen	'leaf'	[dejlɛn]
-wraig	tafarnwraig	'bar maid'	[tavarn^{w}rajg]
	cantwraig	'singer'	[kant^{w}rajg]
	golchwraig	'washerwoman'	[gɔlx^{w}rajg]
-es	arthes	'she-bear'	[arθɛs]
	awdures	'authoress'	[awdɨrɛs]
	Eifftes	'Egyptian (female)'	[ejftɛs]
-fa	allanfa	'exit'	[ałanva]
	cuddfa	'hiding place'	[kɨðva]
	meddygfa	'surgery'	[mɛðəgva]

guessing masculine by itself would give us 64%. The difference is significant, $\chi^2(1,189422) = 684.06$, $p < 0.001$, but the effect size is extremely small: Cramér's $V = 0.05$.

These numbers are based just on the suffixes above, an arbitrary sample. More probative would be a test using all possible suffixes. To do this, we first identified all possible word suffixes that are uniquely associated with one gender or the other in the CEG corpus. For example, *-ldeb* is uniquely associated with masculine words and *-ogaeth* with feminine words. (Even though *-deb* is a morpheme of Welsh associated with the masculine gender, that letter sequence is not uniquely associated with masculine, e.g. *diweddeb* 'cadence' [diwɛ́ðɛb] has indeterminate gender.) Longer potential suffixes were excluded when they already contained a candidate suffix. For example, *-oldeb* is also uniquely associated with masculine words, but contains *-ldeb*, so it is excluded. Finally, for a suffix to be included, it had to be associated with at least two words of the same gender.

This resulted in 1,720 candidate suffixes from the CEG corpus. Counts for each morpheme ranged from 162 to 2. These are distributed in typical Zipfian fashion as shown in Figure 1.

To test these, we have to use nouns not in the CEG corpus and so these were drawn from a publically available electronic dictionary: Nodine (2003). This dictionary contains 24,662 entries, of which 13,894 are nouns. Of these, 1,680 do not appear in the CEG corpus. These latter nouns are distributed across the gender categories as in Table 15.

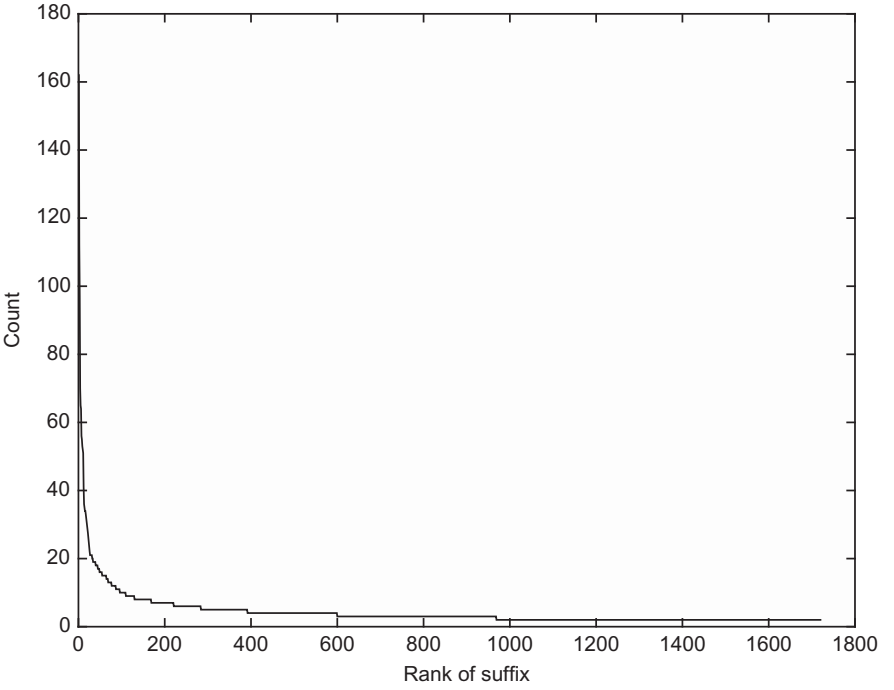

Figure 1: Counts for hypothesized morphemes in CEG corpus.

Table 15: Distribution of gender categories in the Nodine dictionary.

Gender	Count	Percent
Masculine	1,202	72
Feminine	433	26
Indeterminate	45	3

Hypothesized suffixes were tested incrementally, applying more common ones before less common ones. Figure 2 shows the results proceeding from the hypothesized suffixes with the highest counts to those with the lowest. The overall percent correct is given along the left axis and it can be seen that success increases as more and more suffixes are invoked. The curve is fairly steep at the beginning and then levels out, never getting above 50%.

We can combine the hypothesized suffixes strategy with the guess masculine strategy: if a form ends in a hypothesized suffix, guess the appropriate gender; if not, guess masculine. When we do this, the results level out, as shown in Figure 3. This is expected. When only a few suffixes are applied, we use the

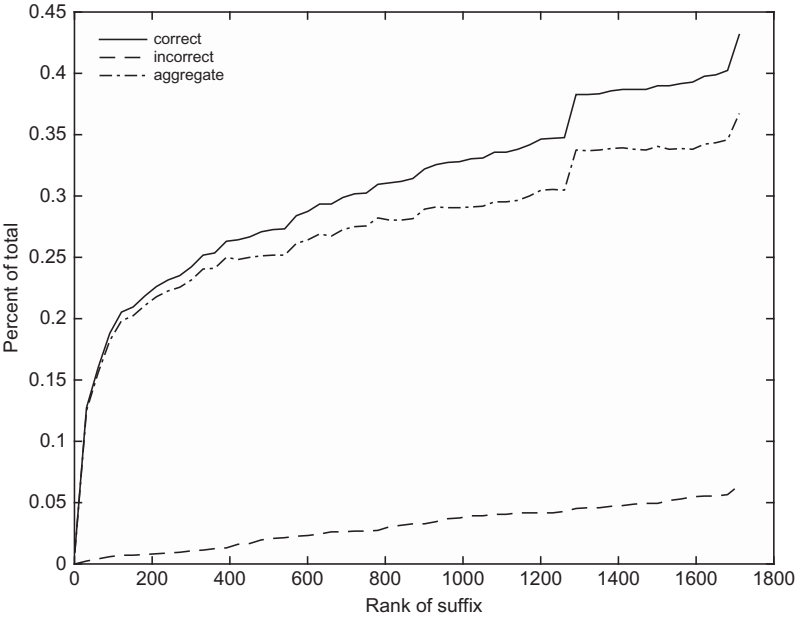

Figure 2: Identifying gender in the Nodine dictionary using hypothesized suffixes.

