Bayesian processing of multilingual documents

Keith Briggs

Keith.Briggs@bt.com

research.btexact.com/teralab/keithbriggs.html

Cavendish Inference Group 2005 Feb 15 1145

cam-2005feb15.tex typeset 2005 February 14 11:04 in pdfLaTeX on a linux system

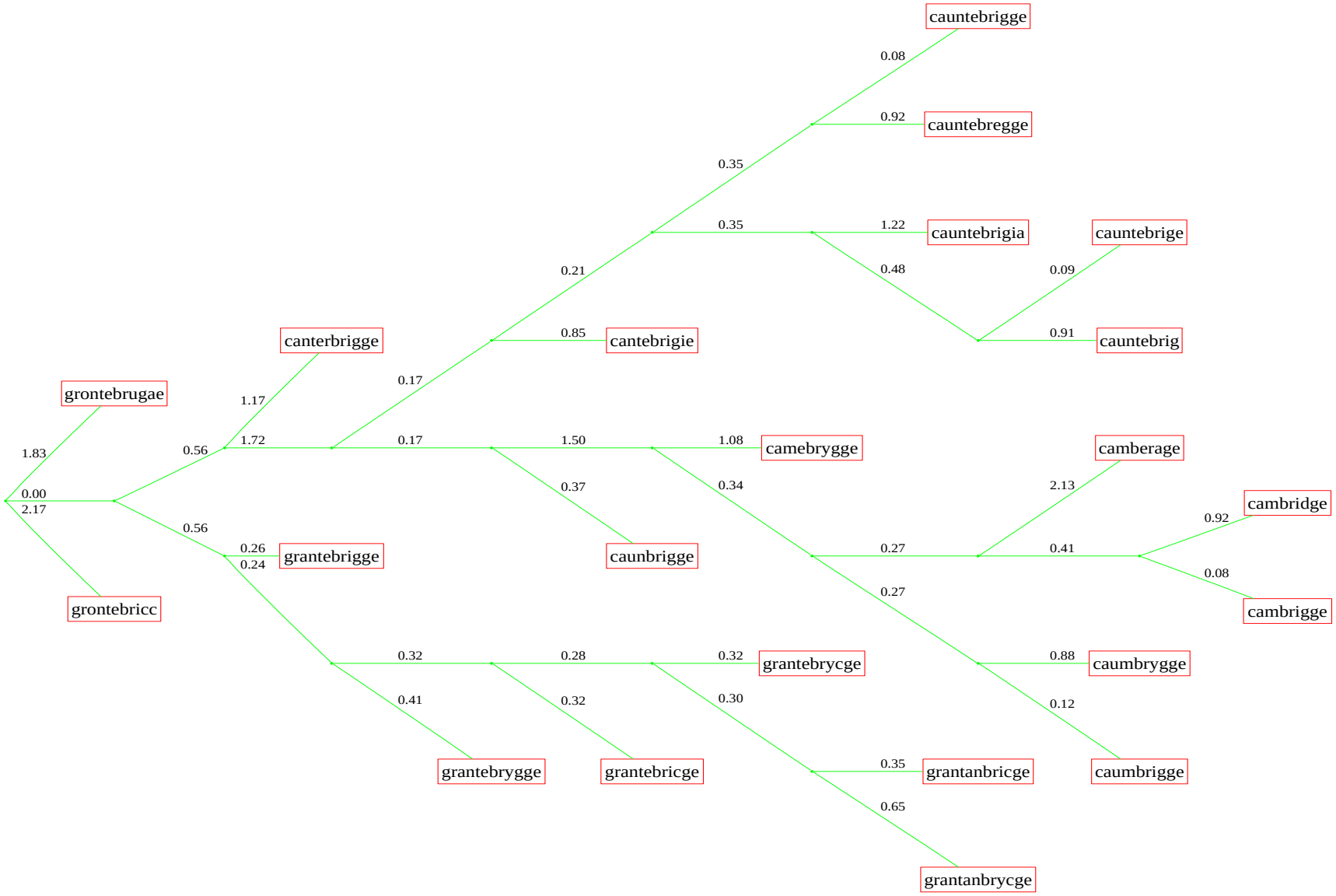
Outline

- The Germanic Lexicon Project
- some problems in multilingual text processing
- Bayesian ideas
- some preliminary 'solutions'

The aim:

to determine how well probabilistic methods work

History of 'Cambridge'



The Germanic Lexicon Project

- www.ling.upenn.edu/~kurisuto/germanic/language_resources.html
- The goal of this project is to create comprehensive online coverage of the lexicons of the early Germanic languages
- All of the data is available free of charge and free of copyright or other intellectual property encumbrance
  - *Wörterbuch der Indogermanischen Sprachen: Dritter Teil: Wortschatz der Germanischen Spracheinheit by Falk, Fick, and Torp (1909)*
  - *Grammar of the Gothic Language by Joseph Wright (1909)*
  - *An Anglo-Saxon dictionary, based on the manuscript collections of the late Joseph Bosworth; edited and enlarged by T. Northcote Toller*
  - *Altsächsiche Grammatik (Old Saxon Grammar) by Johan H. Gallée (1910)*
  - *An Icelandic-English dictionary by Richard Cleasby and Gudbrand Vigfusson (1874)*

Example page

hrægl and ðæt tôhlutan *diviserunt sibi vestimenta mea et super vestem meam miserunt sortem*, Ps. Th. 21, 16. Hié (*the apostles*) ðysne middangeard on twelf tânum tôhluton, and æghwylc ânra heora in ðæm dæle ðe hê mid tân geeode manige þeóde ûrum Drihtne gestreónde, Blickl. Homl. 121, 8.

