Predicting the gender of Welsh nouns

Michael Hammond
U. of Arizona

A. Overview

(1) Welsh gender exhibits three quite interesting properties:
- 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 usually exposed in the morphology of the noun, but more typically in mutation options.

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

(3) Outline:
- a. descriptive generalizations
- b. basic statistical regularities
- c. morphological model (induction of suffixes)
- d. phonological model (letter-based N-grams)
- e. syntactic models (language-specific properties)
- f. conclusion

B. Descriptive generalizations

(4) Welsh has two grammatical genders: masculine and feminine (King, 2003):

Masculine		Feminine	
pen	'head'	llaw	'hand'
ci	'dog'	coes	'leg'
ceffyl	'horse'	cath	'cat'
gobaith	'hope'	ffatri	'factory'
afal	'apple'	almon	'almond'
...		...	

(5) Generally arbitrary, though terms for animals and people often bear the expected gender:

Masculine		Feminine	
dyn	'man'	dynes	'woman'
mab	'son'	merch	'daughter, girl'
tad	'father'	mam	'mother'
tarw	'bull'	buwch	'cow'
brawd	'brother'	chwaer	'sister'
ceffyl	'horse'	caseg	'mare'
ceiliog	'rooster'	iâr	'chicken'
maharen	'ram'	mamog	'ewe'
...		...	

(6) For some words, gender varies across dialects, speakers, or simply cannot be determined:

abid	'habit'	agendor	'abyss'
anhrefn	'disorder'	cleber	'chatter'
cochl	'cloak'	ennyd	'instant'
ffarwel	'farewell'	gwehelyth	'lineage'
tangnefedd	'peace'	uchelgais	'ambition'

(7) Gender is marked indirectly:

a. pronouns
   (f)e/(f)o 'he' vs. *hi* 'she'
   ei+soft mut. vs. *ei*+asp. mut.
b. soft mutation of fem. sg. noun with article or 1
   ci 'dog (masc)' → *y/un ci* vs. *cath* 'cat (fem)' → *y/un gath*
c. soft mutation of adjectives with fem. sg. noun
   ci bach 'little dog' vs. *cath fach* 'little cat'
d. form of 2, 3, and 4
   dau gi, *tri chi*, *pedwar ci* (masc)
   vs. *dwy gath*, *tair cath*, *pedair cath* (fem)
e. form of certain adjectives
   gwyn 'white (masc)', *trwm* 'heavy (masc)', etc. vs.
   gwen 'white (fem)', *trom* 'heavy (fem)', etc.
f. form of demonstratives
   hwn/hwnnw vs. *hon/honno*

C. Basic statistical regularites

(8) The tagged CEG corpus (Ellis et al., 2001) contains 1,223,649 word tokens:

	Noun tokens		Noun types	
Masc.	120,646	64%	5302	69%
Fem.	57,178	30%	2037	27%
Indet.	11,598	6%	303	4%

(9) This overall distribution suggests a benchmark strategy for determining the gender of a noun: *guess masculine*. We'd be right 64% of the time in the CEG corpus.

(10) This is not a general fact about gender systems:

Language	Corpus	Masculine	Feminine	Neuter
Spanish	IULA	69901	72088	NA
		49%	51%	NA
French	Lexique380	26744	18925	NA
		59%	41%	NA
German	CELEX	10786	13688	6005
		35%	45%	20%
Dutch	CELEX	27925	24819	21795
		37%	33%	29%
Russian	Russ. Nat. Corp.	37737473	27962098	14214372
		47%	35%	18%

(11) General comparison:

a. in languages with three genders, neuter is generally underrepresented;
b. Welsh has the most extreme masculine/feminine difference.

D. Morphological models

(12) Frequent suffixes with masculine nouns:

-deb	absenoldeb 'absence', cytundeb 'agreement', diddordeb 'interest'
-iant	methiant 'failure', moliant 'praise', peiriant 'engine'
-yn	(a)deryn 'bird', bathodyn 'badge', blodyn 'flower'
-iad	adolygiad 'review', benthyciad 'borrowing', canlyniad 'consequence'
-wr	adarwr 'bird-catcher', Albanwr 'Scot', arbenigwr 'specialist'
-ydd	ieithydd 'linguist', anarchydd 'anarchist', darllenydd 'reader'
-wch	anialwch 'desert', ariangarwch 'avarice', harddwch 'beauty'
-ter/der	balchder 'pride', anhoffter 'dislike', dyfnder 'depth'
-rwydd	anghofrwydd 'forgetfulness', cwrteisrwydd 'courtesy', hapusrwydd 'happiness'

