

# ChronHib

CHRONOLOGICON  
HIBERNICUM

## Statistical methods in the Old Irish language: A methods point of view of the ChronHib project

Marco A. Aquino-López, David Stifter  
(Maynooth University)

**Statistics in Historical Corpus Linguistics**

**Maynooth University**  
**4-5 Oct 2019**

# Acknowledgements



National University of Ireland Grant Scheme  
for Early Career Academics



Maynooth University Arts and Humanities Institute



European Research Council, H2020 #647351

- Early Irish
- Proto-Goidelic before 4<sup>th</sup> c. A.D.
  - Primitive Irish 4<sup>th</sup>-6<sup>th</sup> c.
  - **Archaic Old Irish** 7<sup>th</sup> c.
  - **Old Irish (Olr.)** 8<sup>th</sup>-9<sup>th</sup> c.
  - Middle Irish (Mlr.) 10<sup>th</sup>-12<sup>th</sup> c.
  - Early Modern Irish 13<sup>th</sup>-16<sup>th</sup> c. (dialectal differentiation of Scottish Gaelic & Manx)
  - dialectal Modern Irish 17<sup>th</sup>-21<sup>st</sup> c. (Munster, Connacht, Ulster; *An Caighdeán Oifigiúil*)



# Challenges & Research Questions

## **Old Irish:**

- a vast amount of text surviving

## **Challenges:**

- few contemporary MSS ('glosses')
- mostly late MS transmission (up to 1000 years after composition)
- unknown dates of composition
- unknown authors

## **ChronHib:**

- **Can we derive narrow time-spans for dating texts from the linguistic variation which they exhibit?**
- **Can we determine the probabilities of dates for specific texts, and define time ranges for the linguistic variation and changes?**



# Statistics in Early Irish

## Frequency statistics, e.g.:

Liam Breatnach, 'The suffixed pronouns in early Irish', *Celtica* 12 (1977).

Warren Cowgill, 'On the Prehistory of Celtic Passive and Deponent Inflection', *Ériu* 34 (1983).

Kim McCone, 'The Nasalizing Relative Clause with Object Antecedent in the Glosses', *Ériu* 31 (1980).

Ruairí Ó hUiginn, 'Complementation in Early Irish: the *verba dicendi*', *Ériu* 49 (1998).

Peter Schrijver, *Studies in the History of Celtic Pronouns and Particles*, Maynooth 1997.

Elisa Roma, 'How Subject Pronouns Spread in Irish', *Ériu* 51 (2000).

Elisa Roma, 'How Many Definiteness Markers per NP in Old Irish?', *Proceedings XIII. ICC Bonn*, 2009.

## Hypothesis testing:

Anthony Harvey, 'The Ogam Inscriptions and their Geminate Consonant Symbols', *Ériu* 38 (1987).

Aaron Griffith, 'The Animacy Hierarchy and the Distribution of the *notae augentes* in Old Irish', *Ériu* 58 (2008).

# Statistics in Early Irish

## Frequency statistics, e.g.:

Liam Breatnach, 'The suffixed pronouns in early Irish', *Celtica* 12 (1977).

Warren Cowgill, 'On the Prehistory of Celtic Passive and Deponent Inflection', *Ériu* 34 (1983).

Kim McCone, 'The Nasalizing Relative Clause with Object Antecedent in the Glosses', *Ériu* 31 (1980).

Ruairí Ó hUiginn, 'Complementation in Early Irish: the *verba dicendi*', *Ériu* 49 (1998).

Peter Schrijver, *Studies in the History of Celtic Pronouns and Particles*, Maynooth 1997.

Elisa Roma, 'How Many Definiteness Markers per NP in Old Irish?', *Proceedings XIII. ICC Bonn*, 2009.

## Hypothesis testing:











Anthony Harvey, 'The Ogam Alphabet and its Component Symbols', *Ériu* 38 (1987).

Aaron Griffith, 'The Animacy Hierarchy and the Distribution of the *notae augentes* in Old Irish', *Ériu* 58 (2008).



**“statistical  
prestidigitation”**

# Corpora in Early Irish Studies

-  *Thesaurus Palaeohibernicus*. 2 vols, ed. Wh. Stokes & J. Strachan, 1901-3
-  *Dictionary of the Irish Language (DIL)*, ed. G. Quin, 1976
-  *electronic Dictionary of the Irish Language (eDIL)*, ed. G. Toner, 2007-
-  *Corpus of Electronic Texts (CELT)*, UCC 1997-
-  *Thesaurus Linguae Hibernicae (TLH)*, UCD 2006-11
-  *Lexicon of Würzburg Glosses*, ed. S. Kavanagh, 2001
-  *Milan Glosses Database*, ed. A. Griffith, 2006-11
-  *Priscian Glosses Database*, ed. B. Bauer, 2010-14
-  *Lexicon of Blathmac's Poems*, ed. S. Barrett, 2013-7
-  *Parsed Old and Middle Irish Corpus (POMIC)*,  
ed. E. Lash, 2011-4
-  *Annals of Ulster Database*, ed. F. Qiu, 2015-7 (ChronHib)

Corpus  
Palaeo  
Hibernicum  
(CorPH –  
110.000  
tokens)

# The Old Irish Article

	neut.	masc.	fem.	total	<i>inna/na</i>
<b>Würzburg</b>	290 (27.8%)	478 (44.1%)	316 (29.2%)	<b>1084</b>	146 (13.5%)
<b>Milan</b>	1274 (28.9%)	1937 (44%)	1194 (27.1%)	<b>4405</b>	859 (19.5%)
<b>St Gall</b>	532 (41.8%)	407 (31.9%)	335 (26.3%)	<b>1274</b>	179 (14.1%)
<b>total</b>	<b>2096</b>	<b>2822</b>	<b>1845</b>	<b>6663</b>	

## OLD-IRISH PARADIGMS

### THE ARTICLE

#### SINGULAR

<i>Masculine</i>	<i>Feminine</i>	<i>Neuter</i>
N. in, int	ind', <sup>1</sup> in', int <sup>2</sup>	a n-
A. in n-, lasin n- etc.	in n-, lasin n- etc.	a n-, lassa n- etc.
G. ind', in', int <sup>2</sup>	<b>inna, na</b>	ind', in', int <sup>2</sup>
D. dond', don', dont <sup>2</sup>	dond', don',	dond', don', dont <sup>2</sup>
cossind', cossin', etc.	cossind', cossin', etc.	cossind', cossin', etc.

#### PLURAL

N. ind', in', int <sup>2</sup>	<b>inna, na</b>	<b>inna, na</b>
A. <b>inna, na</b> lasna, etc.	<b>inna, na</b> lasna, etc.	<b>inna, na</b> lasna, etc.
G. inna n-, na n-	inna n-, na n-	inna n-, na n-
D. donaib cosnaib, etc.	donaib cosnaib, etc.	donaib cosnaib, etc.

#### DUAL

N. in dá'	in dí'	in dá n-
A. in dá'	in dí'	in dá n-
G. in dá'	in dá'	in dá n-
D. don dib n-	don dib n-	don dib n-

<sup>1</sup> indicates that the form lenites.

<sup>2</sup> before ś.

Holger Pedersen, *Vergleichende Keltische Grammatik* I, 1909, 261:

“In der ersten oder einzigen proklitischen Silbe kann der Vokal im Air. nur dann geschwunden sein, wenn er anlautend ist: ... *na rúna* Wb. 8d19 ‘die Geheimnisse’ (*inna rúna*)... Diese Erscheinung ist jedoch im Air. noch verhältnismäßig selten und tritt erst im Mir. in größerem Umfang auf...”

