Language Identification

Scott Martin (Linguistics 384)

Wednesday, April 30, 2008

Using the trigram frequency distribution for Czech and Polish on the back of this handout, you want to identify the language of the following phrase:

```
A to ci leci!
```

1. In the following table, list all the trigrams in the phrase above:

Trigram	Czech	Polish

Trigram	Czech	Polish

2. Using the table on the back, fill out the rest of the table and figure out which language this method predicts for the phrase above.

3. Looking at the table on the back again, find one character that is only used in Czech and one that is only used in Polish.

Partial Trigram Frequency Distribution Table for Czech and Polish

Trigram	Czech	Polish	Trigram	Czech	Polish	Trigram	Czech	Polish
A_o	1	0	_tr	10	8	o_d	20	23
A_p	2	1	_tu	14	13	o_e	5	0
A_r	0	1	_tv	13	0	o_f	0	2
A_t	12	3	chł	0	5	osł	0	4
A_u	1	0	chż	0	1	osť	10	0
A_v	5	0	chá	6	0	osó	0	3
A_w	0	2	chę	0	2	osý	1	0
La_	1	0	chř	1	0	ot,	0	2
Lak	0	1	chů	7	0	ot_	5	2
Laz	6	0	ci!	0	1	ota	4	17
Lec	0	2	ci,	1	2	otc	5	0
Len	0	1	ci.	1	2	ote	1	7
Lid	1	0	ci_	14	20	tně	2	0
Los	0	1	cia	0	29	tní	7	0
Lub	0	1	cic	0	16	tný	1	0
Lži	1	0	ece	1	0	to,	3	8
Lęk	0	1	ech	29	15	to.	0	2
_bę	0	2	eci	1	18	to?	0	3
_bě	3	0	eck	6	2	to_	57	39
_bř	1	0	ecn	1	5	tob	1	0
_bý	1	0	eco	0	1	zi:	1	0
_ca	0	7	ect	2	0	zi_	5	3
_ce	11	4	ecy	0	2	zia	0	6
_ch	24	42	ecz	0	20	zib	0	2
_ci	0	42	i_g	1	8	zic	0	2
_co	12	24	i_h	3	0	zie	0	81
_ką	0	6	i_i	0	16	zig	0	1
_kł	0	1	i_j	8	14	zim	0	5
_ká	1	0	i_k	9	8	zio	0	1
_kę	0	1	i_l	2	5	zis	0	2
_kó	1	0	i_m	14	10	zit	1	0
_kř	13	0	i_n	14	12	ziw	0	7
_la	24	9	i_o	7	9	ły.	0	2
_le	7	14	i_p	26	27	ły_	0	51
_lh	3	0	leb	3	0	łyc	0	2
_oz	1	0	lec	8	9	łyn	0	5
_oś	0	1	led	12	1	łys	0	13
_oż	0	1	leg	0	12	ływ	0	4
_oč	4	0	leh	2	0	śta	0	1
_pa	19	40	lei	0	2	świ	0	22
_pe	7	5	lej	1	7	ść?	0	1
_pi	3	30	lek	12	7	ść_	0	13
_pl	5	2	lel	1	0	ścu	0	1
_po	121	171	lem	6	2	š,_	6	0
_pr	50	105	o_M	2	1	š.*	1	0
_ti	15	0	o_P	11	0	š..	4	0
_tk	0	2	o_S	1	0	š_A	1	0
_tl	1	1	o_a	3	2	š_a	3	0
_tm	1	0	o_b	11	16	3_|	2	0
_to	55	31	o_c	8	21	3_č	2	0