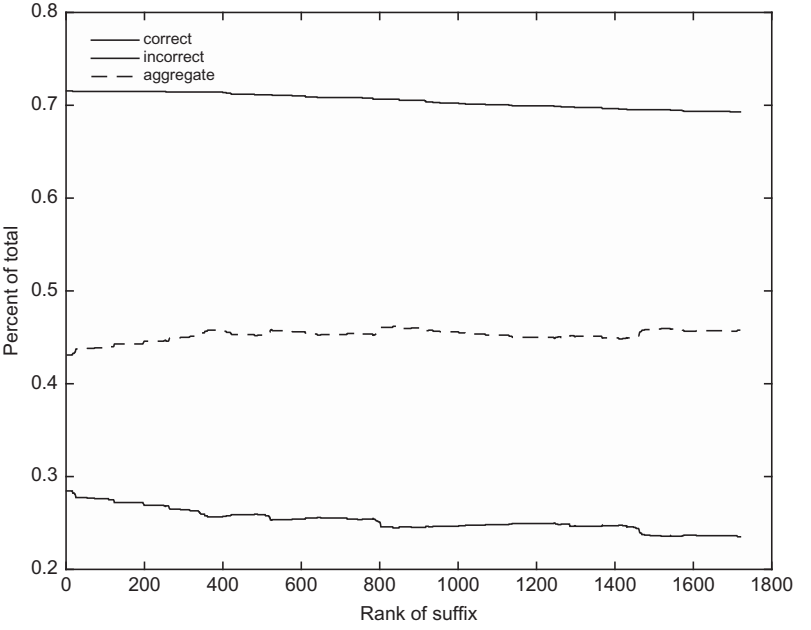

Figure 3: Identifying gender in the Nodine dictionary using hypothesized suffixes and the *guess masculine* strategy.

guess masculine strategy almost exclusively; thus this combined algorithm starts at effectively 72%. As more and more suffixes are applied, success falls off to approximately 69%. The difference is significant, $\chi^2(1,1680) = 4.23$, $p = 0.04$, but the effect size is extremely small: Cramér's $V = 0.03$. We conclude that this does no better than simply guessing masculine.

5 Letter-based *N*-gram models

In this section, we consider an *N*-gram approach.[7] The basic idea is that the general shape of masculine and feminine nouns might differ in terms of what consonants and vowels occur and what order they might occur in. Since Welsh orthography is relatively transparent, we can use orthographic representations as a stand-in for phonological or phonetic ones.

It's possible that the raw frequency of individual sounds (qua letters) might suffice, but we begin with the hypothesis that sound sequences matter, that if the phonology of words has some connection to their gender it won't be just in terms of what sounds occur in a word, but in terms of what sequences of sounds occur. The simplest way to implement this idea is with a letter-based *N*-gram model.

To understand this, we first review a little basic probability theory. We take the relative frequency of an item to be a reasonable estimate of its probability. For example, the probability of *deryn* 'bird' can be estimated using a corpus by dividing the number of occurrences of *deryn* by the total number of words in a corpus. The same thing can be done with word sequences; for example, the probability of *deryn bach* 'little bird' can be estimated by dividing the number of occurrences by the total number of two-word sequences in the corpus. We can schematize these as follows:

(22)
$$\mathrm{p}(\mathrm{deryn}) = \frac{C(\mathrm{deryn})}{C(\mathrm{all\ words})}$$

$$\mathrm{p}(\mathrm{deryn\ bach}) = \frac{C(\mathrm{deryn\ bach})}{C(\mathrm{all\ two\text{-}word\ sequences})}$$

An expression like C(deryn) refers to the count for that word. The last expression above is an instance of joint probability; we can also calculate conditional

7 See Manning and Schütze (1999) for a general discussion of such models.

probabilities. For example, we can calculate the probability of *bach*, given that it co-occurs with or follows *deryn*:

$$(23)\quad \mathrm{p}(\mathrm{bach} \mid \mathrm{deryn}) = \frac{\mathrm{p}(\mathrm{deryn\,bach})}{\mathrm{p}(\mathrm{deryn})}$$

All of these definitions can be relativized to other elements, including individual letters or sequences of letters. For example, we can calculate p(*e*), p(*de*), and p(*e*|*d*) in analogous fashion.

It follows as a matter of algebra (the *chain rule*) that the probability of sequence of elements is equivalent to the product of a set of conditional probabilities:

$$(24)\quad \mathrm{p}(deryn) = p(d) \times \mathrm{p}(e \mid d) \times \mathrm{p}(r \mid de) \times \mathrm{p}(y \mid der) \times \mathrm{p}(n \mid dery)$$

In this case, the probability of each letter is calculated as a function of the probabilities of all preceding material. It is customary to assume that the number of elements that the conditional probabilities are built on is bounded *(limited horizon* property).[8] For example, we might bound it at a single segment:

$$(25)\quad \mathrm{p}(deryn) = \mathrm{p}(d) \times \mathrm{p}(e \mid d) \times \mathrm{p}(r \mid e) \times \mathrm{p}(y \mid r) \times \mathrm{p}(n \mid y)$$

If the context is a single element, in this case a single letter, this is referred to as a *bigram model*. *N*-gram models, including bigram models, are widely used to model the statistical properties of word or letter sequences. A bigram model is the simplest way to include limited sequence information in a probabilistic model of the relationship between word shapes and gender.

If we construct bigram models for different word categories, and those letter sequence probabilities are a good indication of what category a word should belong to, then we should be able to use those models to predict the category of novel words. At first blush, one might think that we simply calculate the probability of a novel word with respect to the different models and choose the category that assigns that word the highest probability.

Implementing this, three separate bigram models were constructed from the CEG corpus, one each for masculine, feminine, and indeterminate nouns. These models were tested against the novel words from the Nodine dictionary. For each novel word, we calculate its probability using the bigrams calculated for

8 This is not a mathematical truth, but a convenient and usual simplification which allows us to continue to use probability theory and to implement such models efficiently (Manning and Schütze 1999).

each gender. The gender of the noun is predicted to be the one which assigns it the highest probability. Out of 1,633 words, the bigram model got 1,077 correct (66%). This is, of course, less than the *guess masculine* strategy would achieve for the Nodine dictionary. The difference is significant, $\chi^2(1, 1633) = 25.14$, $p < 0.001$, but the effect size is small: Cramér's $V = 0.08$.

This simple model is actually mathematically incorrect. Each bigram model is actually a conditional probability, a calculation of the probability of some letter sequence given each of the three gender categories. For some word *w* of unknown gender, the values calculated from each model would let us compare p(*w*|*M*), p(*w*|*F*), and p(*w*|*I*), where *M*, *F*, and *I* correspond to the three gender categories. In point of fact, what we're actually interested in are p(*M*|*w*), p(*F*|*w*), and p(*I*|*w*). That is, our probability models give us the first three quantities, but we really want the second three; we know the likelihood of each letter sequence given each gender, but now we want to know the likelihood of each gender given the letter sequences we've calculated.

We can get these latter values with a simple algebraic manipulation: *Bayes' Law*[9]. We know the following from the definition of conditional probability:

(26) $$\mathrm{p}(w \mid M) = \frac{\mathrm{p}(w, M)}{\mathrm{p}(M)}$$
$$\mathrm{p}(M \mid w) = \frac{\mathrm{p}(w, M)}{\mathrm{p}(w)}$$

These can be converted to the following:

(27) $$\mathrm{p}(w \mid M)p(M) = p(w, M)$$
$$p(M \mid w)p(w) = p(w, M)$$

from which it follows that:

(28) $$\mathrm{p}(M \mid w)p(w) = \mathrm{p}(w \mid M)p(M)$$

This, in turn, means that if we want to calculate p(*M*| *w*), we need only calculate the following.

(29) $$\mathrm{p}(M \mid w) = \frac{\mathrm{p}(w \mid M)p(M)}{\mathrm{p}(w)}$$

9 Manning and Schütze (1999).

We can calculate the same values for the other gender categories.

