

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 PDF \LaTeX ON A LINUX SYSTEM

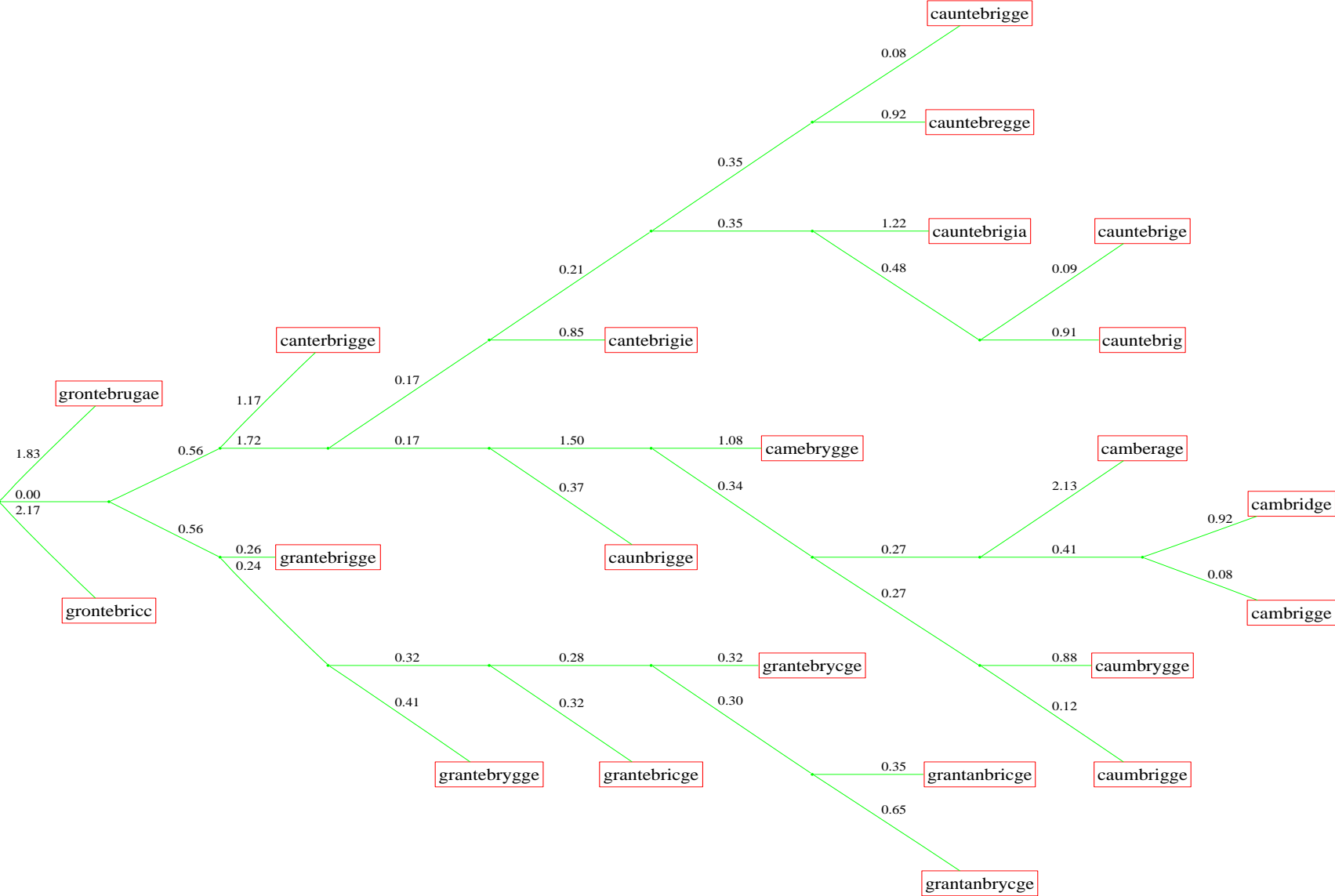
Outline

- ★ 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* ■
 - ▷ *Altsächsische Grammatik (Old Saxon Grammar)* by Johan H. Gallée (1910) ■
 - ▷ *An Icelandic-English dictionary* by Richard Cleasby and Gudbrand Vigfusson (1874) ■

hrægl and ðæt tōhlutan *diviserunt sibi vestimenta mea et super vestem meam miserunt sortem*, Ps. Th. 21, 16. Hié (*the apostles*) ðysne middangeard on twelf tānum tōhluton, and æghwylc ānra heora in ðæm dæle ðe hē mid tān geeode manige þeode ūrum Drihtne gestreónde, Blickl. Homl. 121, 8.

