A Novel Approach to Categorising Extrinsic Lexical Units in a Corpus of Welsh-Language Tweets

Bethan Siân Tovey
Swansea University

Prifysgol Abertawe
Swansea University

CorCenCC

1. What is an Extrinsic Lexical Unit?
2. Extrinsic Lexical Units in Welsh
3. Pragmatic Markers
4. Methodology and Data Analysis
5. Summation

Prifysgol Abertawe
Swansea University

What is an Extrinsic Lexical Unit?

(and why am I calling them that?)

Prifysgol Abertawe
Swansea University

CorCenCC

Terminology

Poplack (2018):
"donor language" → "native language" or "recipient language"

Grosjean (2008):
"base language"
or → "guest language"
"host language"

Prifysgol Abertawe
Swansea University

CorCenCC

NATIVE LANGUAGE

FOREIGN LANGUAGE

Prifysgol Abertawe
Swansea University

CorCenCC

DONOR LANGUAGE

RECIPIENT LANGUAGE

Prifysgol Abertawe
Swansea University

CorCenCC

HOST LANGUAGE

GUEST LANGUAGE

Prifysgol Abertawe
Swansea University

CorCenCC

Intrinsic

Latin *intrinsecus* "internally, from within".

Extrinsic

Latin *intrinsecus* "externally, from outside".

Prifysgol Abertawe
Swansea University

CorCenCC

whole items formed from language-intrinsic morphemes

whole items formed language-extrinsically

Prifysgol Abertawe
Swansea University

CorCenCC

extrinsic-language
morpheme(s)
used to form
longer words

Prifysgol Abertawe
Swansea University

CorCenCC

Whole ELUs

aquarium
eau-de-vie
whisky
feng shui
axolotl

WATER
WATERY
WATERING
WATERED
WATERS

EL morphemes

aqualung
waterable
aquaphobia
vodkas
whiskyish

Prifysgol Abertawe
Swansea University

CorCenCC

extrinsic root + intrinsic root	aqua lung
intrinsic root + extrinsic derivational	water able
extrinsic root + extrinsic root	aqua phobia
extrinsic root + intrinsic derivational	whisky ish
extrinsic root + intrinsic inflectional	vodka s

Prifysgol Abertawe
Swansea University

CorCenCC

extrinsic root + intrinsic root	aqua lung
intrinsic root + extrinsic derivational	water able
extrinsic root + extrinsic root	aqua phobia
extrinsic root + intrinsic derivational	whisky ish
extrinsic root + intrinsic inflectional	vodka s

Prifysgol Abertawe
Swansea University

CorCenCC

Intrinsic Lexical Unit

* developed directly from intrinsic lexical resources

or

* developed from extrinsic unit(s), using intrinsic derivational/compounding processes, to create a unit that dœs not exist in the extrinsic language itself

Prifysgol Abertawe
Swansea University

CorCenCC

Extrinsic Lexical Unit

* has a (near-)identical counterpart in another language with which there is a history of contact (allowing for differences of spelling and inflection)
* has a meaning similar to, or derived from, the meaning of that counterpart

Prifysgol Abertawe
Swansea University

CorCenCC

Extrinsic Lexical Units in Welsh

(borrowings, loanwords, or something else?)

Prifysgol Abertawe
Swansea University

CorCenCC

History of Language Contact

Brythonic: Other Celtic languages; Vulgar Latin

Old Welsh: Other Celtic languages; Old English

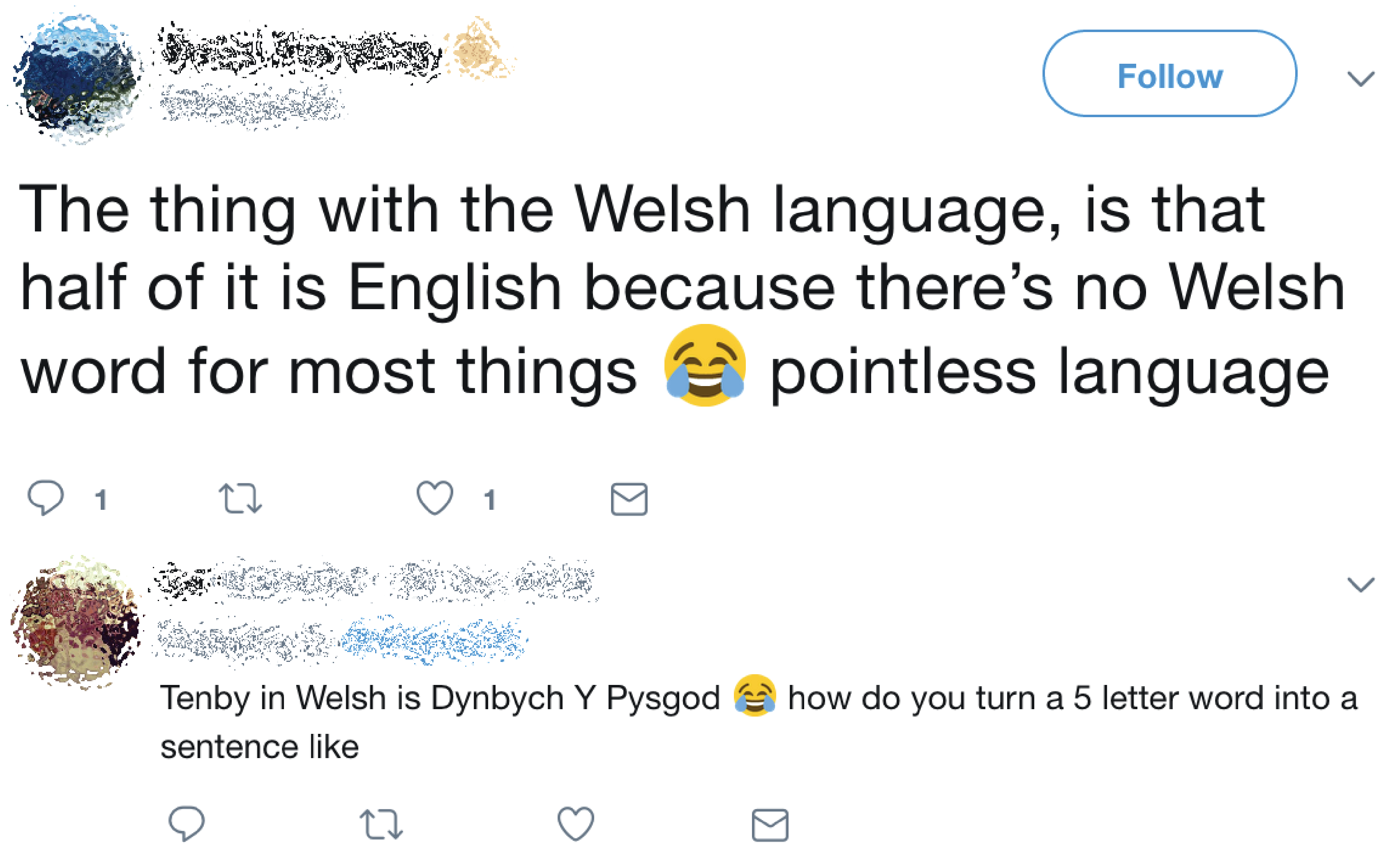
Middle Welsh: Middle English; Flemish; Irish; French