(30) $$\mathrm{p}(M \mid w) = \frac{\mathrm{p}(w \mid M)p(M)}{\mathrm{p}(w)}$$

$$\mathrm{p}(F \mid w) = \frac{p(w \mid F)p(F)}{\mathrm{p}(w)}$$

$$\mathrm{p}(I \mid w) = \frac{\mathrm{p}(w \mid I)p(I)}{\mathrm{p}(w)}$$

To determine the best fit for gender, we compare these values and choose the highest. On this procedure, since the denominators on the right are all the same, they can be eliminated.

(31) $$\mathrm{p}(M \mid w) = p(w \mid M)p(M)$$

$$\mathrm{p}(F \mid w) = p(w \mid F)p(F)$$

$$\mathrm{p}(I \mid w) = p(w \mid I)p(I)$$

In a more compressed form, we are trying to solve for some gender g:

(32) $$\hat{g} = \arg\max_g \mathrm{p}(w \mid g)p(g)$$

The initial model above didn't take into account the relative distribution expected for each gender: p(M), p(F), and p(I). In other words, based on the CEG corpus, we expect masculines to make up 69% of the word types, feminines 27%, and indeterminate items 4%, as in Table 11. We can weight each model's estimate accordingly, multiplying the masculine estimate by 0.69, the feminine by 0.27, and the indeterminate one by 0.06. Using this, we get 1,215 out of 1,633 correct (74%). This is better than what we've gotten with the previous models, but still only marginally better than the *guess masculine* strategy: 72%. The difference is significant, $\chi^2(1, 1633) = 6.53$, $p = 0.01$, but the effect size is extremely small: Cramér's $V = 0.04$.

Cavnar and Trenkle (1994) offer a related approach that we can try as well. Their procedure is widely used for language identification and it might reasonably generalize to the gender problem here. We again construct letter N-grams. In this case, however, for each model we compute counts for all N-gram types up to 5-grams. We then rank-order all the N-grams of different orders *together* and select the top 300 for each model. That rank-ordered list constitutes the model for each gender. For each candidate word, we extract and order the N-grams of that word. We compute an out-of-rank score for the word, the sum of the

differences in ranking of the *N*-grams of the word with respect to the 300 *N*-grams of the model, a measure of how far the ranking of its *N*-grams depart from the language model, and choose the model with the lowest out-of-rank score.

This model performed relatively well, getting 1,142 words correct out of 1,633 (70%). When we weight the models, as we did in the previous case, performance improves to 1,188 correct out of 1,633 (73%). This is marginally better than the guess masculine model (72%), $\chi^2(1, 1633) = 3.46$, $p = 0.06$, Cramér's $V = 0.03$, and marginally worse than the simple weighted bigram model (74%), $\chi^2(1, 1633) = 14.04$, $p < 0.001$, Cramér's $V = 0.05$. The first difference is not significant; the second is, but the effect size is very small.

Summarizing, the *N*-gram approaches perform better than the morphological approaches, but do not significantly outperform the *guess masculine* strategy. If we cast this in psychological terms, this implies that the simplest gender learning model, one that relies on no prior knowledge of Welsh and with the simplest of phonological assumptions, does not suffice.

We next turn to approaches that rely on specific properties of Welsh.

6 Word-level properties of Welsh

In this section, we consider language-specific techniques for identifying gender. Effectively, we build our gender learning algorithms on more sophisticated phonological and morphological assumptions. We focus on two: vowels and consonants.

Consider vowels first. The logic for considering vowels as a way of distinguishing genders is based on gender-marking with adjectives discussed in Section 2.5. If we consider the adjectives that mark gender, the overwhelming majority of them mark it in one of two ways. Either the masculine form contains *w* [u/ʊ] and the feminine *o* [o/ɔ], or the masculine contains *y* [ɨ/ə] and the feminine *e* [e/ɛ]. Here we consider the hypothesis that the same distinction may statistically separate gender in nouns. Examining the distribution of these in noun types in the CEG corpus, we find that there is indeed a difference in their distribution, as shown in Table 16. The vowels *y* and *w* are far more

Table 16: Distribution of vowels in CEG corpus.

	y, w	o,e	Ratio	Types
Masc.	7,318	6,121	1.196	7,946
Fem.	1,724	3,448	0.500	3,121
Indet.	278	439	0.633	498

common with masculines than with the other two categories. The overall distribution of the two vowel categories across the gender categories is significant: $\chi^2(2, 19391) = 708.99$, $p < 0.001$, Cramér's $V = 0.19$. The full distributions for *y*, *w* are plotted in Figure 4. The *x*-axis is the number of instances of these vowels in a word and the *y*-axis is the percent of words that have these numbers.

We can use these facts to build yet another model for predicting gender (a model that will unfortunately be unsatisfactory as well). The data in Figure 4 effectively represent p (*yw*|*g*), where *g* represents the three gender categories.[10]

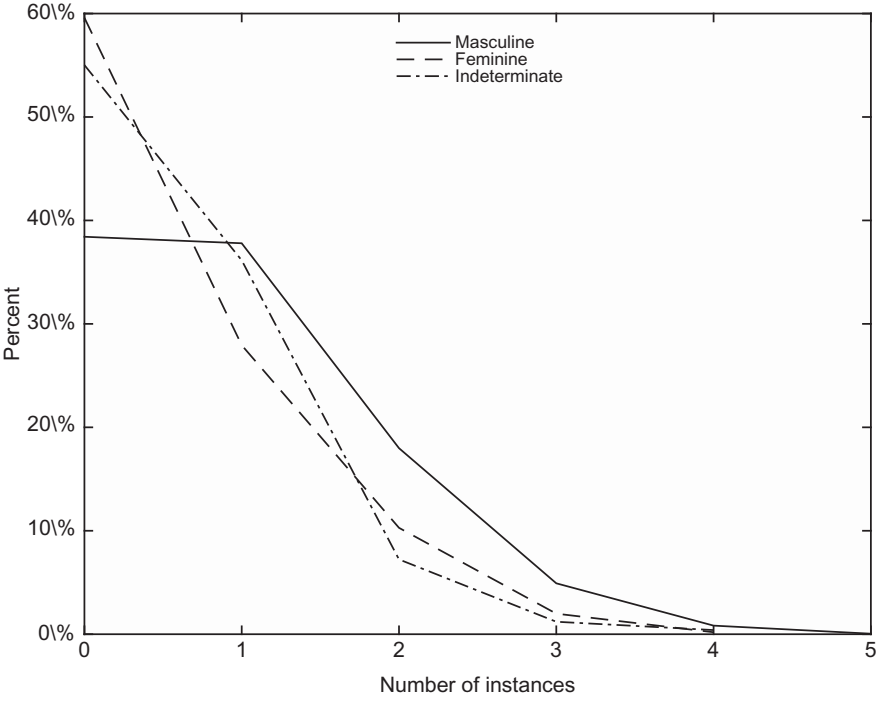

Figure 4: Distribution of *y* and *w* across genders in CEG.

10 To be fully precise, we should write $p(yw|g)$ as $p(y \vee w|g)$, but we use the former to make subsequent formulas simpler.