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

tôhlíce; *adv.* *Toughly, tenaciously*:—Tôhlíce, thōlícæ, tóchtlícæ *uscide, viscide* (*viscide fortiter*, Migne), Txts. 107, 2170. Tólice *huscide*, 69, 1033.

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

OCR on scan of example page

1000 TÓH-LÍC -- TÓ-LICGAN.

hraegl and ðaet tShlutan *diviserunt sibi vestimento mea eí super vestem mearn miserunt soríem*, Ps. Th. 21, 16. Hié (*the apostles*) thysne middanyearð 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, thǫlíce, thǫchtlíce *uscide, viscide* (*viscidefortiter*, Migne), Txts. 107, 2170. Tthlíce *huscide*, 69, 1033.

tó-hlidan; *p.* -hlád, *pl.* -hlidon; *pp.* -hliden *To yawn, gape, open, crack* (intrans.), *split* (intrans.) *asunder*: – Tóhlád seó eor horne *terra dissilui*. 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 eorthscaef 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, \dots, C_k ■
- ★ priors $\pi_1, \pi_2, \dots, \pi_k$ ■
- ★ models $\mathbb{P}_{C_1}, \mathbb{P}_{C_2}, \dots, \mathbb{P}_{C_k}$ ■
- ★ 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 $w = w_1w_2 \dots w_k$ ■

★ reference measure $R_C(w) \equiv p_C(\wedge, w_1)p_C(w_1, w_2) \dots p_C(w_k, \$)$ ■

▷ *this is naïve - it assumes adjacent digrams are statistically independent*

■

★ Dirichlet digram measure $p_C(u, v) = \frac{\#(v|u)}{\sum_r \#(r|u)} \frac{+ \alpha \mu(v)}{+ \alpha}$ ■

★ α is a hyperparameter, and the optimum α should be chosen from tests on various corpora