Modern Welsh: Modern English; BSL; a growing range of other languages

Prifysgol Abertawe
Swansea University

CorCenCC

Attitudes towards English Words

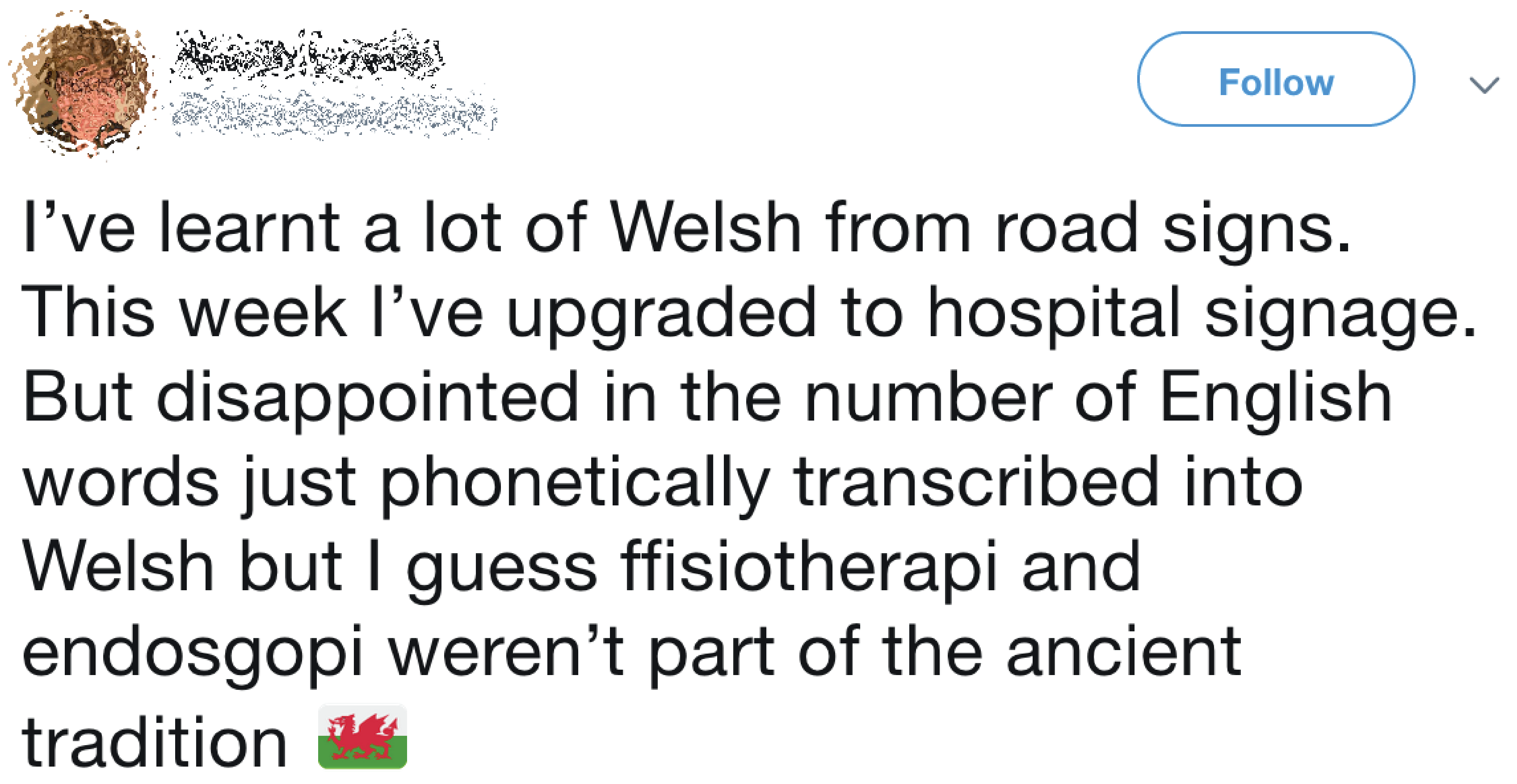


[redacted]