What we need is a model of p (*g*|*yw*), that is, a model that predicts gender from the distribution of *y* and *w*. We can do this just as we did in the previous section using Bayes' Law. We have the following equivalences:

(33) $$p(M \mid yw) = \frac{\mathrm{p}(yw \mid M)\,\mathrm{p}(M)}{\mathrm{p}(yw)}$$

$$\mathrm{p}(F \mid yw) = \frac{\mathrm{p}(yw \mid F)\,\mathrm{p}(F)}{\mathrm{p}(yw)}$$

$$p(I \mid yw) = \frac{\mathrm{p}(yw \mid I)\,\mathrm{p}(I)}{\mathrm{p}(yw)}$$

Since the denominators on the right are the same, they can be simplified to:

(34) $$p(M \mid yw) = \mathrm{p}(yw \mid M)\,\mathrm{p}(M)$$

$$\mathrm{p}(F \mid yw) = \mathrm{p}(yw \mid F)\,\mathrm{p}(F)$$

$$p(I \mid yw) = \mathrm{p}(yw \mid I)\,\mathrm{p}(I)$$

In other words:

(35) $$\hat{g} = \arg\max_g \mathrm{p}(yw \mid g)\mathrm{p}(g)$$

The basic idea is that to determine the gender of a novel word, we first compute the likelihood of its count for *y*, *w* given the three distributions in Figure 4. We know the three gender categories are not equally likely, so we weight each probability by the likelihood of the category. We then choose the category that gives the item the highest probability.

This model is mathematically correct, but the weighting swamps the effect of *y* and *w*. The simple gender probabilities are as given below:

(36) Masc. $$\mathrm{p}(M) = \frac{7946}{7946 + 3121 + 498} = 0.69$$

Fem. $$\mathrm{p}(F) = \frac{3121}{7946 + 3121 + 498} = 0.27$$

Indet. $$\mathrm{p}(I) = \frac{498}{7946 + 3121 + 498} = 0.04$$

We then get the following calculations. Unfortunately, masculine (37) wins in all cases; notice how the masculine values for all counts is highest, meaning that this model is effectively the same as *guess masculine*.

(37) masculine =

n	$p(n \mid M)$	×	$p(M)$	=	$p(M \mid n)$
0	0.38		0.69		0.26
1	0.37		0.69		0.25
2	0.18		0.69		0.12
3	0.04		0.69		0.03
4	0.008		0.69		0.0055
5	0.001		0.69		0.0007

(38) feminine =

n	$p(n \mid F)$	×	$p(F)$	=	$p(F \mid n)$
0	0.59		0.27		0.16
1	0.27		0.27		0.07
2	0.1		0.27		0.03
3	0.02		0.27		0.0054
4	0.002		0.27		0.0005
5	0		0.27		0

(39) indeterminate =

n	$p(n \mid I)$	×	$p(I)$	=	$p(I \mid n)$
0	0.55		0.04	=	0.02
1	0.36		0.04	=	0.01
2	0.07		0.04	=	0.003
3	0.01		0.04	=	0.0004
4	0.004		0.04	=	0.0002
5	0		0.04	=	0

We can incorporate both pairs of vowels into the model as well, but the performance is nearly the same. All combinations of the vowels are assigned masculine gender except one: zero instances of *y* and *w*, and three instances of *o* or *e*. Following the same procedure as above over the distribution of both vowels, we get 1,204 correct and 476 incorrect for 72%, effectively the same as *guess masculine*.

The important take-home message here is that even though there is a significant difference between the gender categories in terms of the distribution of these vowels, that difference is insufficient to have any predictive force when compared with the *guess masculine* strategy. This, in fact, follows mathematically from the fact that we must factor in the relative likelihood of the different gender categories overall when building a predictive model.

Let's now turn to a model built on consonants. Another possibility would be to take advantage of the distribution of soft mutation to determine the gender of a noun. We saw in Section 2.2 that feminine singular nouns undergo soft mutation after the definite article, while masculines do not. What we might hypothesize is that, overall, feminine nouns should exhibit more instances of soft mutation than masculine nouns. This is, in fact, the case. Table 17 shows the number of instances of soft mutation across genders in the CEG corpus. This

Table 17: Soft mutation by gender in the CEG corpus.

Gender	Soft mut.	Tokens	Percent
Masc.	27,552	120,646	23
Fem.	16,829	57,178	29
Indet.	3,083	11,598	27

distribution is significant: $\chi^2(2, 189422) = 913.96, p < 0.001$, Cramér's $V = 0.07$, but the effect size is very small.

To implement this as a predictive model for gender, we need to generalize beyond a binary choice for whether a form is mutated or not. The problem is that the difference is too small at that level of granularity: all genders occur more often in unmutated form. If, however, we consider the relative frequency of mutation, if we consider a candidate form in terms of how often it exhibits soft mutation, then we may have something.

Figure 5 shows the distribution of soft mutation across the three gender categories. Each line represents how often words occur without soft mutation.

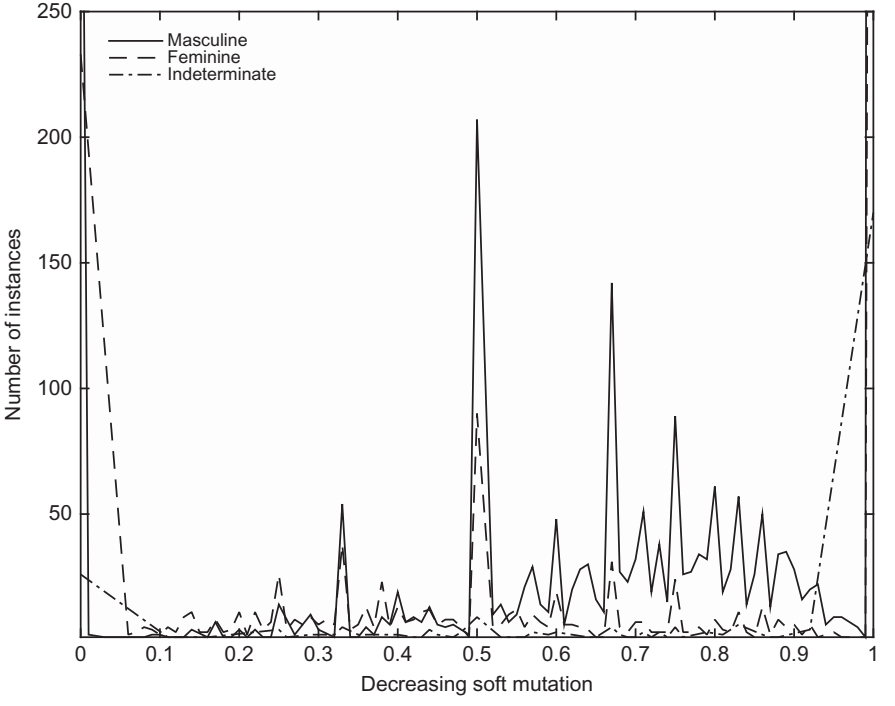

Figure 5: Instances of soft mutation by gender categories.

A form on the right side of the chart rarely undergoes soft mutation; a form on the left often. You can see that masculine forms, while generally most frequent across the continuum, dip below feminine forms when soft mutation is more frequent.

It's a trivial matter to convert this into a model for determining gender. For any candidate noun, we simply determine how often it shows soft mutation and then read the gender off the highest line at that point in Figure 5. If we extract 100 random words, distributed appropriately across the different gender categories, and we do this 100 times, the average success rate is 70.91, again, not very different from simply guessing masculine.