Rudolf Thurneysen, *A Grammar of Old Irish*, 1946, 293:

“Occasionally the initial *i* of disyllabic forms is dropped in absolute anlaut also: ‘**na** beside **inna** .’”

Kim McCone, *Stair na Gaeilge*, 1992, 120:

“Bhi *inna* déshiollach sa tSean-Ghaeilge de ghnáth ach buailtear le corrsampla de *na* giorraithe fiú sna gluaiseanna (...; cf. *le, la* sa Fhraincis < L *ille, illa*).”

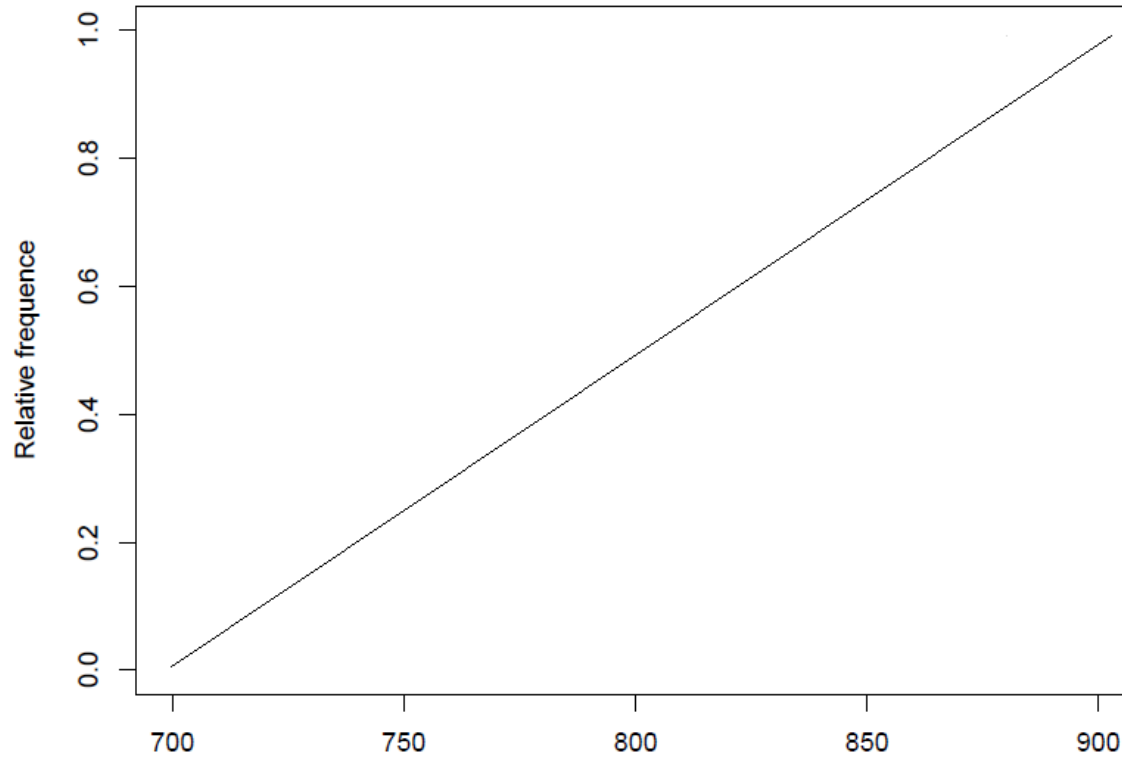
Kim McCone, ‘The Würzburg & Milan Glosses: our Earliest Sources of ‘Middle Irish’’, *Ériu* 36 (1985), 89-90:

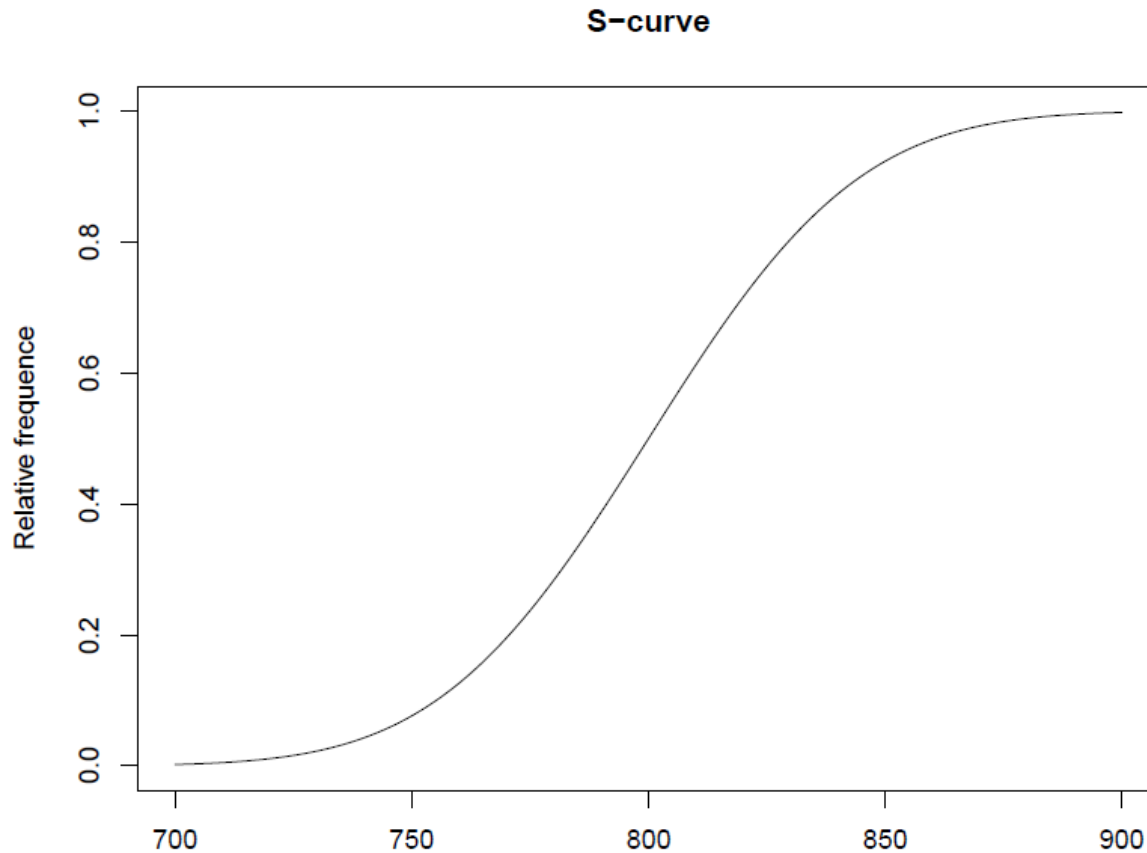
“There are a few instances of *na* for *inna* even in Wb. (12c11, 16a12, 18d14, Ml.18d24, etc.; cf. GOI 293) to show that this reduction had taken place by then, restricting *inna* to a learned register variant that was coming under increasing pressure from *na* and the preponderance of which in the Glosses is probably due at least in part to orthographic conservatism.”