tôh-lîc; *adj. Tough, tenacious.* v. next word.

tôhlîce; *adv. Toughly, tenaciously*:—Tôhlîce, thôlîcae, tôchtlîcae *uscide, viscide* (viscide *fortiter*, Migne), Txts. 107, 2170. Tôlîce *huscide*, 69, 1033.

tô-hlîdan; *p.* -hlâd, *pl.* -hlidon; *pp.* -hliden *To yawn, gape, open, crack* (intrans.), *split* (intrans.) *asunder*:—Tôhlâd seó eorþe *terra dissiluit*, Ors. 3, 3; Swt. 102, 26. Tôhlâd seó eorþe and wæs byrnende fŷr up of ðære eorþan *flamma scisso terrae hiatu eructata*, 4, 2; Swt. 160, 24: Lchdm. iii. 428, 3. Se beorg tôhlâd eorðscræf egeslîc *the hill yawned, an awful cave it grew*, Andr. Kmbl. 3173; An. 1589. Heofonas tôhlidon, Blickl. Homl. 105, 13. Tôhlîdan *dehiscere*, Germ. 400, 482. Biþ ðæt heáfod tôhliden *the head shall be cloven*, Soul Kmbl. 213; Seel. 109. Hié gesâwon swelce se hefon wære tôhliden *coelum scindi velut*

OCR on scan of example page

hraegl and ðaet tShlutan ***diviserunt sibi vestimento mea eí super vestem mearn miserunt soríem, Ps.*** Th. 21, 16. Hié (***the apostles***) thysne middanyeard on twelf tánum tóhluton, and áeghwylcánra heora in thaern dáele the hé niid tán geeode manige horneóde úrum Drihtne gestreónde. Blickl. Homl. 121, 8.

tóh-líc; *adj. Tough, tenacious,* v. next word.

tohlice; *adv. Tougkly, tenaciously* :– Tóhlíce, thðlícae, thðchtlícae ***uscide, viscide (viscídefortiter,*** Migne), Txts. 107, 2170. Tthlíce ***huscide,*** 69, 1033.

tó-hlídan; *p.* -hlád, *pl.* -hlidon; *pp.* -hliden *To yawn, gape, open, crack* (intrans.), *split* (intrans.) *asunder:* – Tóhlád seó eor horne ***terra dissi-luii.*** Ors. 3, 3 ; Swt. 102, 26. Tóhlád seó eor horne and waes bymende f
'yr up of thaere eorpan? ***amma scisso lerrae hiatu eructata,*** 4, 2 ; Swt. 160, 24: Lchdm. iii. 428, 3. Se beorg tthhlád eorthscraef egeslíc ***the hill yawned, an awful cave ii grew,*** Andr. Kmbl. 3173; An. 1589. Heofonas tóhlidon. Blickl. Homl. 105, 13. Tóhltdan ***dehiscere,*** Germ. 400, 482. Dip thaet heáfod tóhliden ***the head shall be cloven,*** Soul Kmbl. 213 ; Seel. 109.

The issues raised

- can we use Bayesian methods to make probabilistic corrections?
- can we identify the language of a particular word or phrase?
- can we detect OCR errors?
- can we usefully make the automatic correction?

Language recognition

- is amazingly easy:
  - *Zeichen*
  - *Teich*
  - *étang*
  - *raftan*
  - *stagnum*
  - *piccolo*
  - *ddydd*
  - *æftercweðan*
  - *riðja*
  - *négy*
- . . . but what information are we using when we do this?
- and how well can we do it when there are errors?

Text classification theory

- could be based on various choices of *features*: words, or n-grams
- corpora $C_1, C_2, \ldots, C_k$
- priors $\pi_1, \pi_2, \ldots, \pi_k$
- models $\mathbb{P}_{C_1}, \mathbb{P}_{C_2}, \ldots, \mathbb{P}_{C_k}$
- if x is an unknown document, the posterior probability that x belongs to C_j is $P(C_j|x) \propto \mathbb{P}_{C_j}\, \pi_j$
- decision rule: choose j to maximize $P(C_j|x)$

Digram measure

- word $w = w_1 w_2 \ldots w_k$
- reference measure $R_C(w) \equiv p_C(\wedge, w_1) p_C(w_1, w_2) \ldots p_C(w_k, \$)$
  - *this is naïve - it assumes adjacent digrams are statistically independent*
- Dirichlet digram measure $p_C(u, v) = \frac{\#(v|u) + \alpha\mu(v)}{\sum_r \#(r|u) + \alpha}$
- α is a hyperparameter, and the optimum α should be chosen from tests on various corpora

