Historia i podstawy kryptografii
Spotkanie 5
Nieco więcej o analizie częstotliwości

Jacek Rogowski

I Liceum Ogólnokształcące
w Łowiczu

Częstotliwość występowania liter

- Częstotliwość występowania liter oblicza się na podstawie analizy tekstów z tzw. **Korpusu języka**, czyli zbioru wielu tekstów z różnych dziedzin napisanych w danym języku.
- Częstotliwość występowania litery oblicza się jako stosunek liczby wystąpień danej litery w ustalonym zbiorze tekstów do liczby wszystkich liter w tych tekstach.
- Dla języka polskiego obliczono podane dalej częstotliwości występowania liter na podstawie ustalonego zbioru zawierającego około 25 milionów słów.
- W obliczeniach bierze się pod uwagę również słowa pochodzenia obcego, co powoduje konieczność uwzględnienia liter z innych alfabetów, np. **q**, **x**, **v**.

Częstotliwości występowania 35 liter alfabetu polskiego

a	ą	b	c	ć	d	e
8,91%	0,99%	1,47%	3,96%	0,40%	3,25%	7,66%
ę	f	g	h	i	j	k
1,11%	0,30%	1,42%	1,08%	8,21%	2,28%	3,51%
l	ł	m	n	ń	o	ó
2,10%	1,82%	2,80%	5,52%	0,20%	7,75%	0,85%
p	q	r	s	ś	t	u
3,13%	0,14%	4,69%	4,32%	0,66%	3,98%	2,50%
v	w	x	y	z	ź	ż
0,04%	4,65%	0,02%	3,76%	5,64%	0,06%	0,83%

Kilka ważnych uwag

- Praktycznie każdy język zachodnioeuropejski posługuje się własną odmianą alfabetu łacińskiego o innej liczbie liter.
- Używanie liter charakterystycznych dla danego języka może ułatwiać łamanie szyfrów.
- Szyfrowanie maszynowe wymaga ciągłej (czyli: bez luk) numeracji liter alfabetu.
- W (siedmiobitowym) standardzie kodowania znaków ASCII ciągłą numerację mają wielkie i małe litery 26 literowego alfabetu angielskiego:

ABCDEFGHIJKLMNOPQRSTUVWXYZ
abcdefghijklmnopqrstuvwxyz

Tablica kodów ASCII dla liter

W konsekwencji wiadomość jawną zapisuje się korzystając z alfabetu 26 literowego uwzględnionego w standardzie ASCII, przy czym litery charakterystyczne dla danego języka zastępuje się ich standardowymi odpowiednikami (np. zamiast **ą** pisze się **a**, zamiast **ć** używa się **c** itd.):