James Carney, ‘The dating of Early Irish verse texts, 500–1000’, *Éigse* 19 (1982–3), 199:

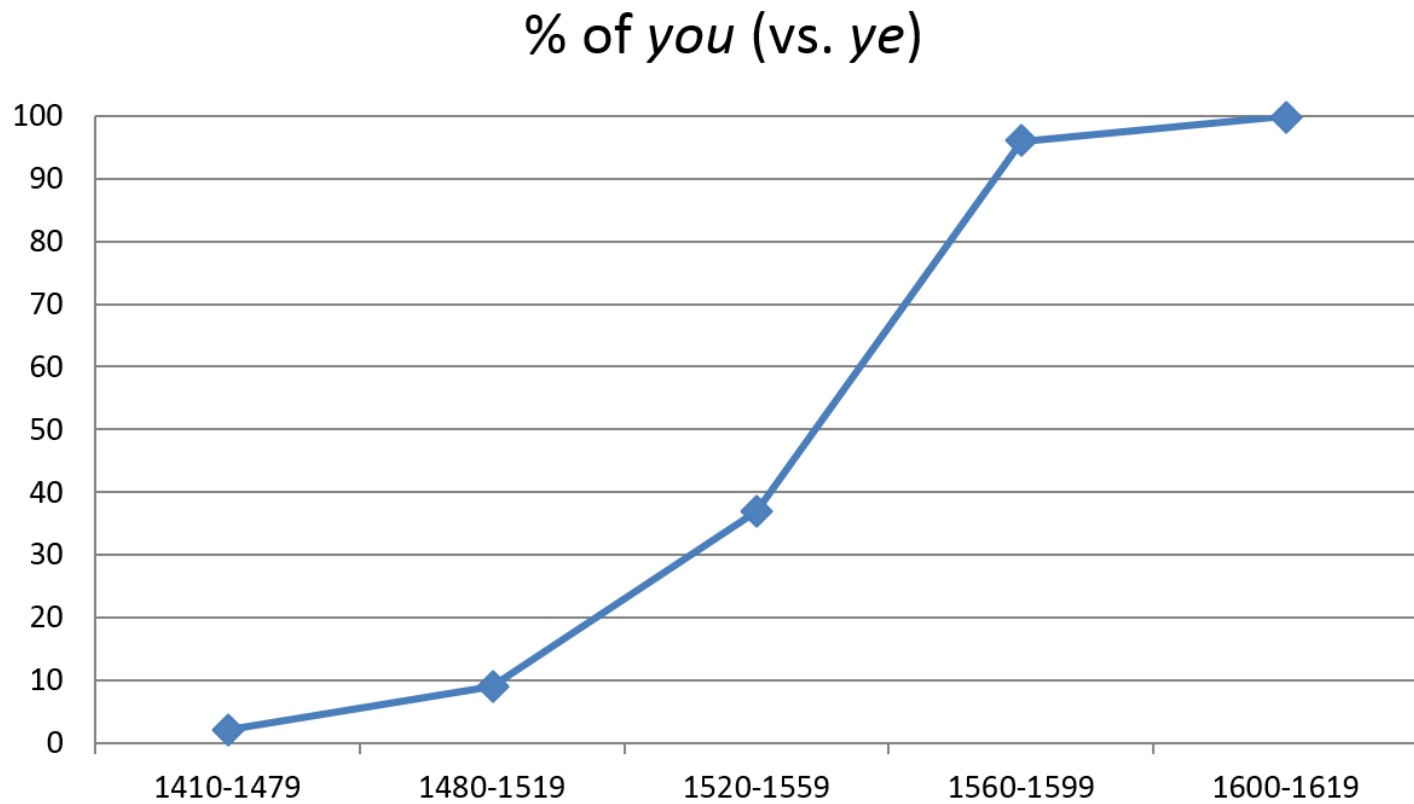
**“We have plainly here a dating criterion of some importance. The graph of *inna* falls slowly and regularly from 100 % in the seventh century to zero from about 900 onwards, a clear example of the linguistic progression that we may expect to find in our list.”**