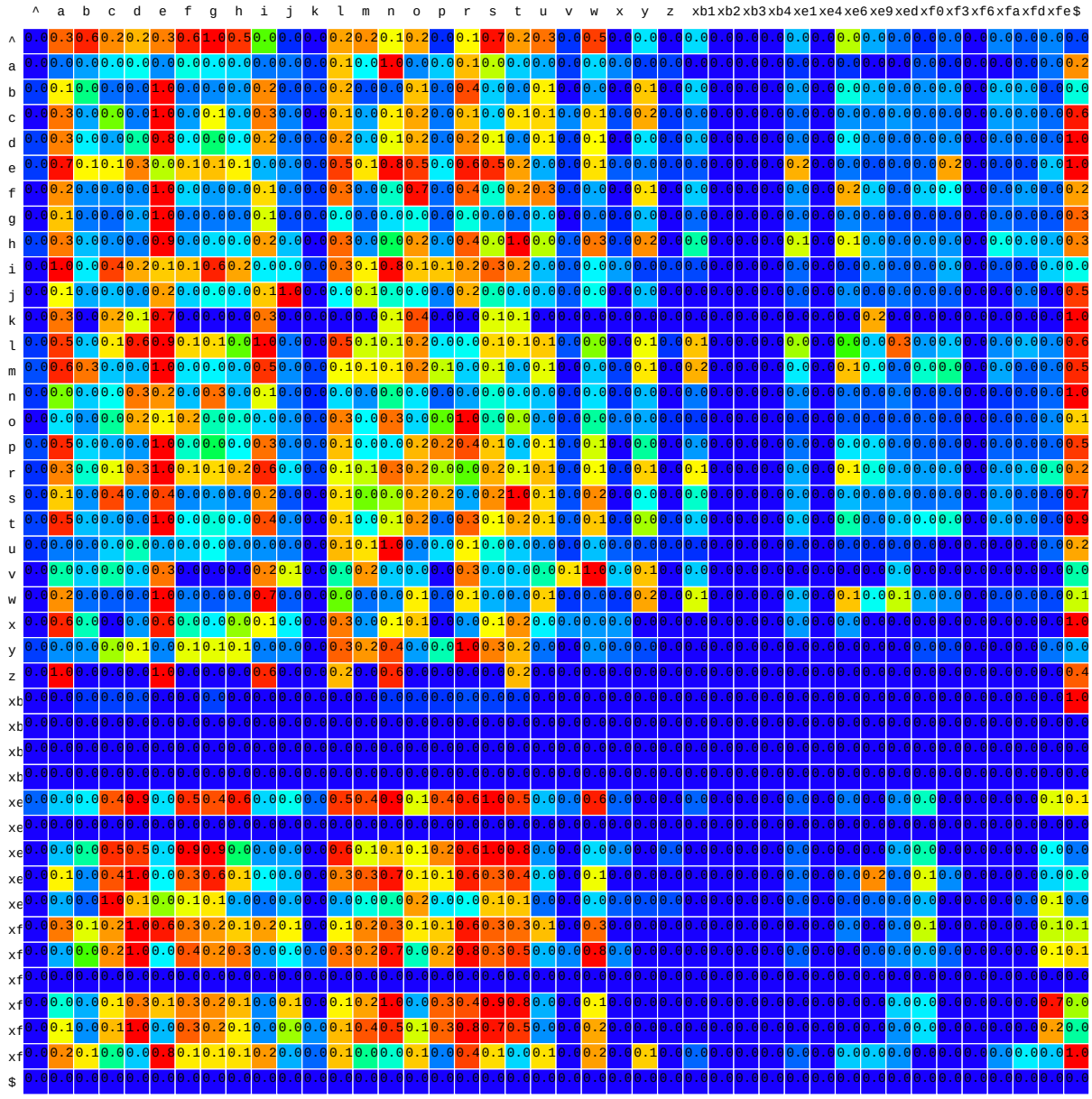
Coding issues

- Only two fixed-width choices - ASCII (1 byte) or Unicode (2 bytes)
- TEX or html are possible, but are not fixed-width
- Unfortunately, ASCII cannot do all characters used in OE or Icelandic
- Therefore, I moved some characters to unneeded ascii positions
  - *e.g. hex* $b1$ *(really the* $\pm$ *sign) for* $\overline{\text{œ}}$