Another way to attack this question is to examine the distribution of initial consonants directly. Hammond (2011) shows that there are significant differences in the distribution of initial consonants in Welsh with respect to gender categories. Figure 6 shows the basic distributional data in terms of percentages; Figure 7 shows it in terms of raw counts.

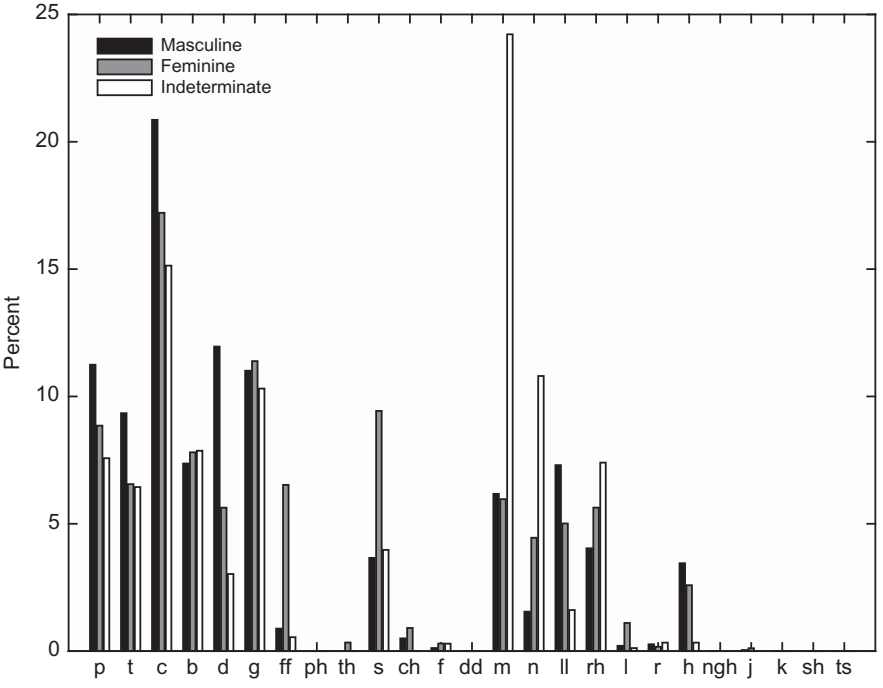

Figure 6: Initial consonant percentages by gender categories.

From Figure 6, we see the basic generalizations that distinguish the gender categories. We see that masculine nouns generally have more initial voiceless

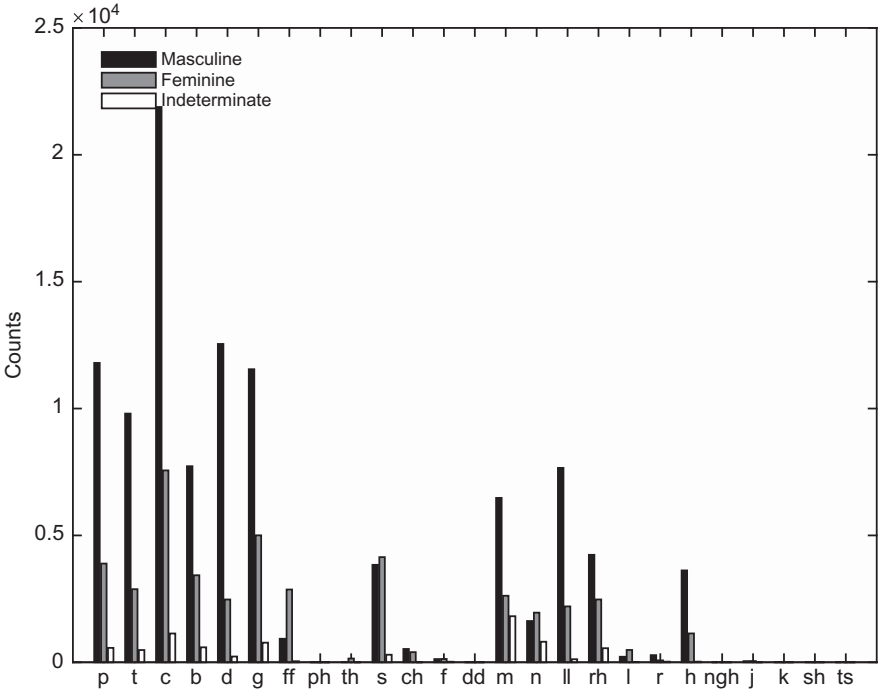

Figure 7: Initial consonant counts by gender categories.

stops and [b], while feminines have a high number of instances of the unmutatable *ff* [f] and *s* [s]. Interestingly, we find a high number of initial nasals *m* [m] and *n* [n] for the indeterminate gender category. From Figure 7 where raw counts are plotted, we see the usual swamping effect from masculines. Masculines outnumber both of the other categories for all consonants except *ff* [f], *s* [s], *n* [n], and *l* [l] (all unmutatable).

If we build a prediction model based on initial consonants, we get only 69% success, effectively the same as *guess masculine*.

In summary, word-internal language-specific properties inspired by the grammatical and lexical system of Welsh fare no better than the *guess masculine* strategy.

7 Syntax

Let's now consider higher-order models, models where we attempt to predict the gender of a noun by looking at the syntactic environment of the noun. We expect this general approach to be rather successful from our description of

the Welsh gender system in Section 2 since we will attempt to predict gender from the language-specific properties that we know indicate gender in the language.

We'll consider several different properties based on the facts there: definite articles, gender specific numbers, demonstratives, adjectives, and plurals.

First, we might expect there to be a difference in the distribution of the definite article based on the fact that it triggers soft mutation with a following feminine noun. This is based on the result in Hammond (2014) that in contexts where a word can mutate, there is an over-representation of words that can show the mutation. This is in fact the case, as seen in Table 18: feminine nouns are more likely to be immediately preceded by a definite article than the other two categories. This distribution is significant: $\chi^2(2, 189422) = 709.88$, $p < 0.001$, Cramér's $V = 0.06$, but the effect size is very small. Alone, this difference is far too small to overpower the masculine bias.

Table 18: Distribution of the definite article by gender.

Gender	Def. art. Freq	Def. art. Count	Overall count
Masc.	0.251	30,272	120,646
Fem.	0.310	17,751	57,178
Indet.	0.256	2,969	11,598

Table 19: Distribution of 2, 3, 4 by gender.

Gender	Num. freq.	Num. count	Overall count
Masc.	0.008	941	120,646
Fem.	0.014	798	57,178
Indet.	0.026	297	11,598

Another marker of gender that we discussed are the numbers that are differentiated by gender: *dau*/*dwy* 'two', *tri*/*tair* 'three', and *pedwar*/*pedair* 'four'. Here again, in Table 19 we see a significant difference with gendered numbers more likely with feminines, but even more likely with the indeterminate category.

Though the difference is small, it is also significant: $\chi^2(2, 189422) = 394.84$, $p < 0.001$, Cramér's $V = 0.05$, but the effect size is very small. This difference is also far too small to overpower the masculine bias.

Demonstratives show a small difference, occurring with greater frequency with the indeterminate gender category and feminines, as seen in Table 20.

Table 20: Distribution of demonstratives by gender.