the rise of *na* acc. to Carney 1982-3



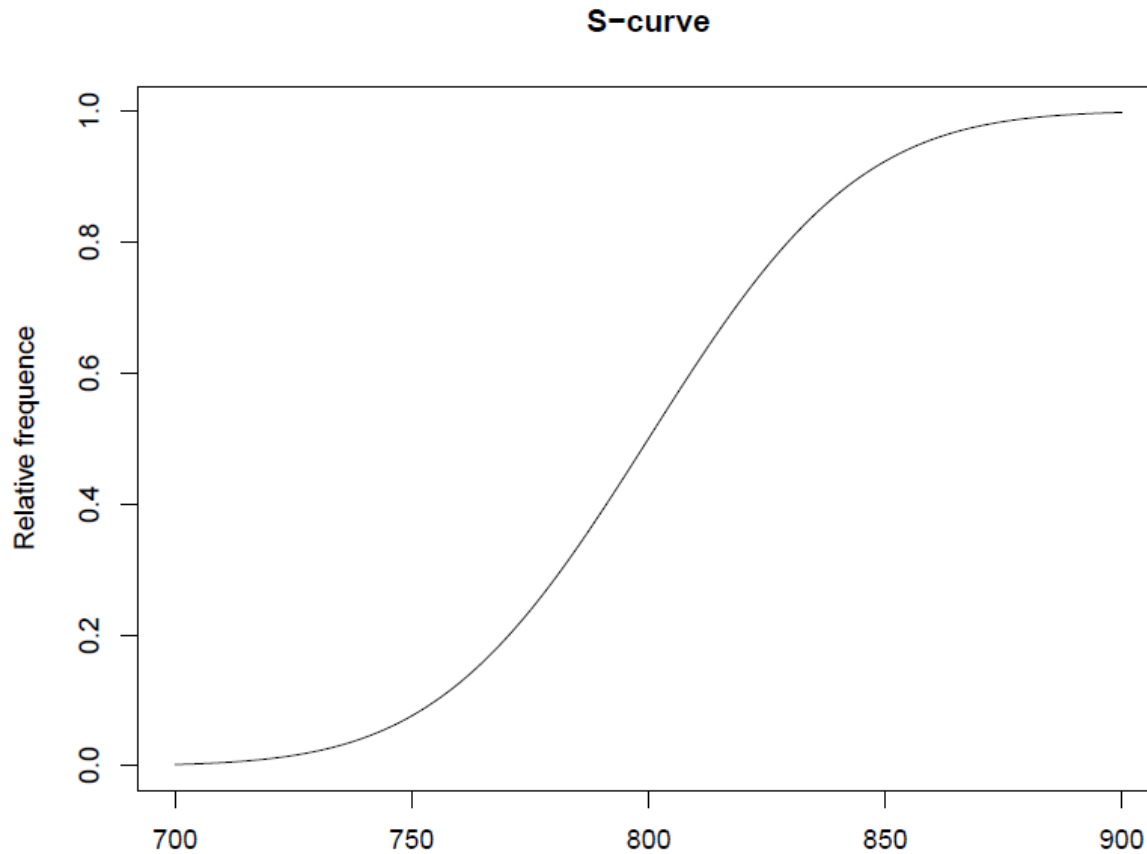


The S-curve is usually used to model the diffusion of linguistic variation



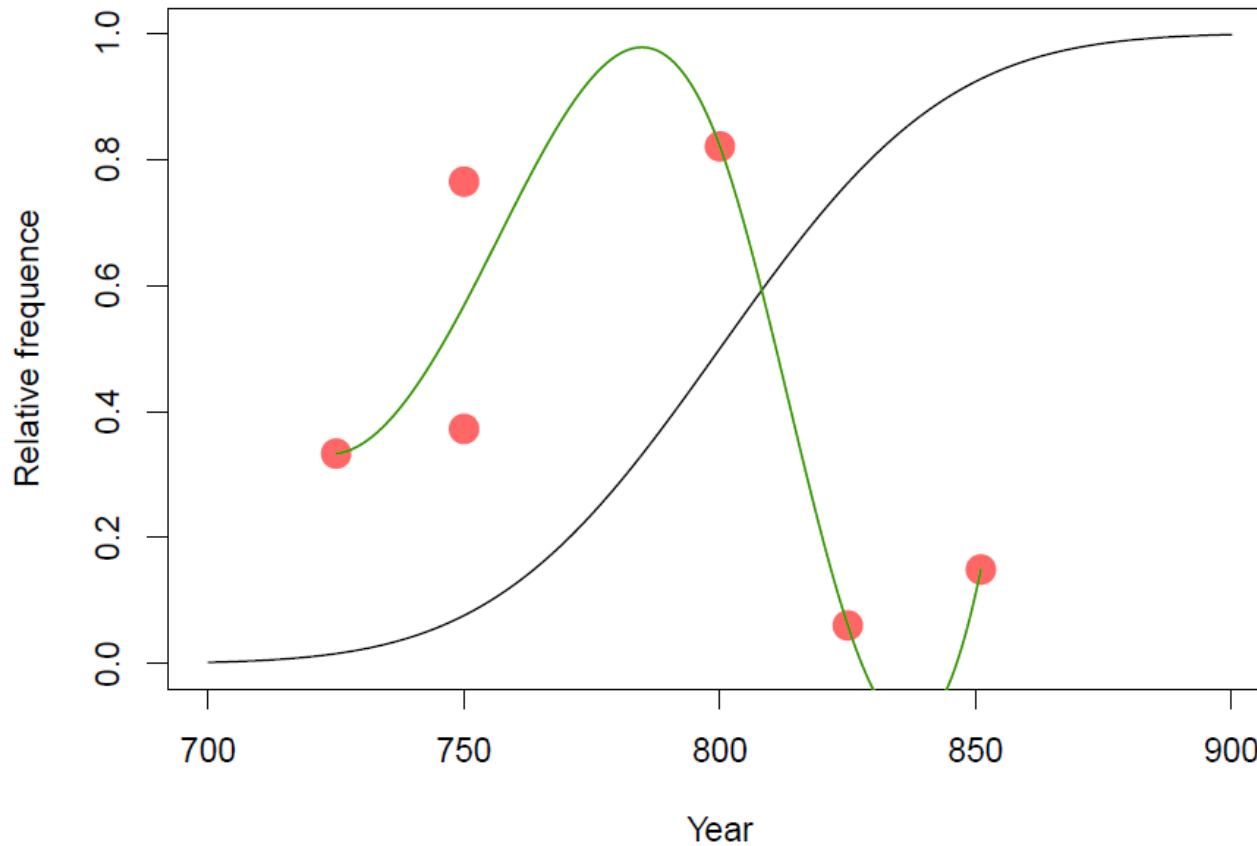
Source: <http://www.helsinki.fi/varieng/series/volumes/16/nevalainen>





adapted after Carney 1982-3

relative frequencies of *na*



# Simple Variations

- These variations can be modelled as a Bernoulli variables, where the probability of observing the younger variant is a function of time  $p(t)$ :

$$X_1 \sim \text{Bernoulli}(p(t)) \quad (1)$$

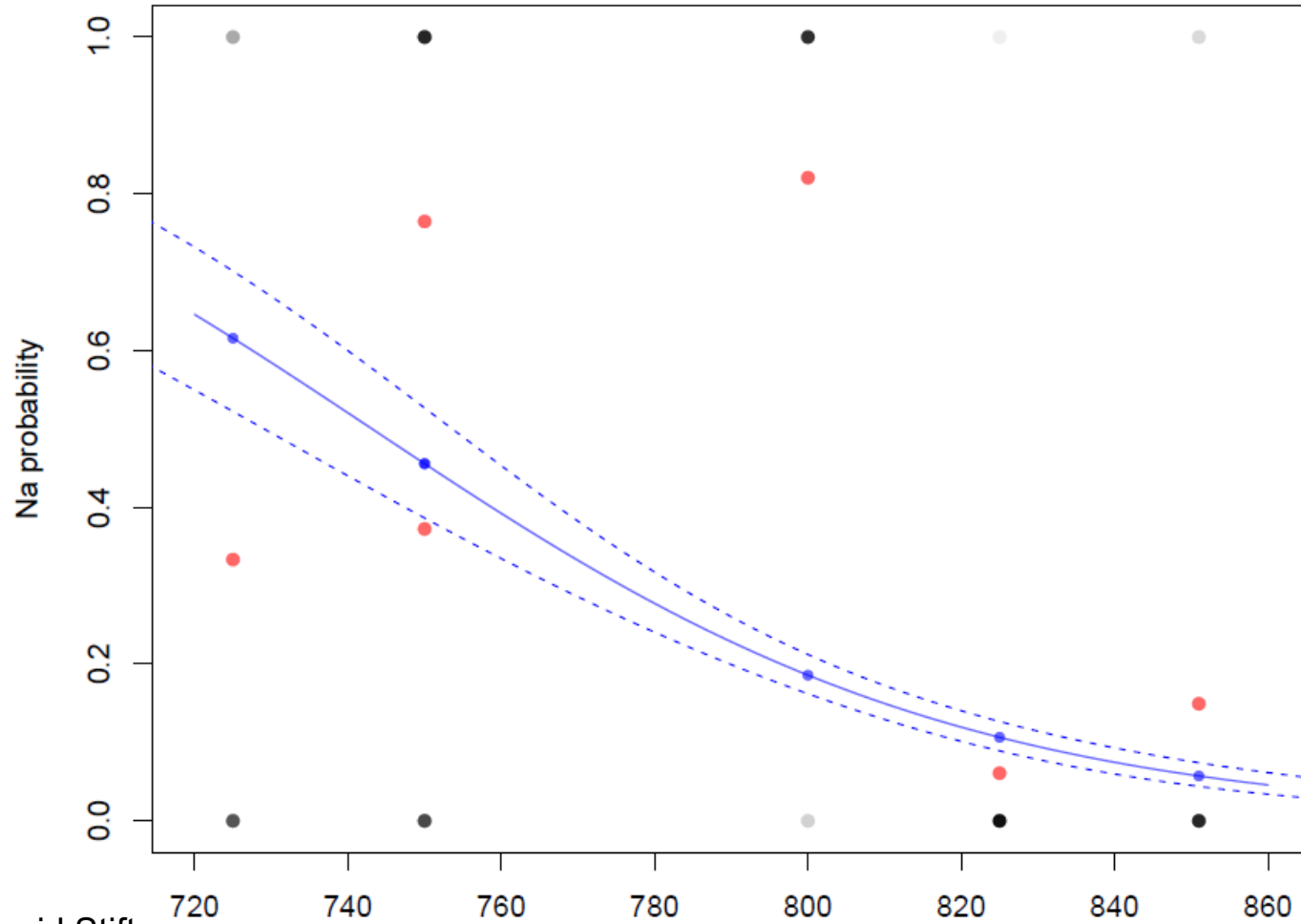
# Inference of Function $p(t)$

To infer the function  $p(t)$ , a linear logistic model can be used.