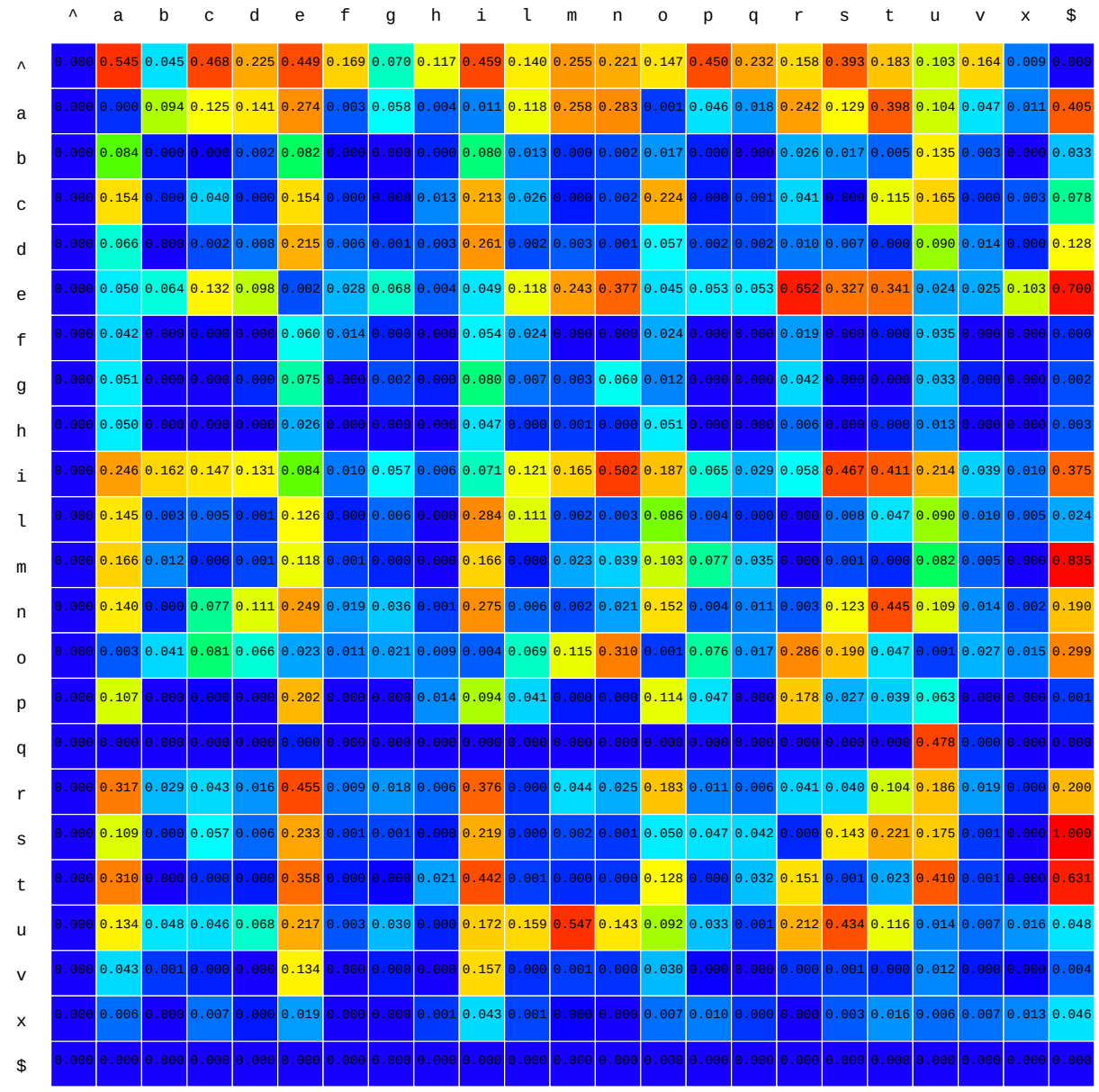
Training

- Collect texts
- split into words; check for obvious errors; fix punctuation and capitalization
- Count trigrams and estimate α