(13) Frequent suffixes with feminine nouns:

-aeth	absenoliaeth 'absence', llofruddiaeth 'murder', cystadleuaeth 'competition'
-en	afallen 'apple tree', cangen 'branch', deilen 'leaf'
-wraig	tafarnwraig 'bar maid', cantwraig 'singer', golchwraig 'washerwoman'
-es	arthes 'she-bear', awdures 'authoress', Eifftes 'Egyptian (female)'
-fa	allanfa 'exit', cuddfa 'hiding place', meddygfa 'surgery'

(14) Morphological model #1:
Using just these affixes, we can correctly assign gender to 37,165 noun tokens (20%) and incorrectly to 8,997 noun tokens (5%) in the CEG corpus.

(15) Morphological model #2:
Combine model #1 with the *guess masculine* strategy. If we can identify a suffix, the guess for gender is based on that; if no affix can be identified, guess masculine. Using this combined strategy, we get 126,694 word tokens correct (67%) and 62,728 word tokens incorrect (33%).

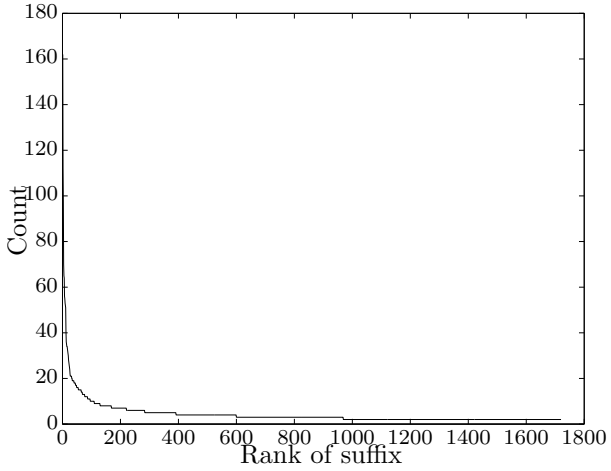
(16) Morphological model #3:

a. using the CEG corpus, find all final letter strings unambiguously associated with either gender and hypothesize that these are suffixes;
b. using an electronic dictionary, Nodine (2003), extract all nouns that do not occur in the first corpus;
c. compare gender assignments predicted by the hypothetical suffixes with actual genders.

(17) #3 is not quite morphology:

a. Some true suffixes don't qualify because they are not gender-unique, e.g. the string *-deb* can occur with the noun *diweddeb* with indeterminate gender;
b. The string *-ldeb* is not a true suffix, but is uniquely associated with masculine.

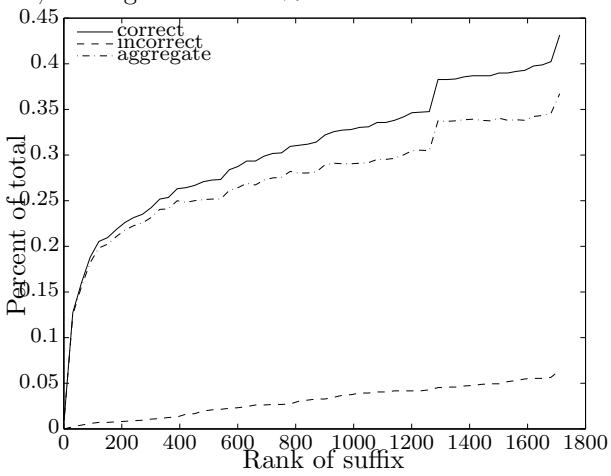
(18) This results in 1720 candidate suffixes from the CEG corpus. Counts for each morpheme ranged from 162 to 2. These are distributed in typical Zipfian fashion:



(19) To test these, we use nouns not in the CEG corpus and so we drew these 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:

Gender	Count	Percent
Masculine	1202	72%
Feminine	433	26%
Indeterminate	45	3%

(20) Testing hypothesized suffixes incrementally, testing more common hypothesized suffixes before less common ones, never gets above 45%:



E. Phonological models: letter-based N-grams

(21) N-gram models are built on the notion of *conditional probability*:

$$\mathrm{p}(rain|clouds) = \frac{\mathrm{p}(rain\ and\ clouds)}{\mathrm{p}(clouds)}$$

(22) A bigram model of words characterizes the probability of a word as the product of the conditional probabilities of its letter sequences:

$$\mathrm{p}(deryn) = \mathrm{p}(d|\#) \times \mathrm{p}(e|d) \times \mathrm{p}(r|e) \times \mathrm{p}(y|r) \times \mathrm{p}(n|y)$$

(23) Building a bigram model of gender in Welsh:
- a. calculate the probability of each gender category overall: P_m, P_f, P_i;
- b. calculate the letter sequence probabilities separately for each gender: M_m, M_f, M_i;
- c. To determine the hypothetical gender of a word, multiply its bigram value by the probability of the gender for each gender and choose the one that has the highest value: $P_m \times M_m$ vs. $P_f \times M_f$ vs. $P_i \times M_i$ (Bayes' Law).