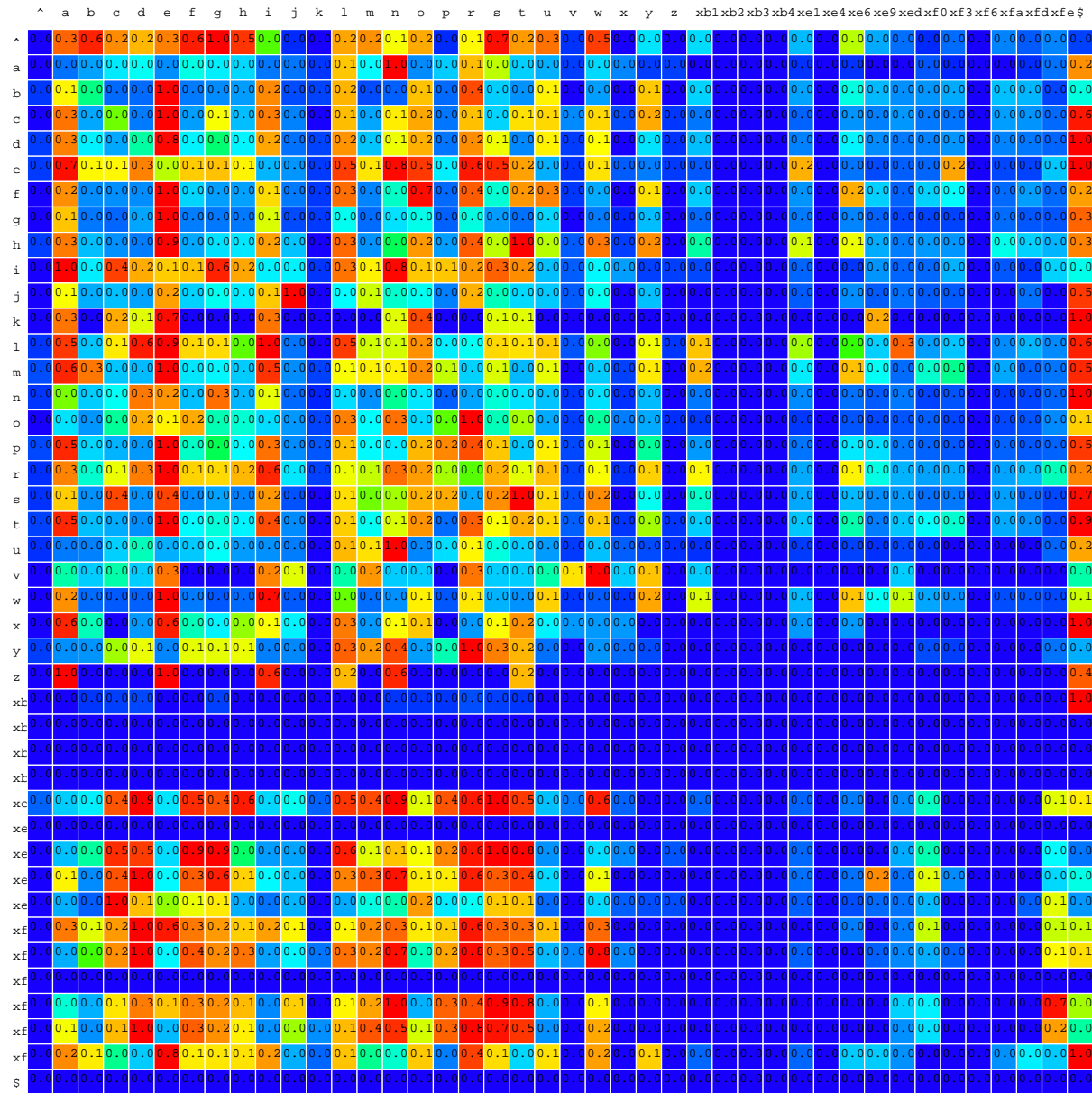
Coding issues

- ★ 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 $\bar{æ}$

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

	^	a	b	c	d	e	f	g	h	i	l	m	n	o	p	q	r	s	t	u	v	x	§
^	0.000	0.545	0.045	0.468	0.225	0.449	0.169	0.070	0.117	0.459	0.140	0.255	0.221	0.147	0.450	0.232	0.158	0.393	0.183	0.103	0.164	0.009	0.000
a	0.000	0.000	0.094	0.125	0.141	0.274	0.003	0.058	0.004	0.011	0.118	0.258	0.283	0.001	0.046	0.018	0.242	0.129	0.398	0.104	0.047	0.011	0.405
b	0.000	0.084	0.000	0.000	0.002	0.082	0.000	0.000	0.000	0.080	0.013	0.000	0.002	0.017	0.000	0.000	0.026	0.017	0.005	0.135	0.003	0.000	0.033
c	0.000	0.154	0.000	0.040	0.000	0.154	0.000	0.000	0.013	0.213	0.026	0.000	0.002	0.224	0.000	0.001	0.041	0.000	0.115	0.165	0.000	0.003	0.078
d	0.000	0.066	0.000	0.002	0.008	0.215	0.006	0.001	0.003	0.261	0.002	0.003	0.001	0.057	0.002	0.002	0.010	0.007	0.000	0.090	0.014	0.000	0.128
e	0.000	0.050	0.064	0.132	0.098	0.002	0.028	0.068	0.004	0.049	0.118	0.243	0.377	0.045	0.053	0.053	0.652	0.327	0.341	0.024	0.025	0.103	0.700
f	0.000	0.042	0.000	0.000	0.000	0.060	0.014	0.000	0.000	0.054	0.024	0.000	0.000	0.024	0.000	0.000	0.019	0.000	0.000	0.035	0.000	0.000	0.000
g	0.000	0.051	0.000	0.000	0.000	0.075	0.000	0.002	0.000	0.080	0.007	0.003	0.060	0.012	0.000	0.000	0.042	0.000	0.000	0.033	0.000	0.000	0.002
h	0.000	0.050	0.000	0.000	0.000	0.026	0.000	0.000	0.000	0.047	0.000	0.001	0.000	0.051	0.000	0.000	0.006	0.000	0.000	0.013	0.000	0.000	0.003
i	0.000	0.246	0.162	0.147	0.131	0.084	0.010	0.057	0.006	0.071	0.121	0.165	0.502	0.187	0.065	0.029	0.058	0.467	0.411	0.214	0.039	0.010	0.375