The thing with the Welsh language, is that half of it is English because there's no Welsh word for most things 😂 pointless language

[redacted]

Tenby in Welsh is Dynbych Y Pysgod 😂 how do you turn a 5 letter word into a sentence like



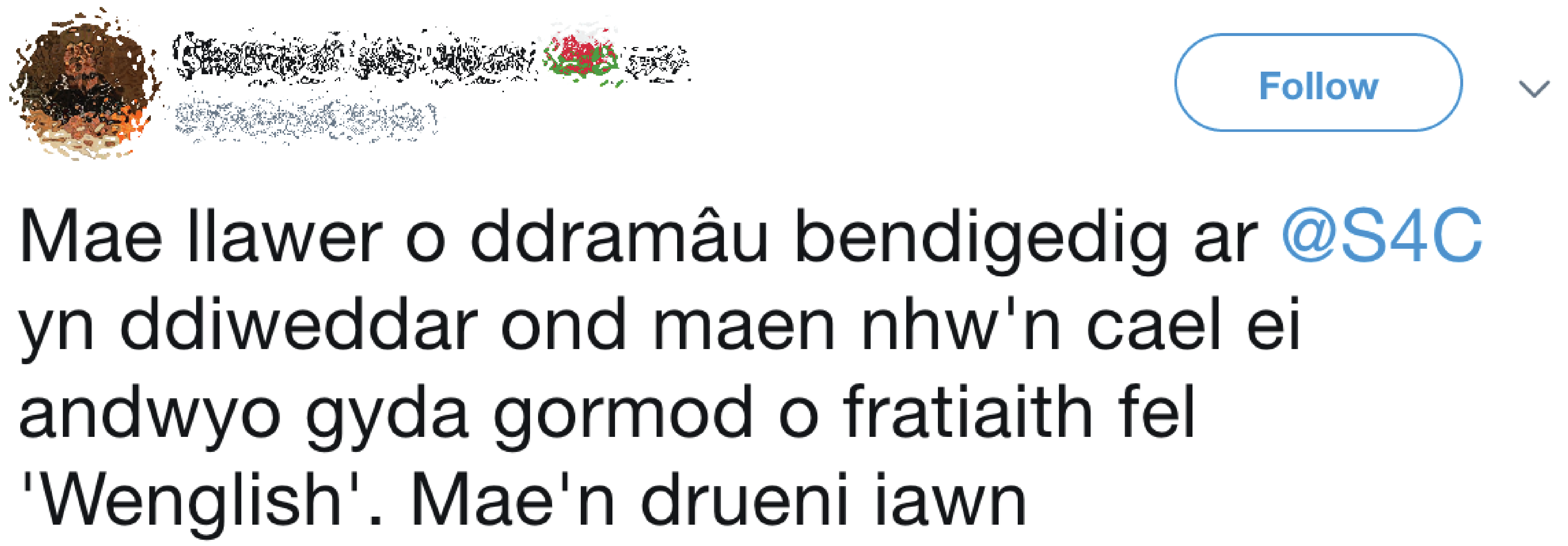
[redacted]

I've learnt a lot of Welsh from road signs. This week I've upgraded to hospital signage. But disappointed in the number of English words just phonetically transcribed into Welsh but I guess ffisiotherapi and endosgopi weren't part of the ancient tradition 🏴󠁧󠁢󠁷󠁬󠁳󠁿

Prifysgol Abertawe
Swansea University

CorCenCC

Attitudes towards English Words



[redacted]

Follow

Mae llawer o ddramâu bendigedig ar @S4C yn ddiweddar ond maen nhw'n cael ei andwyo gyda gormod o fratiaith fel 'Wenglish'. Mae'n drueni iawn

"There are lots of fantastic dramas on S4C recently, but they're all ruined with too much slang like 'Wenglish'. It's a real shame."

Prifysgol Abertawe
Swansea University

Borrowings from English(?)

saim ("grease")
from M.E. "seim" or Old French "saim"

babi ("baby")
from M.E. "baby" - or the other way around?
(Durkin, 2014)

tacsi ("taxi")
from Mod.E. "taxi" - or is it just an internationalism?

Prifysgol Abertawe
Swansea University

CorCenCC

Why Does It matter?

The vocabulary we choose to accept as "Welsh" can influence

* outcomes for children in Welsh-medium education

* interactional competence in Welsh learners

* confidence/willingness to speak Welsh

Prifysgol Abertawe
Swansea University

CorCenCC

Borrowings or Codeswitches?

Poplack and others:

* B and CS are separate categories
* Bs are morphosyntactically and phonologically integrated
* CSs retain extrinsic-langage morphosyntax and phonology (phonological criteria have been modified over time by Poplack)

Myers-Scotton and others:

* B and CS form a continuum
* Bs are frequently-used items
* Both Bs and CSs will be morphosyntactically integrated to some extent
* phonological integration may depend on the social status on the extrinsic language

Prifysgol Abertawe
Swansea University

Borrowings or Codeswitches?

Deuchar, Webb-Davies and Donnelly (2018):

"Listedness" - is the extrinsic lexical unit in a "community dictionary"?

"Listed" verbs: rate of soft mutation similar to that of intrinsic verbs.

"Unlisted" verbs: lower mutation, less frequent.

∴ Borrowings are items with high or low frequency and high integration

Code-switches are items with low frequency and low integration

Prifysgol Abertawe
Swansea University

CorCenCC

Pragmatic Markers

(and why they're a pain in the behind-head)

Prifysgol Abertawe
Swansea University

CorCenCC

What are Pragmatic Markers?