Gender	Dem. freq	Dem. Count	Overall count
Masc.	0.01340	1,617	120,646
Fem.	0.01957	1,119	57,178
Indet.	0.02181	253	11,598

Though this difference is small, it is also significant: $\chi^2(2, 189422) = 123.98$, $p < 0.001$, Cramér's $V = 0.03$, with a very small effect size.

A following adjective is much more likely with feminines than masculines than indeterminate nouns, as seen in Table 21. And though the difference is small, it is also significant, $\chi^2(2, 189422) = 217.17$, $p < 0.001$, Cramér's $V = 0.03$, again with a very small effect size.

Table 21: Distribution of following adjectives by gender.

Gender	Adj. freq.	Adj. count	Overall count
Masc.	0.122	14,770	120,646
Fem.	0.144	8,206	57,178
Indet.	0.104	1,206	11,598

Table 22: Distribution of plural types by gender.

Gender.	Pl. freq.	Pl. count	Overall count
Masc.	0.312	1,608	5,150
Fem.	0.411	788	1,918
Indet.	0.476	186	391

Finally, there is a significant difference in the likelihood of a plural. In Table 22, we compare at the type level, disregarding the frequency of the individual nouns. This is also significant: $\chi^2(2, 7439) = 90.64$, $p < 0.001$, Cramér's $V = 0.11$, with a small effect size.

Summarizing to here, we see significant distributional differences in the various syntactic markers, however, the differences are fairly small and each one alone cannot overcome the masculine bias. In addition, while some of the regularities make sense, it's difficult to come up with a coherent story for why different gender categories stand out for the particular measures they do.

Let's now consider the syntactic cues more directly, not in terms of how frequent this or that marker is, but in terms of how frequent overt unequivocal marking is. For example, with respect to numbers, we ask how often the specific and appropriate gendered number is present to identify the gender of a noun.

Let's consider first the case of definite articles. We've already cited the relative distribution of the article across gender categories in Table 18. How often does each category occur with soft mutation in that environment? This latter figure has to be contextualized some. Some nouns do not begin with consonants that can exhibit the soft mutation. For example, the word *iaith* [jajθ] 'language' is feminine, but the initial glide does not change in soft mutation context, i.e. *yr iaith* [ər jajθ] 'the language'. Table 23 therefore reports the percentage of nouns in each category that begin with consonants that exhibit soft mutation forms when there is a definite article immediately preceding.[11] Raw counts are given in Table 24. As can be seen from the chart in Table 23, the fraction of forms that can undergo soft mutation and do goes up for indeterminate gender items and up further still for feminines. In addition, we can calculate how informative this is. What is the percent of masculine nouns that occur

Table 23: Relative distribution of mutable nouns with the article by gender.

	Masc.	Fem.	Indet.
With article	0.251	0.310	0.256
Mutated with art.	0.009	0.470	0.407
Mutatable with art.	0.746	0.485	0.592
Informativity	0.185	0.146	NA

Table 24: Absolute distribution of mutable nouns with the article by gender.

	Masc.	Fem.	Indet.
With article	30,272	2,969	17,751
Mutated with art.	269	8,340	1,208
Mutatable with art.	22,581	8,609	1,757

11 In principle, it's possible for the definite article to occur further to the left if, for example, a prenominal adjective or number intervenes, e.g. *yr hen gi* [ər heːn gi] 'the old dog', *y tri chi* [ə tri xi] 'the three dogs'. In such cases, the mutation possibilities for the noun depend on the immediate context, as seen here, rather than on the presence of the determiner. Cases like these are then not included in the counts.

with a definite article, have a mutatable initial consonant, and do *not* undergo mutation? Likewise, what is the percent of feminine nouns that occur with a definite article and *do* undergo mutation?

For obvious reasons, we do not calculate an informativity value for the indeterminate gender items. In addition, we do not include indeterminate items in the following charts.

The form of certain adjectives is also a marker for gender, e.g. *ci gwyn* [ki gwɪn] 'white dog' vs. *cath wen* [kaθ wɛn] 'white cat'.[12] This only applies to a very few adjectives and these are not used reliably. Table 25 shows the absolute and relative counts for how often the relevant forms of adjectives occur with masculines and feminines. Certainly the numerical bias corresponds to the descriptive generalizations, but the overall fraction of nouns is quite small.

Table 25: Distribution of gender-marking adjectives by gender.

Noun.	Adjective	Count	Frequency
Masc.	Masc.	598	0.00496
Masc.	Fem.	23	0.00019
Fem.	Masc.	48	0.00084
Fem.	Fem.	313	0.00547

The numbers 2, 3, and 4 also distinguish gender, e.g. *pedwar ci* [pɛdwar ki] 'four dogs' vs. *pedair cath* [pɛdajr kaθ] 'four cats'. These are also relatively rare as shown in Table 26. As with gender-marked adjectives, the numerical bias corresponds to the descriptive generalizations, but the overall fraction of nouns is small.

Table 26: Distribution of 2, 3, 4 by gender.

Noun.	Number	Count	Frequency
Masc.	Masc.	911	0.00755
Masc.	Fem.	30	0.00025
Fem.	Masc.	29	0.00051
Fem.	Fem.	769	0.01345

12 The feminine adjective *gwen* [gwɛn] undergoes soft mutation in this environment as well.

Table 27: Distribution of demonstratives by gender.

Noun.	Dem.	Count	Frequency
Masc.	Masc.	1,374	0.01139
Masc.	Fem.	32	0.00027
Fem.	Masc.	38	0.00066
Fem.	Fem.	965	0.01688

Demonstratives also mark gender, e.g. *y ci hwn* [ə ki hun] 'this dog' vs. *y gath hon* [ə gaθ hɔn] 'this cat'. As noted above, this is a relatively formal construction.[13] Table 27 gives the counts for when the demonstrative immediately follows the noun. As above, the numerical bias corresponds to the descriptive generalizations, but the overall fraction of nouns is small.

Table 28: Distribution of mutable adjectives by gender.

Noun.	Mut'able adj.	Count	Frequency
Masc.	Unmutated	8,230	0.06822
Masc.	Mutated	373	0.00309
Fem.	Unmutated	456	0.00798
Fem.	Mutated	4,290	0.07503

Finally, we consider whether there is an adjective with soft mutation, e.g. *ci mawr* [ki mawr] 'big dog' vs. *cath fawr* [caθ vawr] 'big cat'. See Table 28. Again, the counts here for masculines in particular have to be relativized to just those cases where the adjective can undergo mutation. For example, the adjective *hapus* [hapɨs] 'happy' tells us nothing as its initial consonant does not undergo soft mutation. On the other hand, the adjective *diddorol* [diðɔrɔl] 'interesting' can tell us about gender, since its initial consonant can undergo soft mutation. Here again, there is a distinction with over 6% of masculines showing up with an unmutated adjective that could in principle undergo mutation, and 7% of feminines showing up with a mutated adjective.

Summarizing, we see that in actual use, many of the gender-marking mechanisms of Welsh contribute little to the actual identification of gender.

13 It's also possible for material to intervene between the demonstrative and the noun, e.g. *y ci bach hwn* [ə ki bax hun] 'this little dog (puppy)'. As the distance between the noun and the demonstrative increases, there is an increased need to attend to the syntactic grouping of words. To avoid this complication, we only consider the case where the demonstrative immediately follows the noun.