l	0.000	0.145	0.003	0.005	0.001	0.126	0.000	0.006	0.000	0.284	0.111	0.002	0.003	0.086	0.004	0.000	0.000	0.008	0.047	0.090	0.010	0.005	0.024
m	0.000	0.166	0.012	0.000	0.001	0.118	0.001	0.000	0.000	0.166	0.000	0.023	0.039	0.103	0.077	0.035	0.000	0.001	0.000	0.082	0.005	0.000	0.835
n	0.000	0.140	0.000	0.077	0.111	0.249	0.019	0.036	0.001	0.275	0.006	0.002	0.021	0.152	0.004	0.011	0.003	0.123	0.445	0.109	0.014	0.002	0.190
o	0.000	0.003	0.041	0.081	0.066	0.023	0.011	0.021	0.009	0.004	0.069	0.115	0.310	0.001	0.076	0.017	0.286	0.190	0.047	0.001	0.027	0.015	0.299
p	0.000	0.107	0.000	0.000	0.000	0.202	0.000	0.000	0.014	0.094	0.041	0.000	0.000	0.114	0.047	0.000	0.178	0.027	0.039	0.063	0.000	0.000	0.001
q	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.478	0.000	0.000	0.000
r	0.000	0.317	0.029	0.043	0.016	0.455	0.009	0.018	0.006	0.376	0.000	0.044	0.025	0.183	0.011	0.006	0.041	0.040	0.104	0.186	0.019	0.000	0.200
s	0.000	0.109	0.000	0.057	0.006	0.233	0.001	0.001	0.000	0.219	0.000	0.002	0.001	0.050	0.047	0.042	0.000	0.143	0.221	0.175	0.001	0.000	1.000
t	0.000	0.310	0.000	0.000	0.000	0.358	0.000	0.000	0.021	0.442	0.001	0.000	0.000	0.128	0.000	0.032	0.151	0.001	0.023	0.410	0.001	0.000	0.631
u	0.000	0.134	0.048	0.046	0.068	0.217	0.003	0.030	0.000	0.172	0.159	0.547	0.143	0.092	0.033	0.001	0.212	0.434	0.116	0.014	0.007	0.016	0.048
v	0.000	0.043	0.001	0.000	0.000	0.134	0.000	0.000	0.000	0.157	0.000	0.001	0.000	0.030	0.000	0.000	0.000	0.001	0.000	0.012	0.000	0.000	0.004
x	0.000	0.006	0.000	0.007	0.000	0.019	0.000	0.000	0.001	0.043	0.001	0.000	0.000	0.007	0.010	0.000	0.000	0.003	0.016	0.006	0.007	0.013	0.046
§	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Latin trigrams - a..