Characteristics of PMs:

* units of one or more words which carry metalinguistic information

* any propositional meaning relates to text/context construction

* occur at clausal/phonological/discourse boundaries

* can be removed from a sentence without grammatical violation

Prifysgol Abertawe
Swansea University

CorCenCC

What are Pragmatic Markers?

Cytuno! Dim ond y creme de la creme sy'n deall wedyn! Bach o ffrangeg fyna hefyd #deallYCwbwl haha

"Agreed! Only the creme de la creme understand then! A bit of French there too! #understandItAll haha"

Prifysgol Abertawe
Swansea University

CorCenCC

The Problem of Extrinsic PMs

Stammers (2010)
Tests for morphological integration of English verbs into Welsh:

* Do they show soft mutation in the appropriate contexts?

wnes	*i*	*ddjio*	*ar*	*gyfer*	*priodas*	*ffrindiau*
do-PST-1S	I	DJ-VBZ	for		wedding	friend-PL

"I DJed for [my] friends' wedding"

The Problem of Extrinsic PMs

Stammers (2010)
Tests for morphological integration of English verbs into Welsh:

* Do they show soft mutation in the appropriate contexts?
* Do they appear only in periphrastic constructions, or also with synthetic morphology?

nid	fi	dwîtiodd	hwn
not	me	tweet-PST	this-MASC-SG

"I didn't tweet this"

Prifysgol Abertawe
Swansea University

CorCenCC

The Problem of Extrinsic PMs

The formation of adverbs in Welsh:

* Usually accomplished with an adverbializer particle, "yn", + adjective
* This "yn" triggers soft mutation on the relevant consonants

mae	pob	llyfr	ar	gael	yn	ddigidol
be-3SG-PRS	every	book	for	getting	PTCL-ADVZ	digital

"every book is available digitally"

Prifysgol Abertawe
Swansea University

CorCenCC

The Problem of Extrinsic PMs

BUT English PMs in adverbial form are often borrowed whole:

absoliwtly bril
"absolutely bril[liant]"

Prifysgol Abertawe
Swansea University

The Problem of Extrinsic PMs

Tests for phonological integration of English words into Welsh:

* Almost all Welsh speakers are Welsh-English bilinguals
* Their Welsh-English tends to be pronounced with Welsh phonology
* Extrinsic English-origin items in Welsh largely show no difference from the same items as used in the same speakers' English

Prifysgol Abertawe
Swansea University

CorCenCC

A Potential Alternative?

- Welsh and English orthographic systems are quite different:

cuddle	/ˈkɪðlɛ/ (CY)	/ˈkʌdl/ (EN)
dial	/ˈdiːæl/ (CY)	/ˈdaɪəl/ (EN)
/maɪs/	mice (EN)	maes (CY)

Prifysgol Abertawe
Swansea University

CorCenCC

A Potential Alternative?

* Welsh and English orthographic systems are quite different
* Long-established loanwords are usually orthographically adapted:

sialens	dawns	coffi
"challenge"	"dance"	"coffee"
(1547)	C14	C16
	(from M.E. "daunce")	

Prifysgol Abertawe
Swansea University

CorCenCC

A Potential Alternative?

* Welsh and English orthographic systems are quite different
* Long-established loanwords are usually orthographically adapted
* Orthographic adaptation should affect PMs as much as other words

Prifysgol Abertawe
Swansea University

CorCenCC

Methodology and Data Analysis

(and maybe a few conclusions)

Prifysgol Abertawe
Swansea University

The Corpus

* Roughly 1m words

* Tweets written from 2007-2018 with Welsh as the main language

* All data anonymized before analysis

* Chosen to reflect spontaneous, spoken Welsh as closely as possible

Prifysgol Abertawe
Swansea University

CorCenCC

The Initial Analysis - DERWen

a.k.a.

Dichell Esboniadol Ramadegol y Wenglish (Wenglish Grammatical-Explanatory Device)

Prifysgol Abertawe
Swansea University

CorCenCC

The Initial Analysis - DERWen

* Part-of-speech tagger for code-mixed Welsh-English bilingual data, including non-standard orthography and dialectal forms

* Uses phonological translation of orthographic forms to find English words in Welsh orthography

* Based on initial work by the CorCenCC team

* Coming to GitHub soon...

Prifysgol Abertawe
Swansea University

CorCenCC

A Closer Look at a Selection of PMs

* 60 initial candidate PMs (e.g. *awesome*, *hooray*, *no way*)

* 33 of these were sufficiently prevalent to make it into the final analysis

* Coded for orthography (Welsh, English, Mixed)

* Coded for presence in two major Welsh dictionaries

* Coded for frequency in a reference corpus of English tweets

Prifysgol Abertawe
Swansea University

CorCenCC

A Closer Look at a Selection of PMs

Welsh = orthography entirely adapted to Welsh:

eniwê siriysli ôsym

Prifysgol Abertawe
Swansea University

CorCenCC

A Closer Look at a Selection of PMs

Welsh = orthography entirely adapted to Welsh

English = orthography entirely unadapted to Welsh:

literally   definately   awsum

Prifysgol Abertawe
Swansea University

CorCenCC

A Closer Look at a Selection of PMs

Welsh = orthography entirely adapted to Welsh

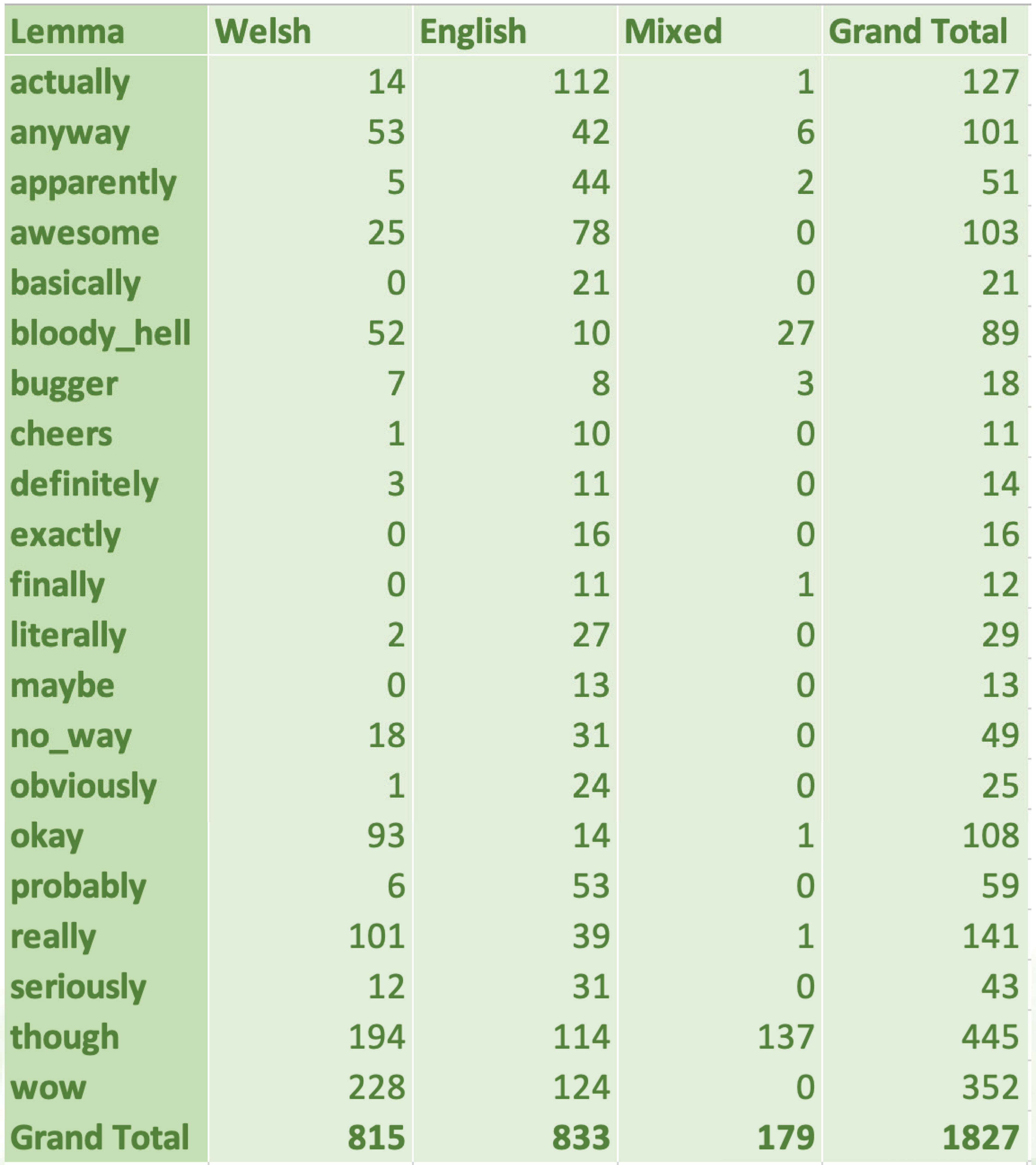
English = orthography entirely unadapted to Welsh

Mixed = features of both orthographies:

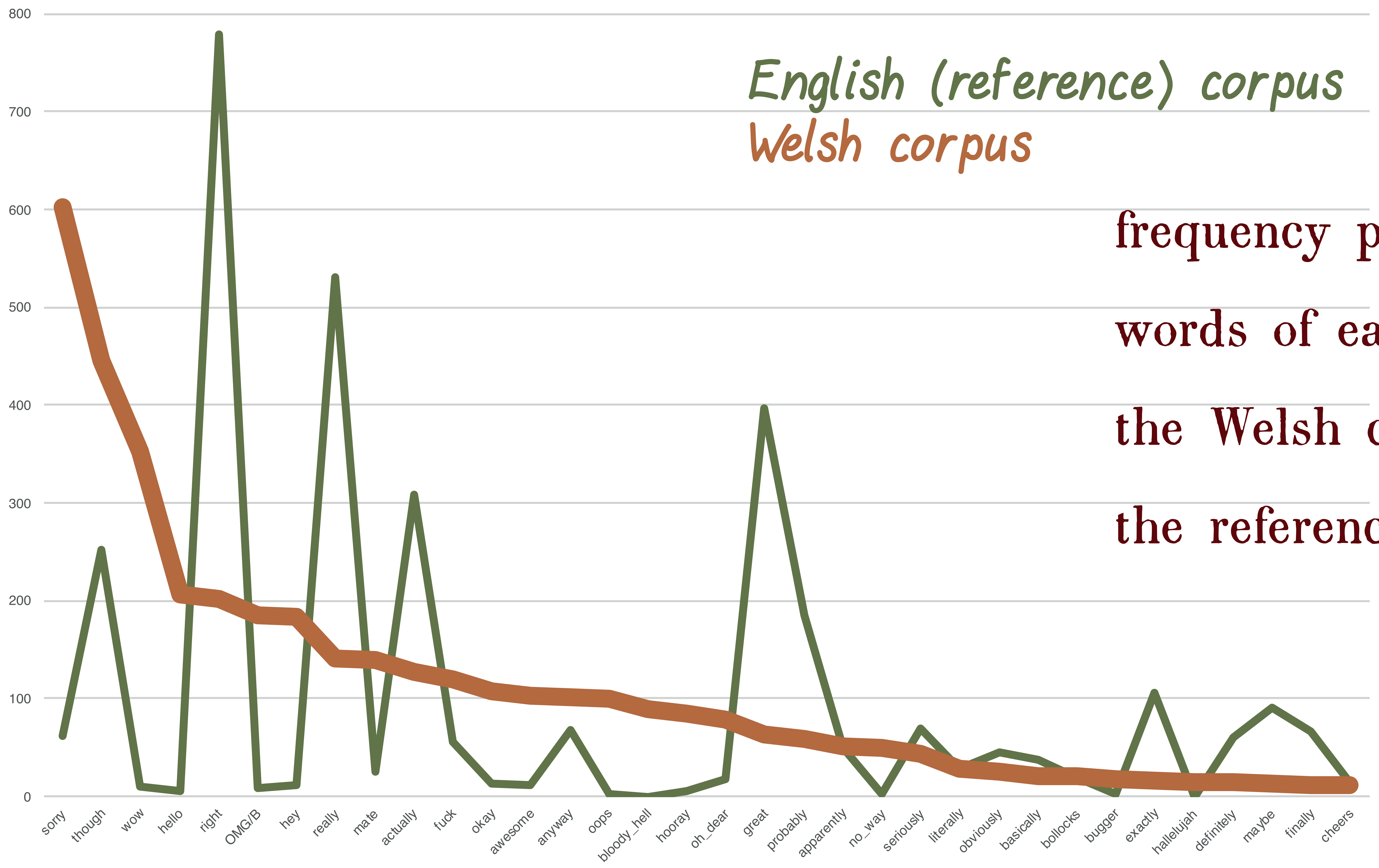
tho   oh diar   finali

Prifysgol Abertawe
Swansea University

CorCenCC



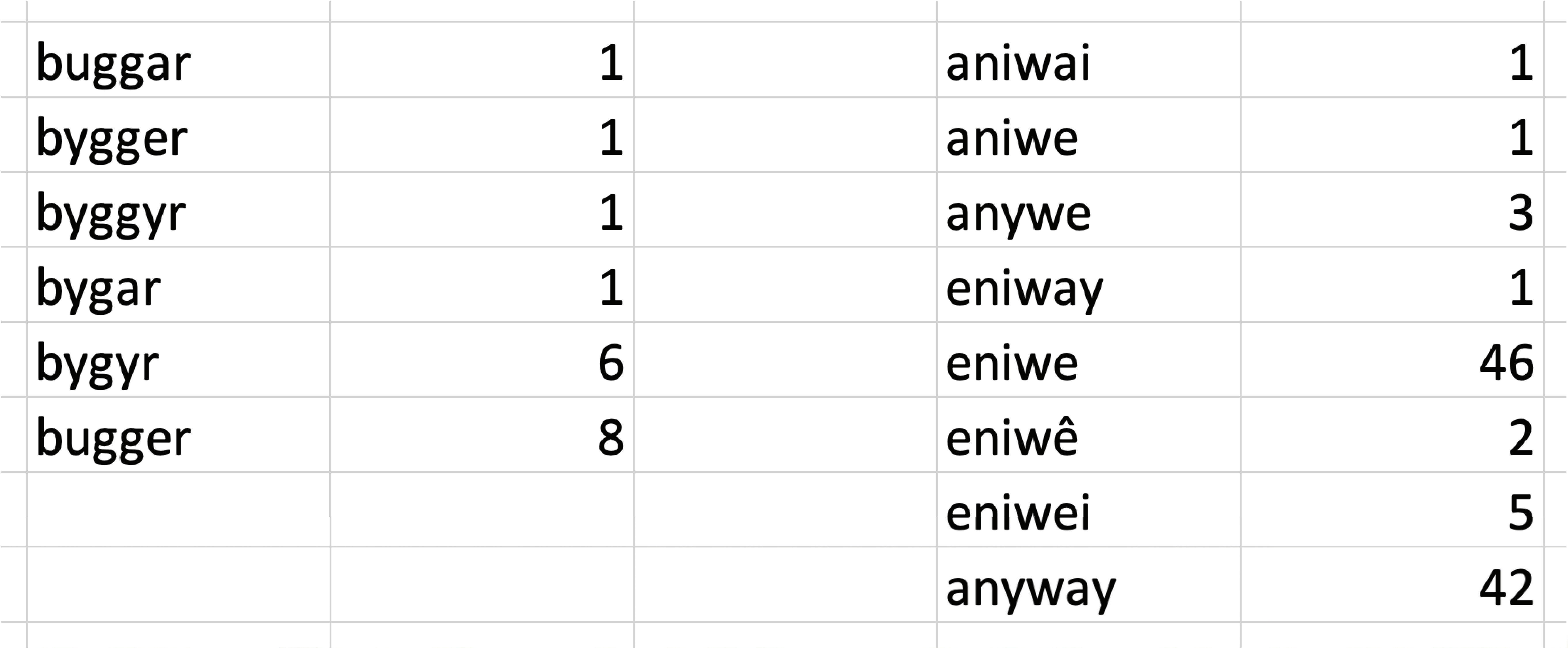
Lemma	Welsh	English	Mixed	Grand Total
actually	14	112	1	127
anyway	53	42	6	101
apparently	5	44	2	51
awesome	25	78	0	103
basically	0	21	0	21
bloody_hell	52	10	27	89
bugger	7	8	3	18
cheers	1	10	0	11
definitely	3	11	0	14
exactly	0	16	0	16
finally	0	11	1	12
literally	2	27	0	29
maybe	0	13	0	13
no_way	18	31	0	49
obviously	1	24	0	25
okay	93	14	1	108
probably	6	53	0	59
really	101	39	1	141
seriously	12	31	0	43
though	194	114	137	445
wow	228	124	0	352
Grand Total	**815**	**833**	**179**	**1827**



frequency per million words of each PM in the Welsh corpus and the reference corpus