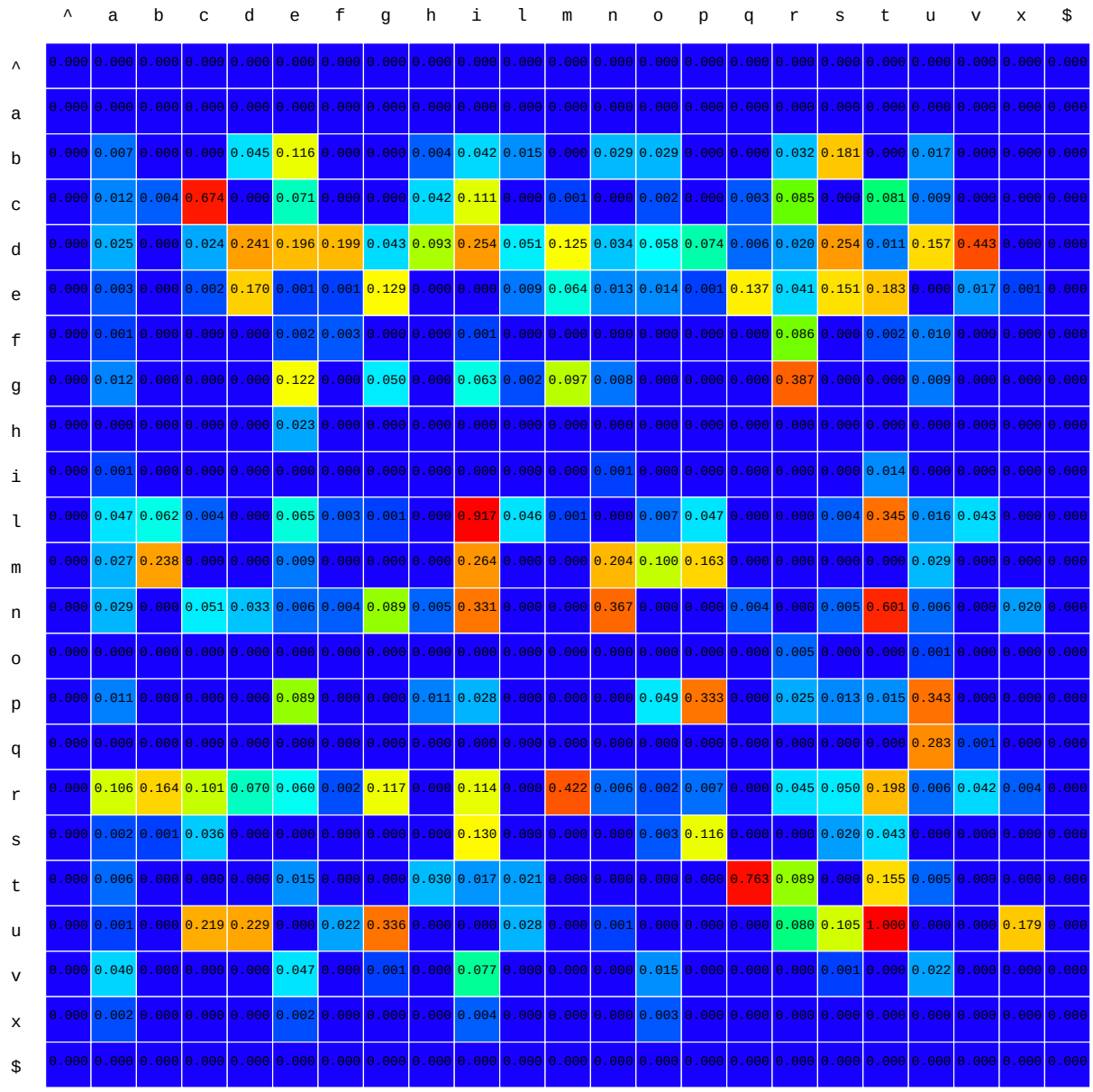
Example digram measure for Old English



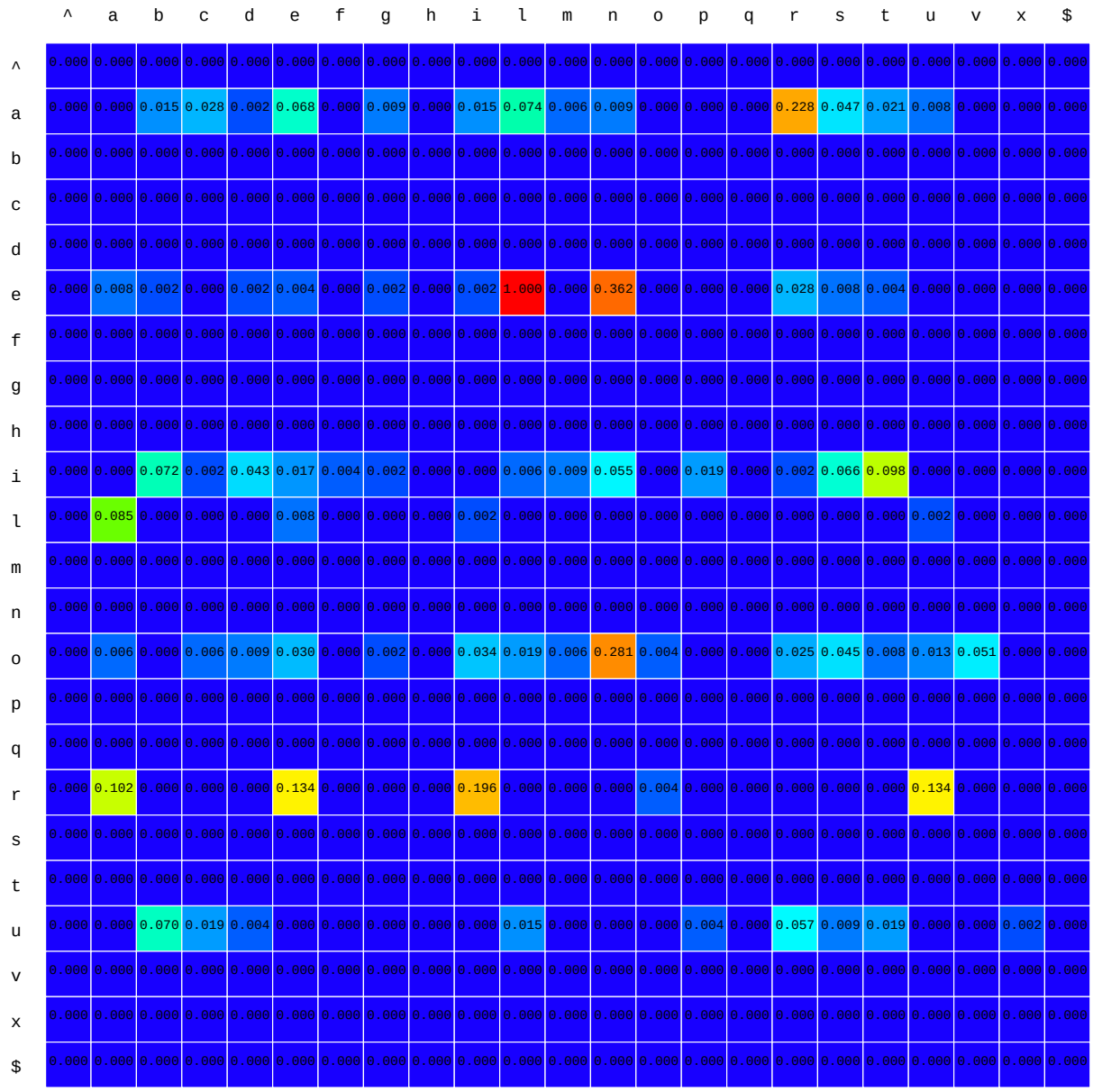
Example digram measure for Latin