	^	a	b	c	d	e	f	g	h	i	l	m	n	o	p	q	r	s	t	u	v	x	
^	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
a	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
b	0.000	0.007	0.000	0.000	0.045	0.116	0.000	0.000	0.004	0.042	0.015	0.000	0.029	0.029	0.000	0.000	0.032	0.181	0.000	0.017	0.000	0.000	0.000
c	0.000	0.012	0.004	0.674	0.000	0.071	0.000	0.000	0.042	0.111	0.000	0.001	0.000	0.002	0.000	0.003	0.085	0.000	0.081	0.009	0.000	0.000	0.000
d	0.000	0.025	0.000	0.024	0.241	0.196	0.199	0.043	0.093	0.254	0.051	0.125	0.034	0.058	0.074	0.006	0.020	0.254	0.011	0.157	0.443	0.000	0.000
e	0.000	0.003	0.000	0.002	0.170	0.001	0.001	0.129	0.000	0.000	0.009	0.064	0.013	0.014	0.001	0.137	0.041	0.151	0.183	0.000	0.017	0.001	0.000
f	0.000	0.001	0.000	0.000	0.000	0.002	0.003	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.086	0.000	0.002	0.010	0.000	0.000	0.000
g	0.000	0.012	0.000	0.000	0.000	0.122	0.000	0.050	0.000	0.063	0.002	0.097	0.008	0.000	0.000	0.000	0.387	0.000	0.000	0.009	0.000	0.000	0.000
h	0.000	0.000	0.000	0.000	0.000	0.023	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
i	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.014	0.000	0.000	0.000	0.000
l	0.000	0.047	0.062	0.004	0.000	0.065	0.003	0.001	0.000	0.917	0.046	0.001	0.000	0.007	0.047	0.000	0.000	0.004	0.345	0.016	0.043	0.000	0.000
m	0.000	0.027	0.238	0.000	0.000	0.009	0.000	0.000	0.000	0.264	0.000	0.000	0.204	0.100	0.163	0.000	0.000	0.000	0.000	0.029	0.000	0.000	0.000
n	0.000	0.029	0.000	0.051	0.033	0.006	0.004	0.089	0.005	0.331	0.000	0.000	0.367	0.000	0.000	0.004	0.000	0.005	0.601	0.006	0.000	0.020	0.000
o	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.001	0.000	0.000	0.000
p	0.000	0.011	0.000	0.000	0.000	0.089	0.000	0.000	0.011	0.028	0.000	0.000	0.000	0.049	0.333	0.000	0.025	0.013	0.015	0.343	0.000	0.000	0.000
q	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.283	0.001	0.000	0.000
r	0.000	0.106	0.164	0.101	0.070	0.060	0.002	0.117	0.000	0.114	0.000	0.422	0.006	0.002	0.007	0.000	0.045	0.050	0.198	0.006	0.042	0.004	0.000
s	0.000	0.002	0.001	0.036	0.000	0.000	0.000	0.000	0.000	0.130	0.000	0.000	0.000	0.003	0.116	0.000	0.000	0.020	0.043	0.000	0.000	0.000	0.000
t	0.000	0.006	0.000	0.000	0.000	0.015	0.000	0.000	0.030	0.017	0.021	0.000	0.000	0.000	0.000	0.763	0.089	0.000	0.155	0.005	0.000	0.000	0.000
u	0.000	0.001	0.000	0.219	0.229	0.000	0.022	0.336	0.000	0.000	0.028	0.000	0.001	0.000	0.000	0.000	0.080	0.105	1.000	0.000	0.000	0.179	0.000
v	0.000	0.040	0.000	0.000	0.000	0.047	0.000	0.001	0.000	0.077	0.000	0.000	0.000	0.015	0.000	0.000	0.000	0.001	0.000	0.022	0.000	0.000	0.000
x	0.000	0.002	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Latin trigrams - b..

	^	a	b	c	d	e	f	g	h	i	l	m	n	o	p	q	r	s	t	u	v	x	§
^	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
a	0.000	0.000	0.015	0.028	0.002	0.068	0.000	0.009	0.000	0.015	0.074	0.006	0.009	0.000	0.000	0.000	0.228	0.047	0.021	0.008	0.000	0.000	0.000
b	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
c	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
d	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
e	0.000	0.008	0.002	0.000	0.002	0.004	0.000	0.002	0.000	0.002	1.000	0.000	0.362	0.000	0.000	0.000	0.028	0.008	0.004	0.000	0.000	0.000	0.000
f	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
g	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
h	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
i	0.000	0.000	0.072	0.002	0.043	0.017	0.004	0.002	0.000	0.000	0.006	0.009	0.055	0.000	0.019	0.000	0.002	0.066	0.098	0.000	0.000	0.000	0.000
l	0.000	0.085	0.000	0.000	0.000	0.008	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.000
m	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
n	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
o	0.000	0.006	0.000	0.006	0.009	0.030	0.000	0.002	0.000	0.034	0.019	0.006	0.281	0.004	0.000	0.000	0.025	0.045	0.008	0.013	0.051	0.000	0.000
p	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
q	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
r	0.000	0.102	0.000	0.000	0.000	0.134	0.000	0.000	0.000	0.196	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.000	0.000	0.134	0.000	0.000	0.000
s	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
t	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u	0.000	0.000	0.070	0.019	0.004	0.000	0.000	0.000	0.000	0.000	0.015	0.000	0.000	0.000	0.004	0.000	0.057	0.009	0.019	0.000	0.000	0.002	0.000
v	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
x	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
§	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Latin trigrams - c..