$$\log\left(\frac{p(t)}{1-p(t)}\right) = \beta_0 + \beta_1 t \quad (2)$$

$$p(t) = \frac{e^{\beta_0 + \beta_1 t}}{1 + e^{\beta_0 + \beta_1 t}} \quad (3)$$

# Logistic Regression



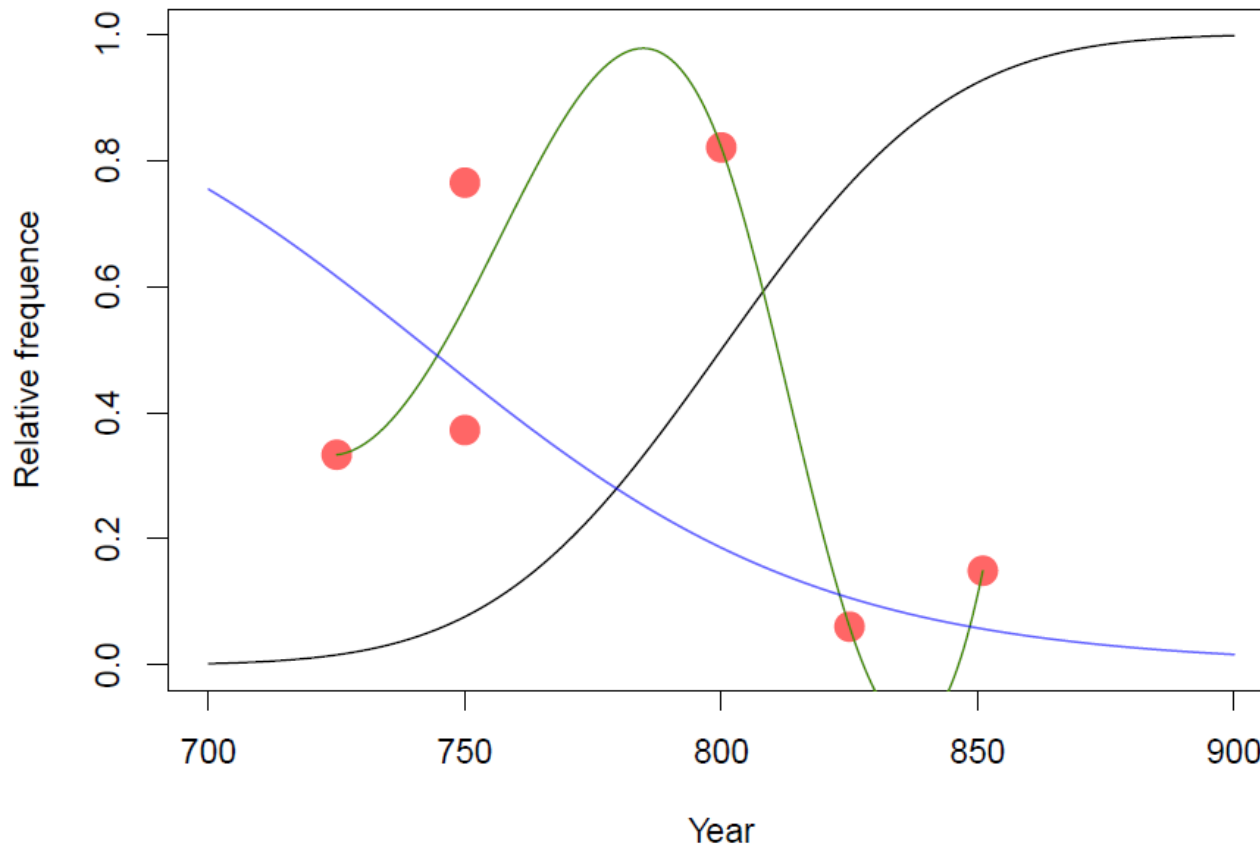
Data: David Stifter

Model: Marco A. Aquino-López

# *inna vs. na*

Label	Year	na	inna	R. frequency
Poem 01	800	32	7	82%
Poem 02	750	26	8	76.5%
Poem 03	725	1	2	33.3%
Wurzburg	750	54	91	37.2%
Milan	825	51	788	6%
St Gall	857	26	148	14.9%

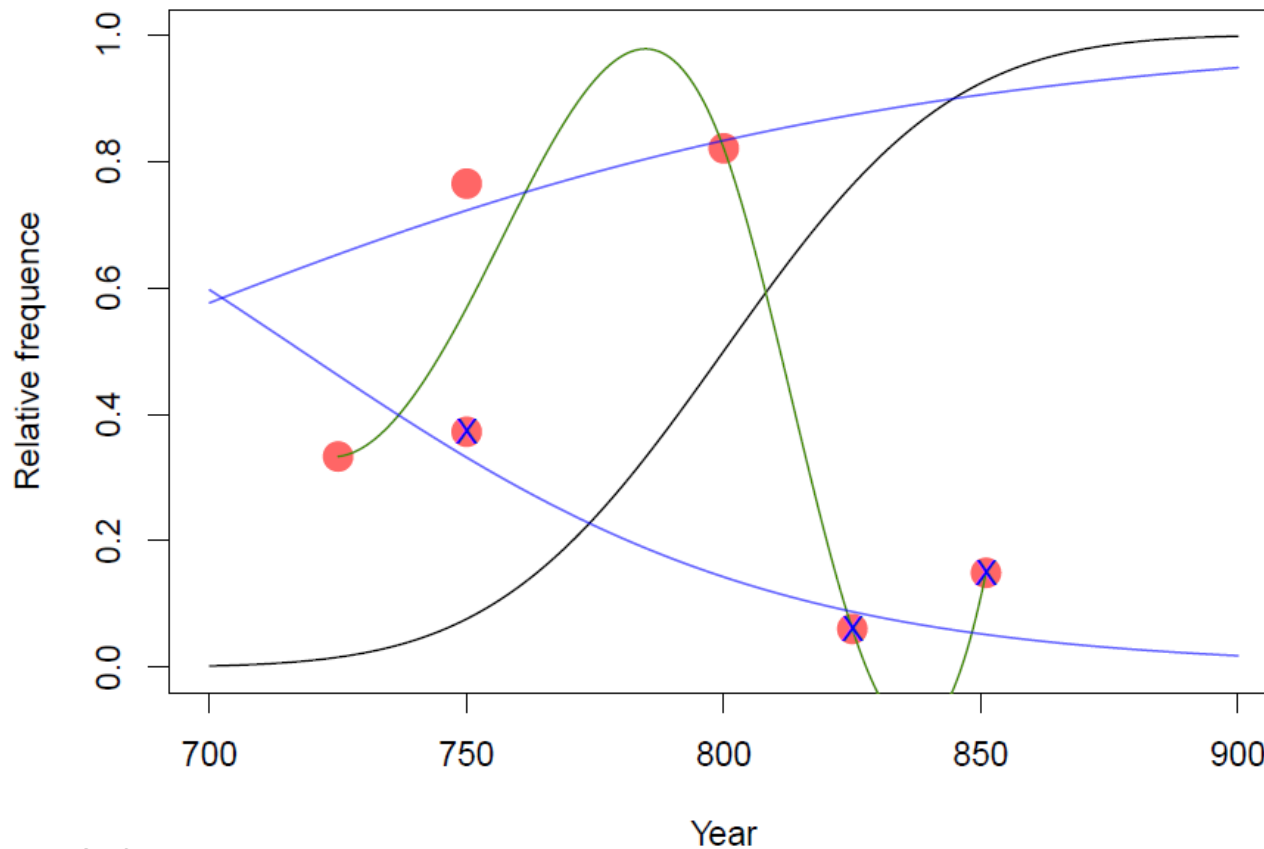
relative frequencies of *na*



Data: David Stifter

Model: Marco A. Aquino-López

relative frequencies of *na*



Data: David Stifter

Model: Marco A. Aquino-López



# inna vs. na (Details)

<i>Würzburg</i>	masc.	neut.	fem.	total
gen.sg.			<b>35</b> [22 : 13 (37%)]	<b>35</b> [22 (63%) : 13 ( <b>37%</b> )]
gen.pl.	<b>43</b> [37 : 6 (14%)]	<b>8</b> [4 : 4 (50%)]	<b>17</b> [11 : 6 (35%)]	<b>68</b> [52 (76%) : 16 ( <b>24%</b> )]
nom.pl.	<b>3</b> [2 : 1 (33%)]	<b>7</b> [2 : 5 (71%)]	<b>7</b> [3 : 4 (57%)]	<b>17</b> [7 (41%) : 10 ( <b>59%</b> )]
acc.pl.	<b>10</b> [2 : 8 (80%)]	<b>7</b> [5 : 2 (29%)]	<b>8</b> [3 : 5 (63%)]	<b>25</b> [10 (40%) : 15 ( <b>60%</b> )]
total	<b>56</b> [41 : 15 (27%)]	<b>22</b> [11 : 11 (50%)]	<b>67</b> [39 : 28 (42%)]	<b>145</b> [91 (63%) : 54 (37%)]