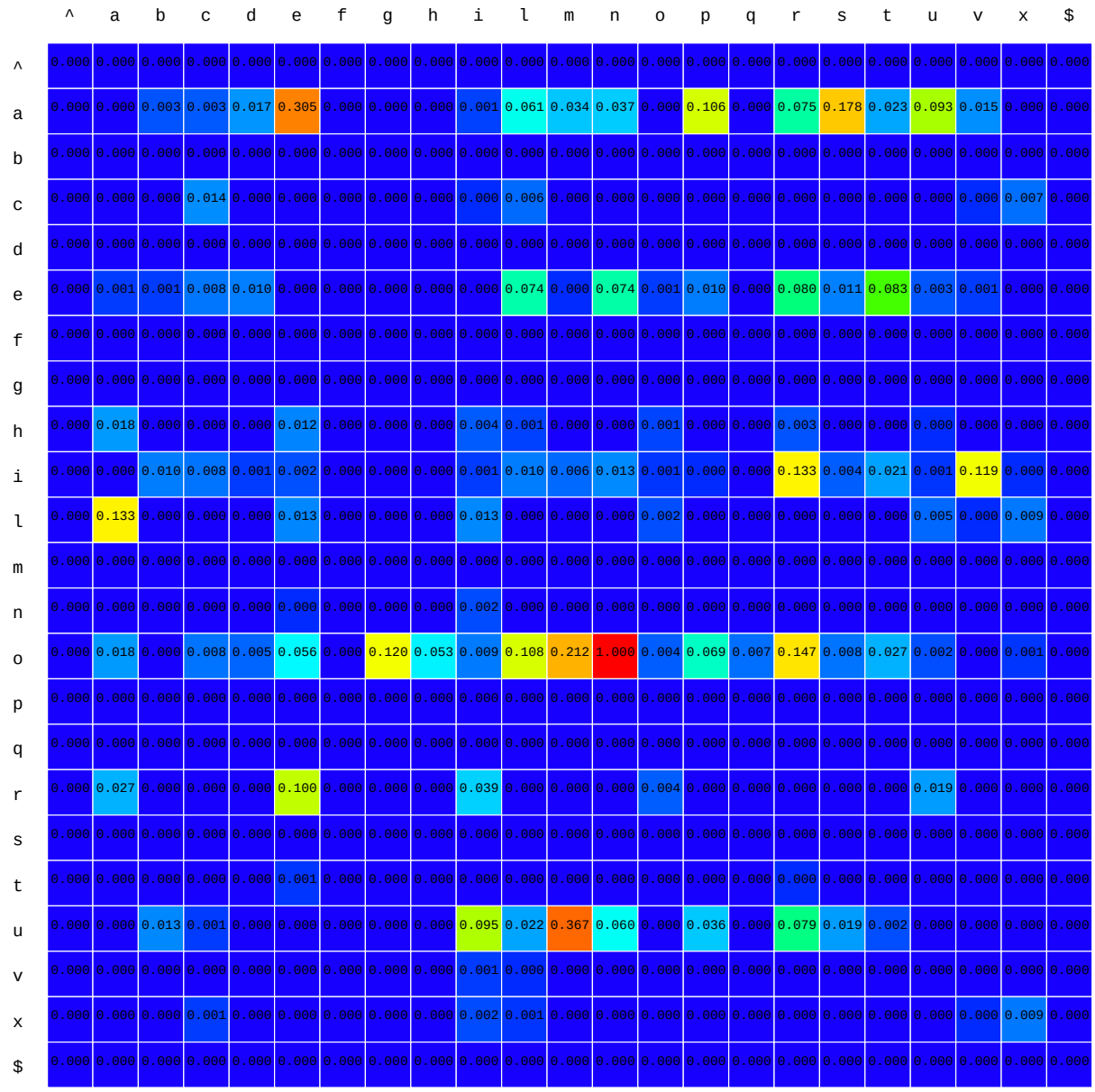
Latin trigrams - a..



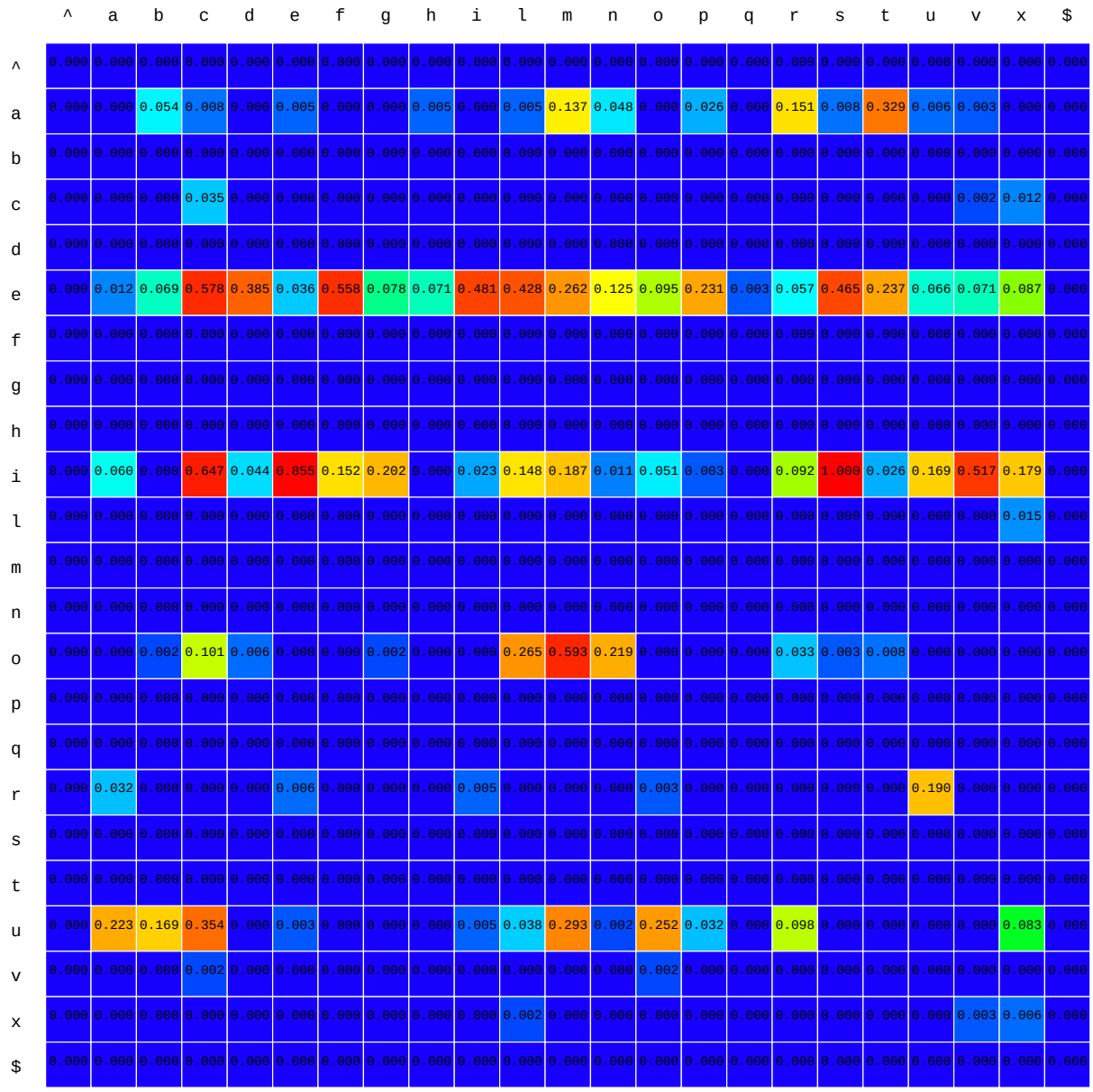
Latin trigrams - b..