	^	a	b	c	d	e	f	g	h	i	l	m	n	o	p	q	r	s	t	u	v	x	
^	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
a	0.000	0.000	0.003	0.003	0.017	0.305	0.000	0.000	0.000	0.001	0.061	0.034	0.037	0.000	0.106	0.000	0.075	0.178	0.023	0.093	0.015	0.000	0.000
b	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
c	0.000	0.000	0.000	0.014	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.007	0.000
d	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
e	0.000	0.001	0.001	0.008	0.010	0.000	0.000	0.000	0.000	0.000	0.074	0.000	0.074	0.001	0.010	0.000	0.080	0.011	0.083	0.003	0.001	0.000	0.000
f	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
g	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
h	0.000	0.018	0.000	0.000	0.000	0.012	0.000	0.000	0.000	0.004	0.001	0.000	0.000	0.001	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000
i	0.000	0.000	0.010	0.008	0.001	0.002	0.000	0.000	0.000	0.001	0.010	0.006	0.013	0.001	0.000	0.000	0.133	0.004	0.021	0.001	0.119	0.000	0.000
l	0.000	0.133	0.000	0.000	0.000	0.013	0.000	0.000	0.000	0.013	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.005	0.000	0.009	0.000
m	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
n	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
o	0.000	0.018	0.000	0.008	0.005	0.056	0.000	0.120	0.053	0.009	0.108	0.212	1.000	0.004	0.069	0.007	0.147	0.008	0.027	0.002	0.000	0.001	0.000
p	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
q	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
r	0.000	0.027	0.000	0.000	0.000	0.100	0.000	0.000	0.000	0.039	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.019	0.000	0.000
s	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
t	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u	0.000	0.000	0.013	0.001	0.000	0.000	0.000	0.000	0.000	0.095	0.022	0.367	0.060	0.000	0.036	0.000	0.079	0.019	0.002	0.000	0.000	0.000	0.000
v	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
x	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.009
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Latin trigrams - d..

	^	a	b	c	d	e	f	g	h	i	l	m	n	o	p	q	r	s	t	u	v	x	§
^	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
a	0.000	0.000	0.054	0.008	0.000	0.005	0.000	0.000	0.005	0.000	0.005	0.137	0.048	0.000	0.026	0.000	0.151	0.008	0.329	0.006	0.003	0.000	0.000
b	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
c	0.000	0.000	0.000	0.035	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.012	0.000
d	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
e	0.000	0.012	0.069	0.578	0.385	0.036	0.558	0.078	0.071	0.481	0.428	0.262	0.125	0.095	0.231	0.003	0.057	0.465	0.237	0.066	0.071	0.087	0.000
f	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
g	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
h	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
i	0.000	0.060	0.000	0.647	0.044	0.855	0.152	0.202	0.000	0.023	0.148	0.187	0.011	0.051	0.003	0.000	0.092	1.000	0.026	0.169	0.517	0.179	0.000
l	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.015	0.000
m	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
n	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
o	0.000	0.000	0.002	0.101	0.006	0.000	0.000	0.002	0.000	0.000	0.265	0.593	0.219	0.000	0.000	0.000	0.033	0.003	0.008	0.000	0.000	0.000	0.000
p	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
q	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
r	0.000	0.032	0.000	0.000	0.000	0.006	0.000	0.000	0.000	0.005	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.190	0.000	0.000
s	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
t	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u	0.000	0.223	0.169	0.354	0.000	0.003	0.000	0.000	0.000	0.005	0.038	0.293	0.002	0.252	0.032	0.000	0.098	0.000	0.000	0.000	0.000	0.083	0.000
v	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
x	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.006	0.000
§	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Latin trigrams - e..