<i>Milan</i>	masc.	neut.	fem.	total
gen.sg.			<b>185</b> [ <b>179 : 6 (3%)</b> ]	<b>185</b> [ <b>179 : 6 (3%)</b> ]
gen.pl.	<b>207</b> [195 : 12 (6%)]	<b>36</b> [ <b>36 : 0 (0%)</b> ]	<b>84</b> [80 : 4 (5%)]	<b>327</b> [311 : 16 (5%)]
nom.pl.	<b>4</b> [4 : 0 (0%)]	<b>65</b> [ <b>59 : 6 (9%)</b> ]	<b>50</b> [47 : 3 (6%)]	<b>119</b> [110 : 9 (8%)]
acc.pl.	<b>89</b> [ <b>80 : 9 (10%)</b> ]	<b>78</b> [ <b>70 : 8 (10%)</b> ]	<b>41</b> [38 : 3 (7%)]	<b>208</b> [ <b>188 : 20 (10%)</b> ]
total	<b>300</b> [279 : 21 (7%)]	<b>179</b> [165 : 14 (8%)]	<b>360</b> [344 : 16 (4%)]	<b>839</b> [788 : 51 (6%)]

<i>St Gall</i>	masc.	neut.	fem.	total
gen.sg.			<b>46</b> [ <b>43 : 3 (7%)</b> ]	<b>46</b> [ <b>43 : 3 (7%)</b> ]
gen.pl.	<b>17</b> [ <b>16 : 1 (6%)</b> ]	<b>20</b> [ <b>18 : 2 (10%)</b> ]	<b>25</b> [ <b>19 : 6 (24%)</b> ]	<b>62</b> [53 : 9 (15%)]
nom.pl.	<b>2</b> [2 : 0 (0%)]	<b>19</b> [15 : 4 (11%)]	<b>23</b> [ <b>14 : 9 (39%)</b> ]	<b>44</b> [ <b>31 : 13 (30%)</b> ]
acc.pl.	<b>5</b> [4 : 1 (20%)]	<b>16</b> [ <b>16 : 0 (0%)</b> ]	<b>1</b> [1 : 0 (0%)]	<b>22</b> [ <b>21 : 1 (5%)</b> ]
total	<b>24</b> [22 : 2 (8%)]	<b>55</b> [49 : 6 (11%)]	<b>95</b> [77 : 18 (19%)]	<b>174</b> [148 : 26 (15%)]

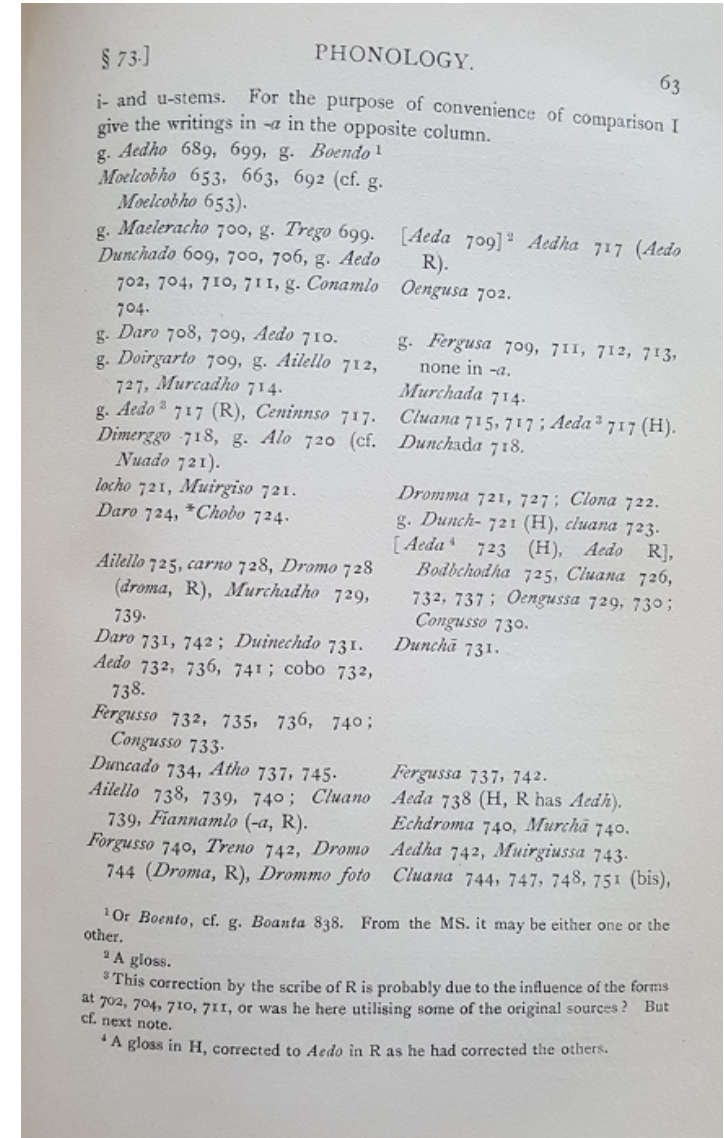
# Merger of -o and -a

gen.sg. of i- and u-stem nouns:

- Early Old Irish: -o
- Late Old Irish: -a

Tomás Ó Máille, *The Language of the Annals of Ulster*, 1910, 62-68:

- c. 830s



# Date Ranges of Texts

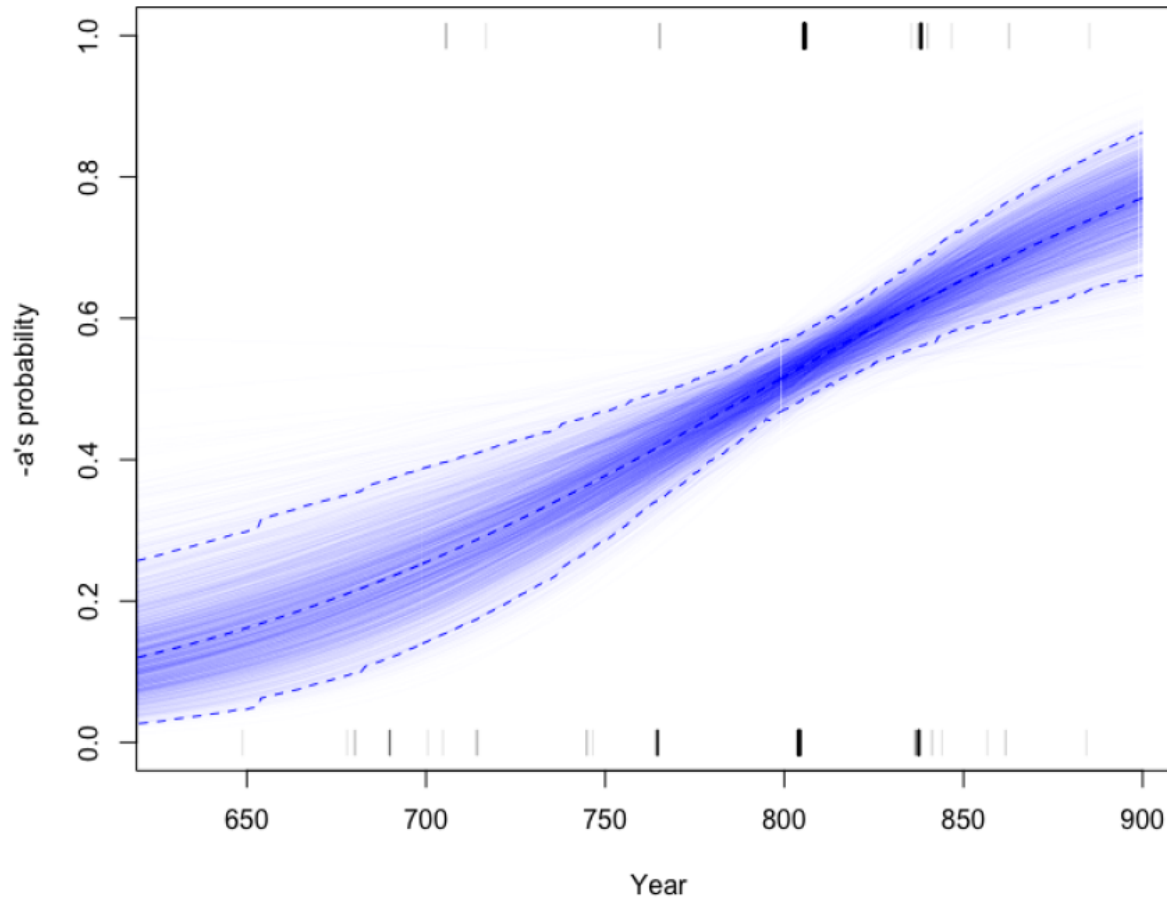
	D	E	F	G	H	I	J	K	
1	title	Ms.	Philological dating	Date Ms Low	Date Ms Upp	?Date Text/Glos	?Date Text/Glos	Weight	Note
74	Naples_Charisius	Biblioteca Nazionale, IV	MS: [VAQ: 7./8. Jh. (LINDSAY, N	800	900	800	900		The date c
75	St_Gall_Prudentius	Stiftsbibliothek Cod. Sa	MS/Glosses: [VAQ - HS: 9. Jh. (S	850	850	850	877		For ms. ar
76	Paris_De_Locis_Sanctis	Bibliothèque Nationale d	MS: [VAQ - 9. Jh. (BIELER, 30)]	800	900	800	842	800	The date c
77	Zurich_De_Locis_Sanctis	Zentralbibliothek, Rh. 73	MS: [VAQ - 1. H. 9. Jh. / „vor	800	842	800	842	800	The date c
78	Einsiedeln_Computus	Einsiedeln, Stiftsbiblioth	MS/Glosses: [VAQ - HS: „IX. Jh.,	867	900	689	719		For dating
79	Bern_Donatus	Burgerbibliothek, Cod. 2	MS: [VAQ - Ende 8. Jh. (LOWE,	775	800	685	715	715	The ms. is
80	Würzburg_Glosses (prima	Universitätsbibliothek, M	MS: [VAQ - HS: „nach der Mitte	750	800	685	715	715	For dating
81	Würzburg_Glosses (main	Universitätsbibliothek, M	MS: [VAQ - HS: „nach der Mitte	750	800	750	770		The upper

# Bayesian Logistic Regression

- Provides probabilities
- Allows to model the uncertainty in the year of origin of the variation

$$\begin{aligned} \log\left(\frac{p(t_i)}{1 - p(t_i)}\right) &= \beta_0 + \beta_1 t_i, \\ t_i &= \mathcal{U}(l_l, l_u), \\ \text{Priors} \\ \beta_0 &\sim \mathcal{N}(0, .01), \\ \beta_1 &\sim \mathcal{N}(0, .01). \end{aligned}$$

# Merger of -o and -a



Data: Fangzhe Qiu  
Model: Marco A. Aquino-López

# Demonstrative *-so/-se*

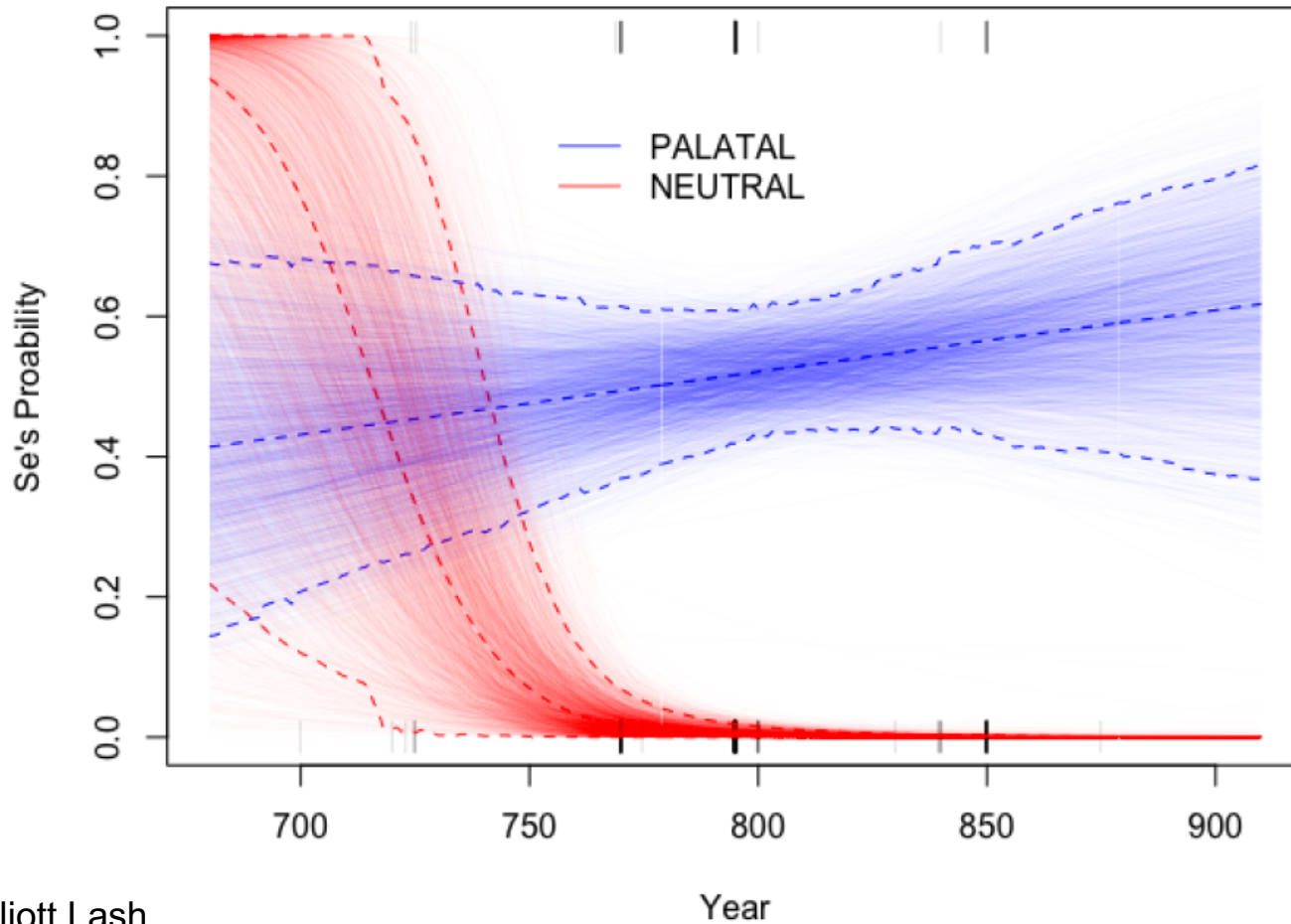
*Grammar of Old Irish*, 300:

- ‘...the enclitic particles **so** and **sa** [...], after palatal auslaut usually **se**, **seo**, and **sea**...’