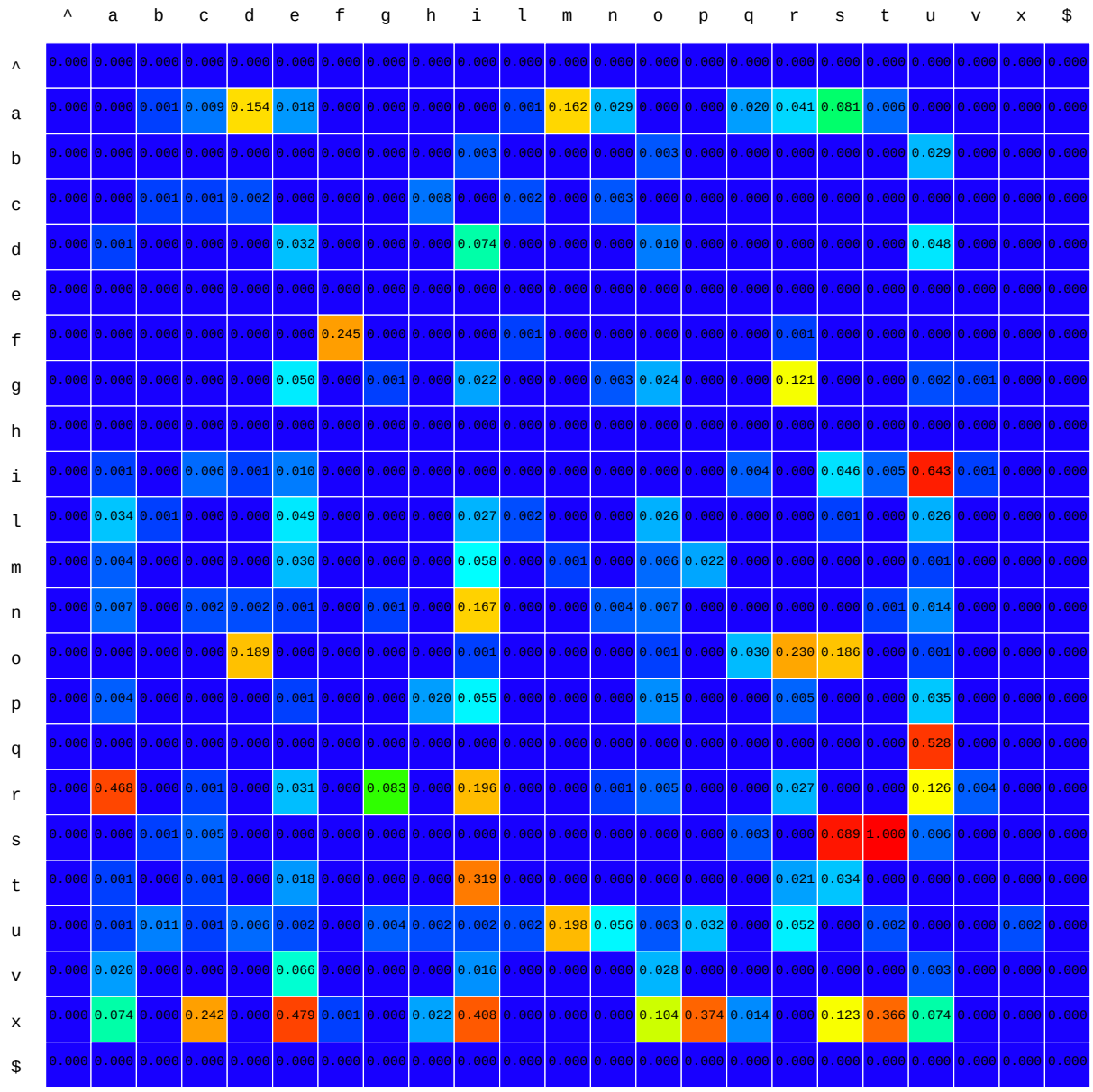
Latin trigrams - c..



Latin trigrams - d..



Latin trigrams - e..



Latin verbs

Fourth conjugation, indicative voice, active mood

PRESENT	**audio**	**audis**	**audit**	**audimus**	**auditis**	**audiunt**
PREFECT	**audivi**	**audivisti**	**audivit**	**audivimus**	**audivistis**	**audiverunt**
IMPERFECT	**audiebam**	**audiebas**	**audiebat**	**audiebamus**	**audiebatis**	**audiebant**
PLUPERFECT	**audiveram**	**audiveras**	**audiverat**	**audiveramus**	**audiveratis**	**audiverant**
FUTURE	**audiam**	**audies**	**audiet**	**audiemus**	**audietis**	**audient**
FUTURE PERFECT	**audivero**	**audiveris**	**audiverit**	**audiverimus**	**audiveritis**	**audiverint**