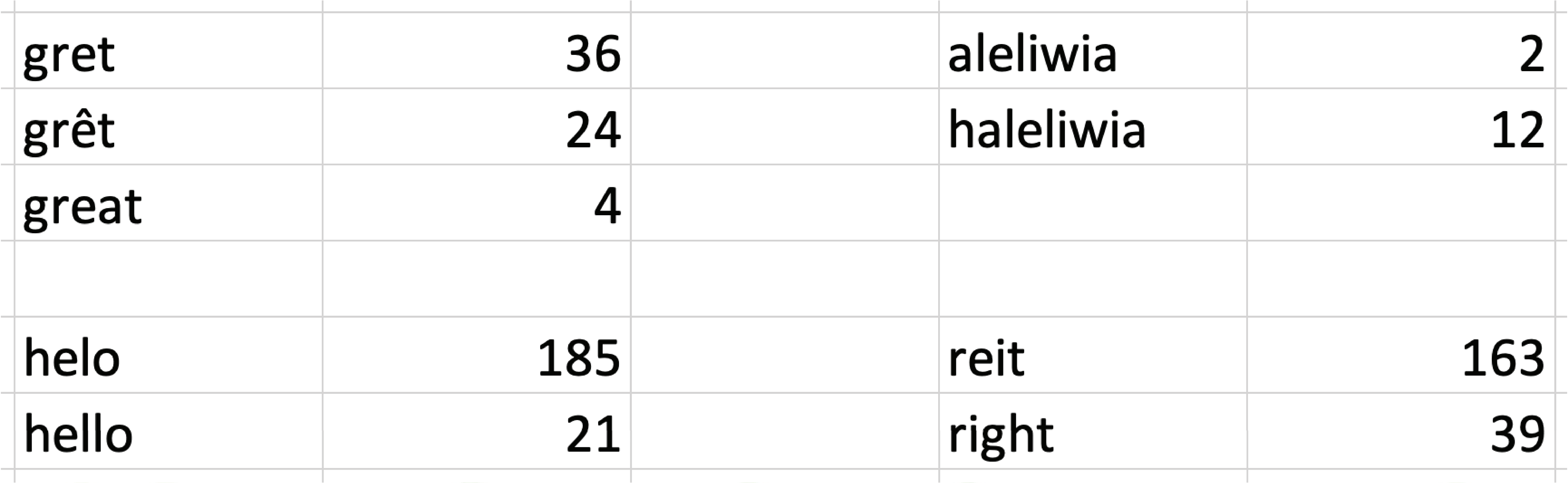
Prifysgol Abertawe
Swansea University

CorCenCC

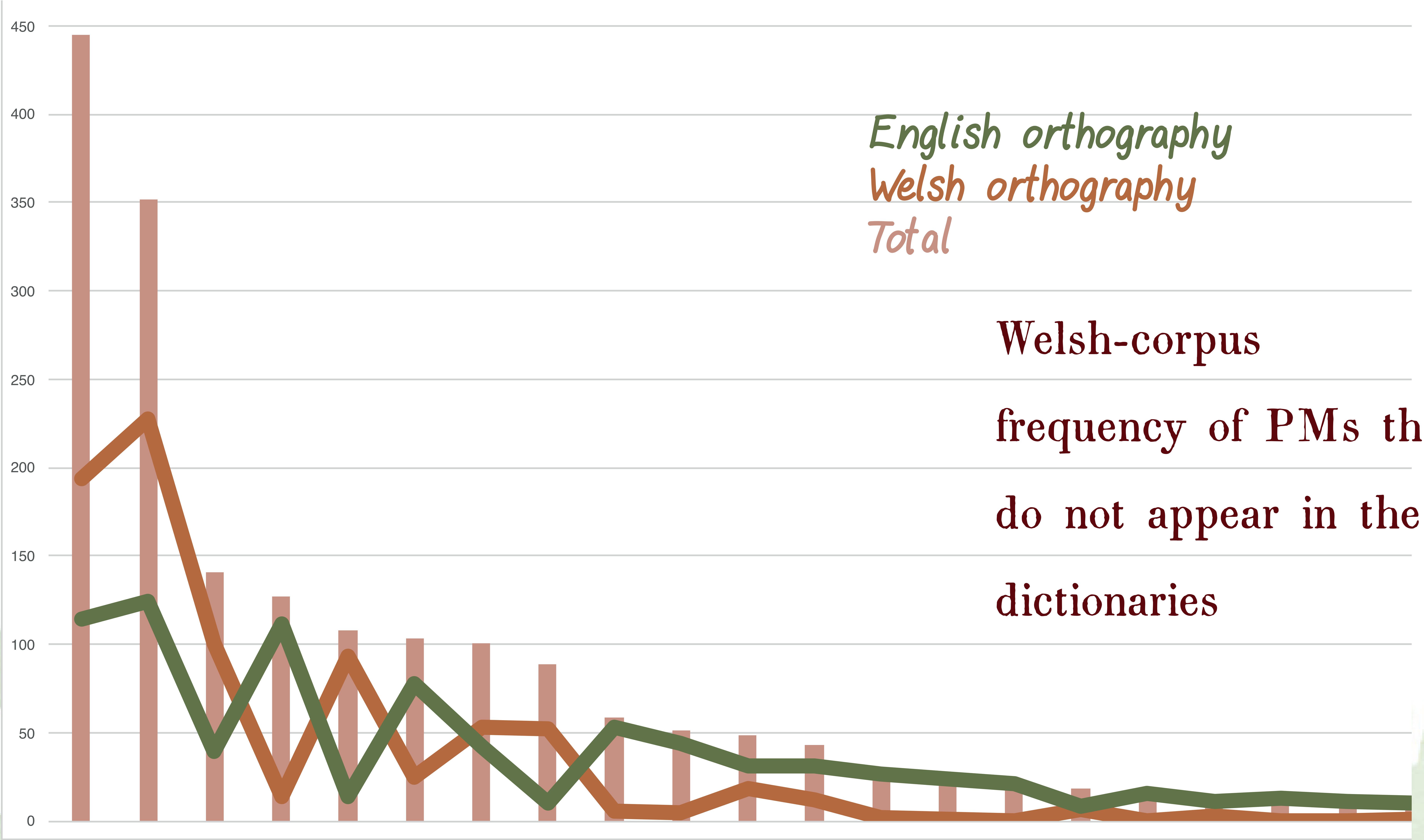


buggar	1		aniwai	1
bygger	1		aniwe	1
byggyr	1		anywe	3
bygar	1		eniway	1
bygyr	6		eniwe	46
bugger	8		eniwê	2
			eniwei	5
			anyway	42

Prifysgol Abertawe
Swansea University



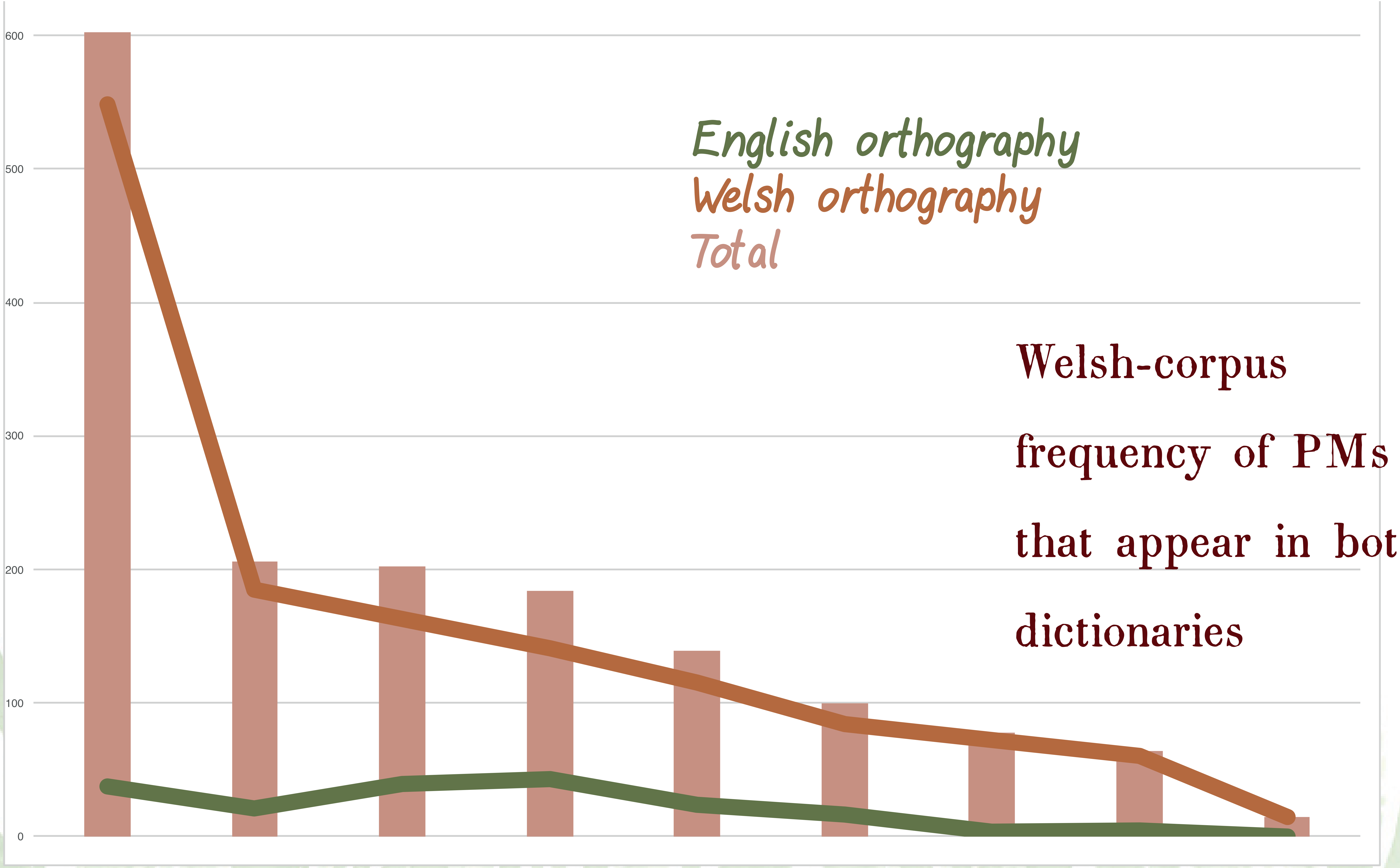
gret	36		aleliwia	2
grêt	24		haleliwia	12
great	4			
helo	185		reit	163
hello	21		right	39



Welsh-corpus frequency of PMs that do not appear in the dictionaries

Prifysgol Abertawe
Swansea University

CorCenCC



Welsh-corpus frequency of PMs that appear in both dictionaries

Prifysgol Abertawe
Swansea University

CorCenCC

Summation

(and what's next)

Prifysgol Abertawe
Swansea University

CorCenCC

Diolch! Tige tank!

Prifysgol Abertawe
Swansea University

CorCenCC