Legend:

- **blue:** after ‘palatal’ sounds (= palatalised/slender consonants, front vowels)
- **red:** after ‘non-palatal’ sounds (= non-palatalised/broad consonants, back vowels)

# Demonstrative *-so/-se*

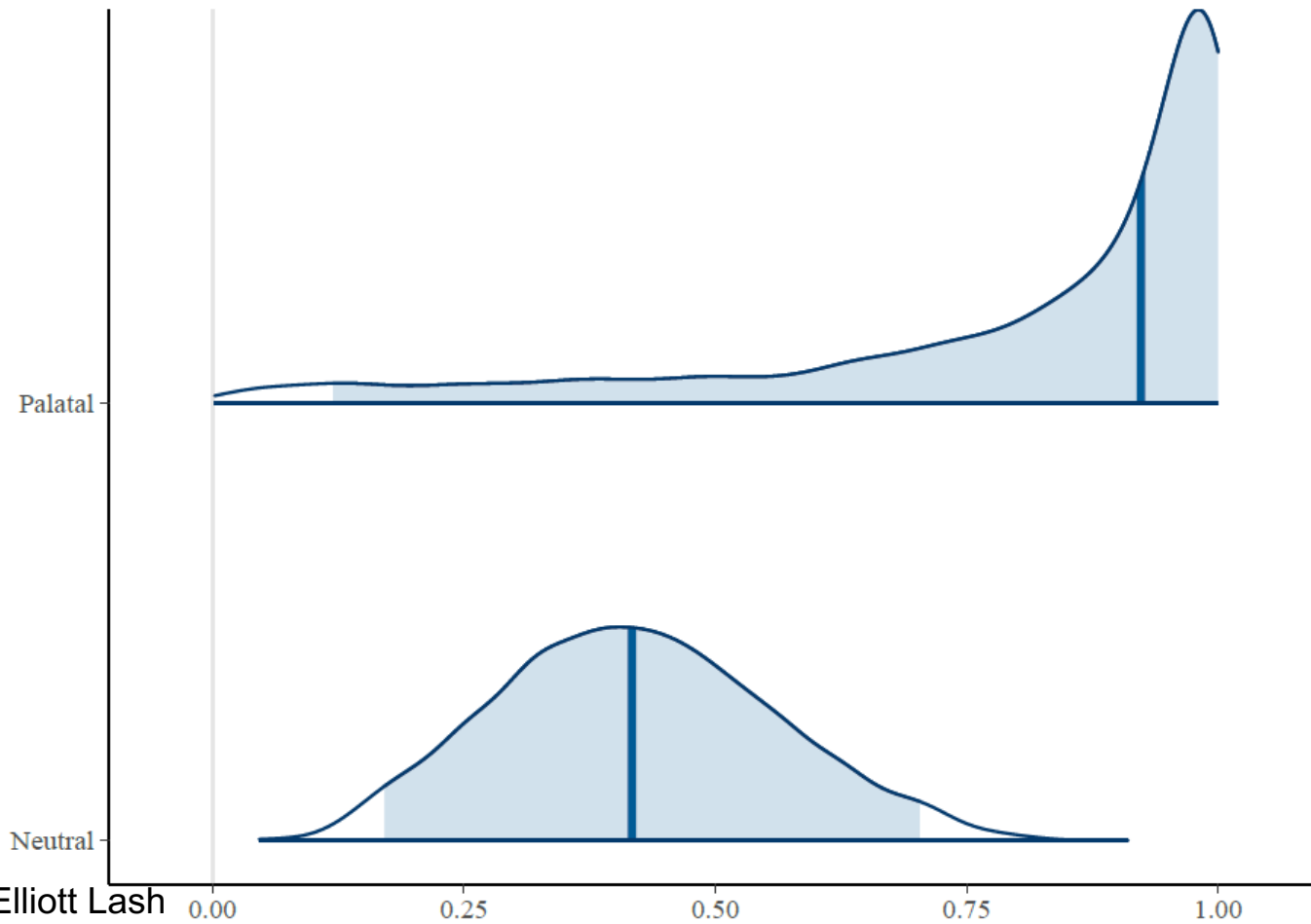


Data: Elliott Lash

Model: Marco A. Aquino-López

# Demonstrative *-sol/-se*

Posterior distributions  
with medians and 95% intervals



Data: Elliott Lash

Model: Marco A. Aquino-López



# Interim Linguistic Results

- Old Irish shows more variation than usually believed (no ‘standard’ language)
- more micro-variation between texts
- development can be modelled along typological expectations
- more precision about the developments over times
- greater insights into how the language functions



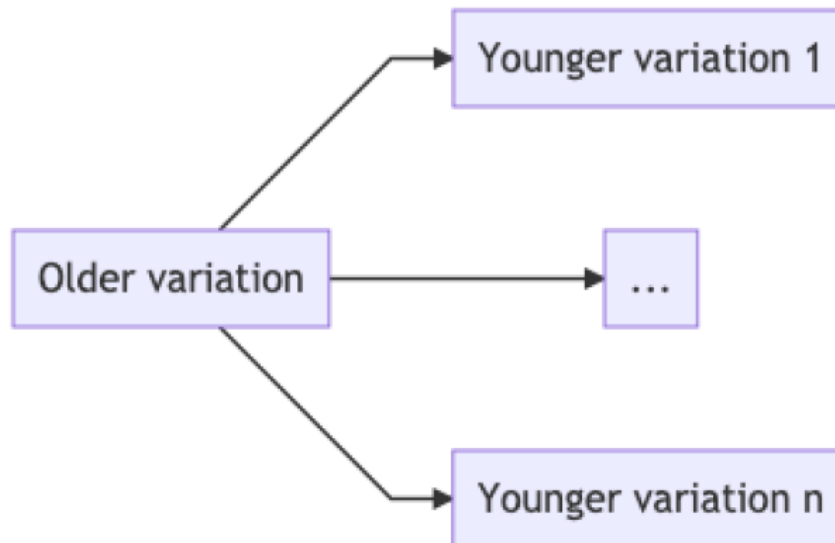
**Ad·tlochammar  
indithim dúib!**

# Multinomial Regression

multiple variations (multinomial regression)

$$Y_i \sim \text{Multinomial}(Y_i \mid \pi_{ij}),$$
$$p(t_{ij}) = \frac{\exp(t_i \beta_j)}{\sum_{k=1}^J \exp(t_i \beta_k)},$$

for  $j = 1, \dots, J - 1$ . (4)



# 1<sup>st</sup> Singular Ending

**Verb type 3a**

**Verb type 3b**

- 0 -imm
- 1 -u
- 2 <-u->
- 3 -ur