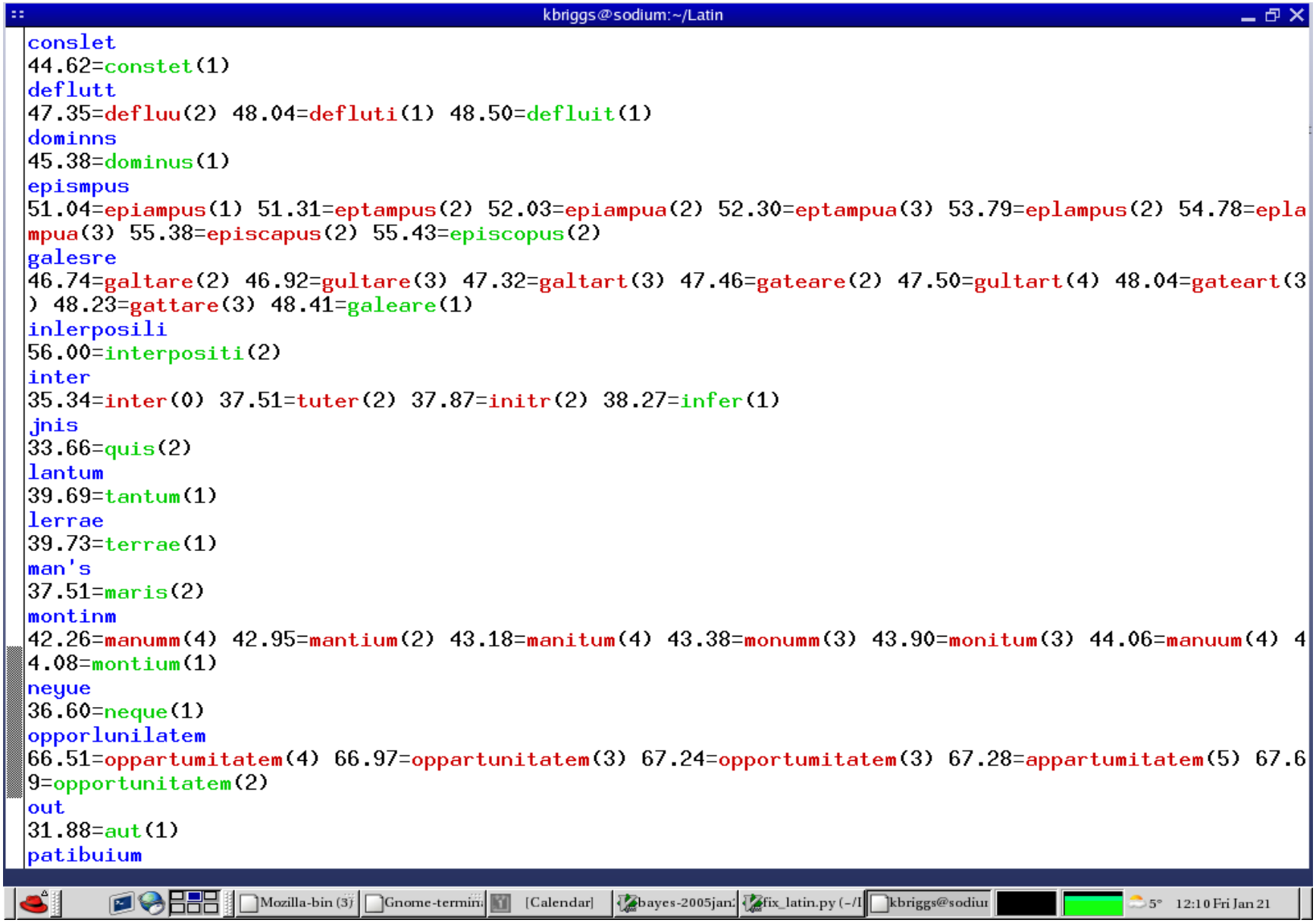
subjunctive

PRESENT	**audiam**	**audias**	**audiat**	**audiamus**	**audiatis**	**audiant**
PERFECT	**audiverim**	**audiveris**	**audiverit**	**audiverimus**	**audiveritis**	**audiverint**
IMPERFECT	**audirem**	**audires**	**audiret**	**audiremus**	**audiretis**	**audirent**
PLUPERFECT	**audivissem**	**audivisses**	**audivisset**	**audivissemus**	**audivissetis**	**audivissent**

Spelling correction

- Idea: keep a list of common errors (perhaps with priors)
- Try all corrections and sort them by likelihood
- Give the users a list of the few most likely to select from
- Could use heuristics: likelihood 'jumps'

Screenshot


kbriggs@sodium:~/Latin
conslet
44.62=constet(1)
deflutt
47.35=defluu(2) 48.04=defluti(1) 48.50=defluit(1)
dominns
45.38=dominus(1)
epismpus
51.04=epiampus(1) 51.31=eptampus(2) 52.03=epiampua(2) 52.30=eptampua(3) 53.79=eplampus(2) 54.78=epla
mpua(3) 55.38=episcapus(2) 55.43=episcopus(2)
galesre
46.74=galtare(2) 46.92=gultare(3) 47.32=galtart(3) 47.46=gateare(2) 47.50=gultart(4) 48.04=gateart(3
) 48.23=gattare(3) 48.41=galeare(1)
inlerposili
56.00=interpositi(2)
inter
35.34=inter(0) 37.51=tuter(2) 37.87=initr(2) 38.27=infer(1)
jnis
33.66=quis(2)
lantum
39.69=tantum(1)
lerrae
39.73=terrae(1)
man's
37.51=maris(2)
montinm
42.26=manumm(4) 42.95=mantium(2) 43.18=manitum(4) 43.38=monumm(3) 43.90=monitum(3) 44.06=manuum(4) 4
4.08=montium(1)
neyue
36.60=neque(1)
opporlunilatem
66.51=oppartumitatem(4) 66.97=oppartunitatem(3) 67.24=opportumitatem(3) 67.28=appartumitatem(5) 67.6
9=opportunitatem(2)
out
31.88=aut(1)
patibuium
Mozilla-bin (3)
Gnome-termini
[Calendar]
bayes-2005jan:
fix_latin.py (~/l
kbriggs@sodiu
5°
12:10 Fri Jan 21