(24) Using this, we get 1,215 out of 1,633 novel words in the Nodine dictionary correct (74%). This is better than what we've gotten with the previous models, but still only marginally better than what the *guess masculine* strategy would yield: 72%.

(25) Interim summary:
- a. N-gram approaches perform better than the morphological approaches, but do not significantly outperform the *guess masculine* strategy.
- b. Cast 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.

F. Syntax

(26) How often does the possibility of soft mutation after the definite article reveal the gender of a noun?

	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

(27) How often does the form of the adjective (e.g. *gwyn* vs. *wen*) reveal the gender of a noun?

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

(28) How often does the form of the number (e.g. *dau* vs. *dwy*) reveal the gender of a noun?

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

(29) How often does the form of the demonstrative (e.g. *hwn* vs. *hon*) reveal the gender of a noun?

Noun.	Dem.	Count	Frequency
Masc.	Masc.	1374	0.01139
Masc.	Fem.	32	0.00027
Fem.	Masc.	38	0.00066
Fem.	Fem.	965	0.01688

(30) How often does the form of the adjective (e.g. *mawr* vs. *fawr*) reveal the gender of a noun?

Noun.	Mut'able adj.	Count	Frequency
Masc.	unmutated	8230	0.06822
Masc.	mutated	373	0.00309
Fem.	unmutated	456	0.00798
Fem.	mutated	4290	0.07503

(31) Putting it together:

a. Using all language-specific cues, the gender of 37% of word *tokens* can be identified.
b. Using all language-specific cues, the gender of 36% of word *types* can be identified.
c. If we use the *guess masculine* strategy on remaining items, then we can successfully identify the gender of 91% of word types.

G. Conclusion

(32) General conclusions:

a. morphology does not suffice to identify gender;
b. phonology does not suffice to identify gender;
c. only language-specific properties—taken together—suffice to do better than simply guessing masculine for all words.

(33) Predictions:

a. gender will only be acquired reliably when the rest of the grammar is under control;
b. if the rest of the grammar—and specifically the mutation system—cannot be used to identify gender, then the gender system will be lost;
c. we expect the class of words with indeterminate gender to coincide with words whose gender cannot be identified on the basis of language-specific properties;
d. if the gender system is lost, what we expect is the masculine pattern to generalize and the feminine pattern to typify an ever-shrinking class of exceptional words.

(34) Methodological conclusions:
Statistical techniques from statistical natural language processing can enrich our understanding of language.

H. 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.

Coleman, John & Janet Pierrehumbert (1997). Stochastic phonological grammars and acceptability. In *Computational Phonology: Third meeting of the ACL Special Interest Group in Computational Phonology*. Somerset: Association for Computational Linguistics, 49–56.

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*. Association for Computational Linguistics, 40–47.

Davies, P., J. Stammers, E. Robert, S. Wynn Lloyd, M. C. Parafita Couto, J. Herring & M. Deuchar (2009). Creating code-switching corpora: Welsh-English and Spanish-English. Paper presented at ISB7, Utrecht, Netherlands.

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`.

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 (2013). Input optimization in English. *Journal of the Phonetic Society of Japan* **17**:1–12.

Hammond, Michael (in press). Calculating syllable count automatically from fixed meter poetry in English and Welsh. *Journal of Literary and Linguistic Computing*.

Hannahs, S. J. (2013). *The Phonology of Welsh*. Oxford: Oxford University Press.

Jones, Mari C (1998). *Language Obsolescence and Revitalization: Linguistic Change in Two Sociolinguistically Contrasting Welsh Communities*. Oxford: Clarendon 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 Dec. 31, 2013.

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*. Sumerville, MA: Cascadilla Press, 852–874.

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*. Sumerville, MA: Cascadilla Press, 2233–2257.

Thomas, Enlli Môn & Virginia C. Mueller Gathercole (2007). Children's productive command of grammatical gender and mutation in Welsh: An alternative to rule-based learning. *First Language* **27**:251–278.

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