	^	a	b	c	d	e	f	g	h	i	l	m	n	o	p	q	r	s	t	u	v	x	ç
^	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
a	0.000	0.000	0.001	0.009	0.154	0.018	0.000	0.000	0.000	0.000	0.001	0.162	0.029	0.000	0.000	0.020	0.041	0.081	0.006	0.000	0.000	0.000	0.000
b	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.029	0.000	0.000	0.000
c	0.000	0.000	0.001	0.001	0.002	0.000	0.000	0.000	0.008	0.000	0.002	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
d	0.000	0.001	0.000	0.000	0.000	0.032	0.000	0.000	0.000	0.074	0.000	0.000	0.010	0.000	0.000	0.000	0.000	0.000	0.000	0.048	0.000	0.000	0.000
e	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
f	0.000	0.000	0.000	0.000	0.000	0.000	0.245	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000
g	0.000	0.000	0.000	0.000	0.000	0.050	0.000	0.001	0.000	0.022	0.000	0.000	0.003	0.024	0.000	0.000	0.121	0.000	0.000	0.002	0.001	0.000	0.000
h	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
i	0.000	0.001	0.000	0.006	0.001	0.010	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.000	0.046	0.005	0.643	0.001	0.000	0.000
l	0.000	0.034	0.001	0.000	0.000	0.049	0.000	0.000	0.000	0.027	0.002	0.000	0.000	0.026	0.000	0.000	0.000	0.001	0.000	0.026	0.000	0.000	0.000
m	0.000	0.004	0.000	0.000	0.000	0.030	0.000	0.000	0.000	0.058	0.000	0.001	0.000	0.006	0.022	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000
n	0.000	0.007	0.000	0.002	0.002	0.001	0.000	0.001	0.000	0.167	0.000	0.000	0.004	0.007	0.000	0.000	0.000	0.000	0.001	0.014	0.000	0.000	0.000
o	0.000	0.000	0.000	0.000	0.189	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.000	0.030	0.230	0.186	0.000	0.001	0.000	0.000	0.000
p	0.000	0.004	0.000	0.000	0.000	0.001	0.000	0.000	0.020	0.055	0.000	0.000	0.000	0.015	0.000	0.000	0.005	0.000	0.000	0.035	0.000	0.000	0.000
q	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.528	0.000	0.000	0.000
r	0.000	0.468	0.000	0.001	0.000	0.031	0.000	0.083	0.000	0.196	0.000	0.000	0.001	0.005	0.000	0.000	0.027	0.000	0.000	0.126	0.004	0.000	0.000
s	0.000	0.000	0.001	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.689	1.000	0.006	0.000	0.000	0.000
t	0.000	0.001	0.000	0.001	0.000	0.018	0.000	0.000	0.000	0.319	0.000	0.000	0.000	0.000	0.000	0.000	0.021	0.034	0.000	0.000	0.000	0.000	0.000
u	0.000	0.001	0.011	0.001	0.006	0.002	0.000	0.004	0.002	0.002	0.002	0.198	0.056	0.003	0.032	0.000	0.052	0.000	0.002	0.000	0.000	0.002	0.000
v	0.000	0.020	0.000	0.000	0.000	0.066	0.000	0.000	0.000	0.016	0.000	0.000	0.000	0.028	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.000	0.000
x	0.000	0.074	0.000	0.242	0.000	0.479	0.001	0.000	0.022	0.408	0.000	0.000	0.000	0.104	0.374	0.014	0.000	0.123	0.366	0.074	0.000	0.000	0.000
ç	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

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=eplam-
pua(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)
jniss
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)
neque
36.60=neque(1)
opporlunilatam
66.51=oppartumitatem(4) 66.97=oppartunitatem(3) 67.24=opportunitatem(3) 67.28=appartumitatem(5) 67.6
9=opportunitatem(2)
out
31.88=aut(1)
patibuium
```