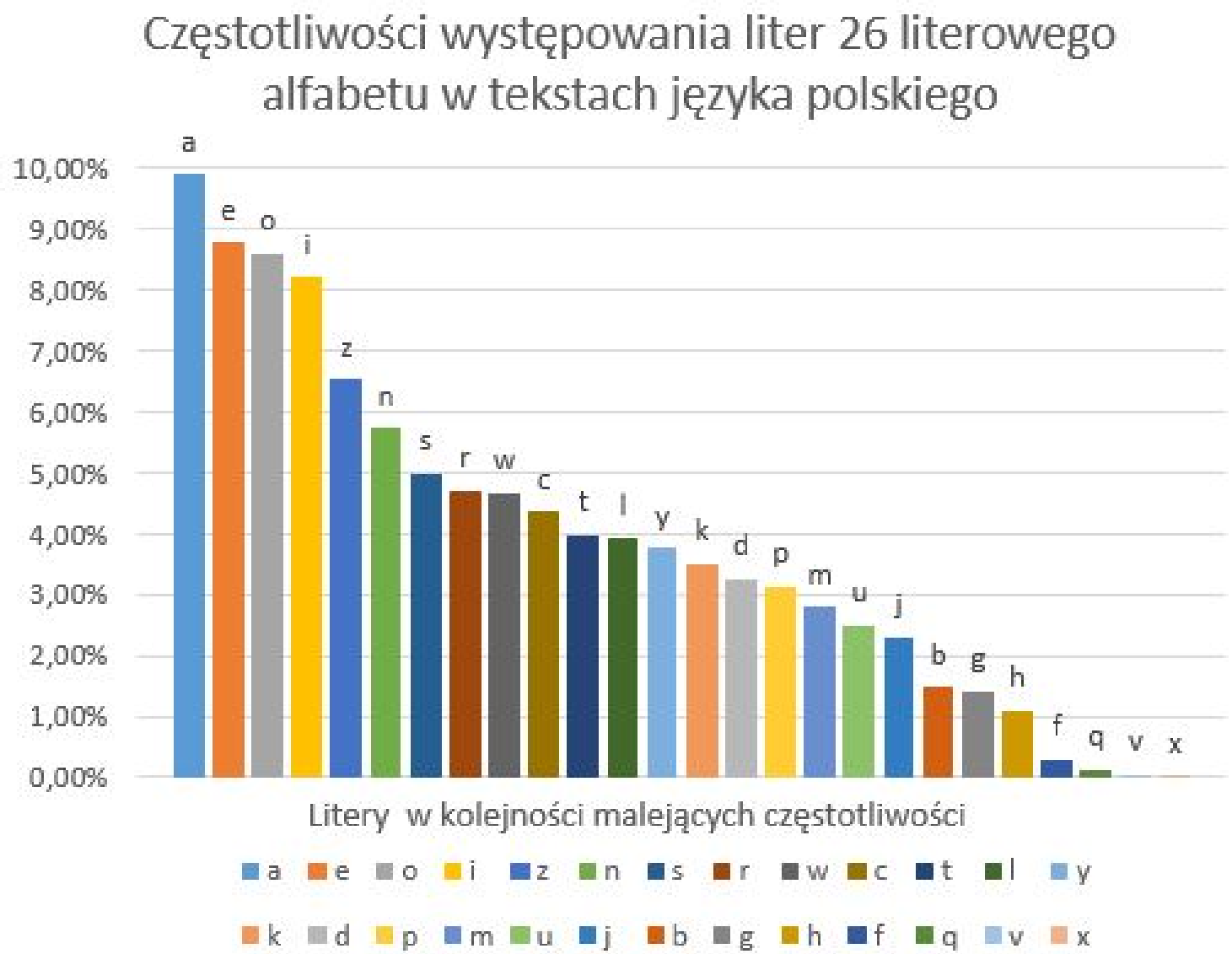
Kod	Litera	Kod	Litera	Kod	Litera	Kod	Litera
65	A	78	N	97	a	110	n
66	B	79	O	98	b	111	o
67	C	80	P	99	c	112	p
68	D	81	Q	100	d	113	q
69	E	82	R	101	e	114	r
70	F	83	S	102	f	115	s
71	G	84	T	103	g	116	t
72	H	85	U	104	h	117	u
73	I	86	V	105	i	118	v
74	J	87	W	106	j	119	w
75	K	88	X	107	k	120	x
76	L	89	Y	108	l	121	y
77	M	90	Z	109	m	122	z