Only the mutation after the article and the mutation of adjectives seem to contribute at a substantive level.

However, this does not take the whole panoply of gender marking into account. That is, do the cues for gender distribute themselves independently or do they overlap? Intuitively, gendered numbers might only occur when gendered adjectives occur, in which case, we don't gain from their distribution. Alternatively, they could not overlap at all, in which case, we learn as much as possible from each.

We calculated this for the cues we've discussed for the CEG corpus and find that the gender of 37% of word tokens can be identified. If we go further and consider word types – treating the gender of a word type as identified if the gender of one of its tokens has been – then the gender of 66% of word types can be identified. Finally, if we use the *guess masculine* strategy on remaining items, then we can successfully identify the gender of 91% of word types. This is far better than our previous models. The difference is significant, $\chi^2(1, 4543) = 1762.5$, $p < 0.001$, Cramér's $V = 0.32$, and the effect size is large.

Summarizing, in this section we reviewed the language-specific cues for gender. We've seen that some of them occur with such infrequency as to be essentially useless. Nonetheless, when they are combined across word types, then we achieve a score compatible to the best scores we've achieved in previous sections. When we combine the language-specific cues with respect to word types, using the *guess masculine* strategy on the remaining items, the score is the highest yet.

Two points should be emphasized here. First, the success of language-specific strategies in Welsh relies critically on the distribution of soft mutation. This means that gender can only be learned if aspects of the syntax of the language are already learned. Moreover, if soft mutation is not used as described, then many of these cues become far less useful. To the extent that the mutation system exhibits variability, it follows that the gender system will as well.

Second, we've gotten this level of success only by including a guessing component, essentially by assuming that if this is what speakers are doing, they are – at some level of abstraction – guessing. This guessing can, of course, be reduced if the corpus is larger which, in the case of a real learner, is certainly the case. That said, guessing plays an integral role and this leads to two consequences. First, if there is guessing, there will be misguessing. We would then expect a set of words to exhibit ambiguous gender. Second, the *guess masculine* strategy is successful because there is a numerical asymmetry in the genders. However, if the *guess masculine* strategy is then used, we would expect the numerical asymmetry to increase.

We then predict that less frequent words, words that do not occur frequently in contexts where gender is marked to become masculine over time.

8 Models for the residue

We've seen that using language-specific cues we can identify the gender for about two thirds of the singular nouns in the CEG corpus, when we aggregate those cues over word types. We saw, in addition, that we can identify the gender of over 90% of the word types when we use the *guess masculine* strategy on the remainder. Let's now look more closely at the remainder words and see if some combination of our other strategies might not do better. Recall that we have gone through these possibilities:

Morphology using actual or hypothesized suffixes associated with the genders. Hypothesized suffixes are extracted from training corpus.

Letter-based bigrams We build bigram models for each gender category.

Cavnar and Trenkle (1994) We build *N*-grams up to 5, rank-order then, and then select the top 300 for each gender category. Out-of-rank scores are used to assess which category is best.

Vowels Distribution of the vowels that most often distinguish gender-marked adjectives are calculated for each gender category.

Soft mutation Relative frequency of soft mutation is used to characterize the gender categories.

Initial consonant Relative frequency of different initial consonants is used to characterize gender categories.

We got various different success rates with these. The most successful was the Cavnar and Trenkle approach. In fact, we can see all the others, except for soft mutation, as related. The morphology approach, letter-based bigram approach, vowel approach, and initial consonant approach all distinguish genders based on frequent letters or letter sequences at different points in the word. We've already seen that the Cavnar and Trenkle approach outperforms all of these, so will pursue a version of it here.

The Cavnar and Trenkle approach collects all *N*-grams and chooses the 300 highest-ranked for each language model. This does not take into account the fact that in our task we know in advance that there are only two models at stake. Hence, we can choose the highest-ranked *N*-grams that are most effective in

distinguishing the two genders. We do this by taking the count for each *N*-gram for each gender and subtracting the count for the same *N*-gram for the other gender. We then choose the 300 highest-ranking *N*-grams for each gender.

This method requires that we set aside some portion of the initial training corpus so that the nouns we use do not also figure in the training. Using the CEG corpus, we reserve the first 1,000,000 words for training and the remaining 223,649 words for testing. For actual testing, we then use only those nouns in the testing portion that do not also occur in the training portion.

A necessary consequence of this is that test items will be relatively low frequency and will each be relatively infrequent. Therefore, the language-particular cues will be less successful overall than they were when we used the entire corpus. There are 494 novel noun types in the test corpus and language-specific cues allow us to determine the gender of 185 of the types (37%). For the remaining nouns, we use our adapted Cavnar and Trenkle algorithm, which correctly identifies 237. Putting it all together, we get 85% success. If we had just used the *guess masculine* strategy on the remainder, we would have gotten only slightly less: 84%. Calculating this in terms of tokens, we get 89% (644 out of 723 tokens).

9 Conclusion

Language-specific cues largely based on the distribution of soft mutation are the best indication of gender in Welsh. When we put them all together, we see a large and significant effect. Other potential cues, including *N*-gram models, morphology, and feminine forms of numbers and adjectives all are negligible when compared with a simple-minded *guess masculine* strategy. To the extent that we see statistical effects for these other cues, they are typically quite small.

We can draw several conclusions. If our models are a reflection of how gender is learned and used by people, then the soft mutation is essential to the survival of the Welsh gender system. To the extent that that system vacillates or weakens, the cues for gender weaken and we predict that nouns will move to the masculine category. If that continues, then at some point the feminine category will no longer be a general category of nouns, but an exceptional suppletive class of nouns.

We can also conclude that nouns should move to the masculine category when there is insufficient evidence to put them in the feminine category: we expect nouns to move into the masculine category in the absence of direct evidence for feminine gender. This means that less frequent words should become masculine. In fact, this is the case, the average frequency of the gender categories differ, as seen in Table 29. Frequencies are not typically distributed

Table 29: Average frequency of nouns by gender.

Gender	Frequency
Masc.	22.755
Fem.	28.070
Indet.	38.277

normally, so log values were used. (Homogeneity of variance is also not satisfied by these data. Levene's test: $F(2, 7639) = 10.2522$, $p < 0.001$.) An ANOVA on log-transformed values shows a significant effect of gender: $F(2, 7639) = 709.52$, MSE = 1749.93, $p < 0.001$. Since there is a significant effect, we did pairwise *t*-tests to see which comparisons were significant. Masculine is significantly different from feminine, $t(7337) = -8.01$, $p < 0.001$, and from the indeterminate category, $t(5603) = -4.85$, $p < 0.001$, but feminine is not significantly different from the indeterminate category, $t(2338) = -1.22$, $p = 0.224$.

We also expect that there should be indeterminate words, words where for some speakers language-specific cues have sufficed to put the word in the feminine category, while for other speakers, those cues have not sufficed and the *guess masculine* strategy has been employed.

Overall, our results establish that the *guess masculine* strategy quite likely plays a role in the determination of gender categories. We've shown that this follows from the indirect nature of language-specific gender-marking in the language (1c). This in turn predicts that the number of masculines should be increasing (1a) and that there should be a set of words where guessing has led to conflicting or indeterminate gender assignment (1).