Częstotliwości występowania 26 liter alfabetu polskiego

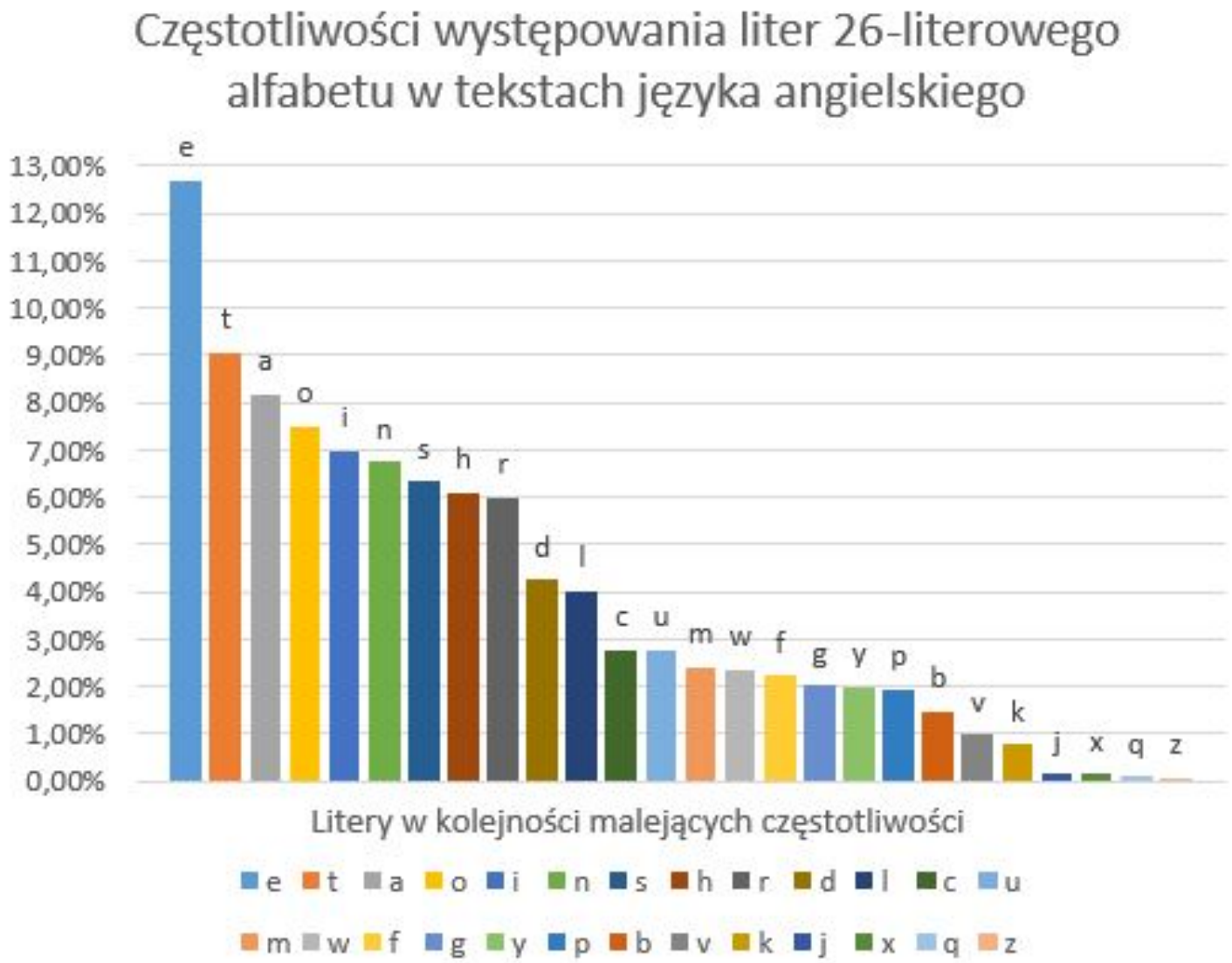
Częstotliwości występowania liter 26 literowego alfabetu polskiego otrzymuje się sumując częstotliwości dla „podobnych” liter w alfabecie 35 literowym.

a	b	c	d	e	f	g
9,90%	1,47%	4,36%	3,25%	8,77%	0,30%	1,42%
h	**i**	**j**	**k**	**l**	**m**	**n**
1,08%	8,21%	2,28%	3,51%	3,92%	2,80%	5,72%
o	**p**	**q**	**r**	**s**	**t**	**u**
8,60%	3,13%	0,14%	4,69%	4,98%	3,98%	2,50%
v	**w**	**x**	**y**	**z**		
0,04%	4,65%	0,02%	3,76%	6,53%		

Histogram częstotliwości dla alfabetu polskiego



Histogram częstotliwości dla alfabetu angielskiego



Ustalanie języka wiadomości jawnej i typu szyfru

- Każdy język ma charakterystyczny dla niego histogram częstotliwości występowania liter.
- Histogram częstotliwości występowania liter w wiadomości zaszyfrowanej ma kształt podobny do histogramu dla danego języka — fakt ten pozwala ustalić, w jakim języku była zapisana wiadomość jawna.
- W przypadku szyfru przestawieniowego kolejne słupki w histogramie dla wiadomości tajnej mają takie same, lub prawie takie same, etykiety literowe, jak kolejne słupki w histogramie dla danego języka.
- W przypadku monoalfabetycznego szyfru podstawieniowego etykiety literowe słupków w histogramie dla wiadomości tajnej są zdecydowanie różne od etykiet literowych słupków w histogramie dla danego języka.