This is also supported from acquisition data. Gathercole and Thomas (2001) show that children are generally better in producing gender cues for masculine nouns than for feminine nouns.[14] Gathercole and Thomas's data suggest that nouns tend to be unmutated (generalizing the masculine pattern) and adjectives tend to be mutated (generalizing the feminine pattern). More to the point, however, they show that children acquire grammatical gender in a piecemeal fashion, not generalizing the two gender patterns to nonsense items. This suggests that up until some point in the acquisition process, mutation patterns are treated as exceptions, rather than as a coherent system for distinguishing two classes of nouns.

14 See also Thomas and Gathercole (2005a) and Thomas and Gathercole (2005b).

This is quite consistent with the facts we've reviewed here and the analysis we've developed. The cues for gender are diffuse and indirect, and it is hard to learn and it takes children some time to do so. There is a protracted period where the soft mutation patterns are not seen as marking two distinct noun categories. Rather, mutation is treated as an exceptional property of individual words.

We can also compare our results to those achieved in the computational domain. Cucerzan and Yarowsky (2003) approach gender from a practical computational perspective: how do we determine the gender of words in any language from an unannotated corpus. Their algorithm has three parts. First, some number of 'seed' nouns of both genders are adopted, say 30 of each. There is no analog to this in our algorithm.

Second, the word-level contexts for each group are determined and reliable contexts for each gender category are found. This is basically a statistical process whereby of all occurring contexts, we find those that most sharply diverge for the two classes of seed items. Once we've identified those contexts, we use them to identify other nouns as members of each gender category. This step is analogous to our identification of the language-specific cues for gender in Welsh.

Third, they build a suffix trie[15] to identify what character sequences at the ends of words can be reliably associated with which gender. This is analogous to our second morphological model in Section 4. Their approach is richer as they don't just consider suffix spans that are uniquely associated with each gender, but suffix spans that can be probabilistically associated with each gender. This enables them to make predictions in more cases than our model.

Cucerzan and Yarowsky test their model on a number of languages (Romanian, French, Spanish, Slovene, and Swedish) and achieve excellent results. They do not test their system on Welsh, however, so this would be an excellent next step. None of the languages Cucerzan and Yarowsky treat are Celtic and none of them have anything like a mutation system for expressing gender. It's unclear then whether they would achieve the same kind of coverage with Welsh or a similar language.

Ayoun (2010) undertakes a similar corpus investigation for French. She shows that fully 49.76% of noun tokens are not marked for gender in her corpus and discusses the implications of this for acquisition of gender categories by second-language learners. As we saw above in Table 12, however, French does not have the same kind of masculine-feminine numerical asymmetry, so it is not

15 A suffix trie is a data structure for storing values associated with strings. Each node in the trie shares a common suffix; all its daughters each potential preceding material.

clear how successful a *guess masculine* strategy would be in French. As we would expect, French does not have a significant class of words like the Welsh indeterminate gender class. Our prediction, based on the Welsh comparison, is that a guessing strategy for French would not be the most successful.

Ours is not the first to apply Bayesian reasoning to problems in the acquisition of morphology. Daland et al. (2007) examine the distribution of defective paradigms in Russian from the perspective of how they might be learned.[16] They model various learning scenarios with explicit Bayesian assumptions and show that, under specific assumptions about how learning proceeds, defective paradigms can be a normal and stable part of a morphological system. This notion of stable defectiveness is quite similar to our results about the indeterminate gender category. That there is such a set of words is a necessary consequence of the nature of Welsh gender cues and the statistical distribution of those cues.

Finally, it's important to point out that our results have been obtained by applying techniques from statistical natural language processing to the gender data. To the extent that our results are interesting theoretically or empirically, or that our results might be useful in a pedagogical or revitalization context, this suggests that as linguists it might be profitable for us to look more closely at the statistical NLP literature for other insights we might obtain with the tools developed there.

Acknowledgements: Thanks to audiences in Tucson and Aberystwyth. Thanks also to Diane Ohala, Dalila Ayoun, Lionel Mathieu, several anonymous reviewers, and the editor for helpful feedback at various stages. All errors are my own.

References

Ayoun, Dalila. 2010. Corpus data: Shedding the light on French grammatical gender or not. *Eurosla Yearbook* 10. 119–141.

Cavnar, William B. & John M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of sdair-94, third annual symposium on document analysis and information retrieval*, 161–175.

Cucerzan, Silviu & David Yarowsky. 2003. Minimally supervised induction of grammatical gender. In *Proceedings of the 2003 conference of the north American chapter of the association for computational linguistics on human language technology-volume 1*, 40–47. Association for Computational Linguistics.

16 See also Sims (in press).

Daland, Robert, Andrea D. Sims & Janet Pierrehumbert. 2007. Much ado about nothing: a social network model of Russian paradigmatic gaps. *Proceedings of the Annual Meeting of the Association of Computational Linguistics* 45. 936–943.

Dorian, Nancy C. 1976. Gender in a terminal Gaelic dialect. *Scottish Gaelic Studies* 12. 279–282.

Ellis, N. C., C. O'Dochartaigh, W. Hicks, M. Morgan & N. Laporte. 2001. Cronfa electroneg o Gymraeg (CEG): A 1 million word lexical database and frequency count for Welsh. http://www.bangor.ac.uk/canolfanbedwyr/ceg.php.en (accessed 5 October 2015).

Gathercole, Virginia C. Mueller & Enlli Môn Thomas. 2001. The acquisition of grammatical gender in Welsh. *Journal of Celtic Language Learning* 6. 53–87.

Hammond, Michael. 2011. Welsh mutations and statistical phonotactics. In Andrew Carnie (ed.), *Formal approaches to Celtic linguistics*, 337–358. Newcastle upon Tyne: Cambridge Scholars Publishing.

Hammond, Michael. 2014. Phonological complexity and input optimization. *Phonological Studies* 17. 85–94.

Hannahs, S. J. 2013. *The phonology of Welsh*. Oxford: Oxford University Press.

King, Gareth. 2003. *Modern Welsh: a comprehensive grammar*. London: Routledge.

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Morgan, T.J. 1952. *Y treigladau a'u cystrawen*. Caerdydd (Cardiff): Gwasg Prifysgol Cymru.

Nodine, Mark H. 2003. Welsh to English lexicon. http://www.cs.cf.ac.uk/fun/welsh/LexiconWE.html (accessed 31 December 2013).

Sims, Andrea D. in press. *Inflectional defectiveness*. Cambridge: Cambridge University Press.

Thomas, Enlli Môn & Virginia C. Mueller Gathercole. 2005a. Minority language survival: Input factors influencing the acquisition of Welsh. In James Cohen, Kara T. McAlister, Kellie Rolstad & Jeff MacSwan (eds.), *Isb4: Proceedings of the 4th International Symposium on Bilingualism*, 852–874. Somerville, MA: Cascadilla Press.

Thomas, Enlli Môn & Virginia C. Mueller Gathercole. 2005b. Minority language survival: Obsolescence or survival for Welsh in the face of English dominance. In James Cohen, Kara T. McAlister, Kellie Rolstad & Jeff MacSwan (eds.), *Isb4: Proceedings of the 4th International Symposium on Bilingualism*, 2233–2257. Somerville, MA: Cascadilla Press.

Watkins, T. Arwyn. 1961. *Ieithyddiaeth: Agweddau ar astudio iaith*. Caerdydd (Cardiff): Gwasg Prifysgol Cymru.