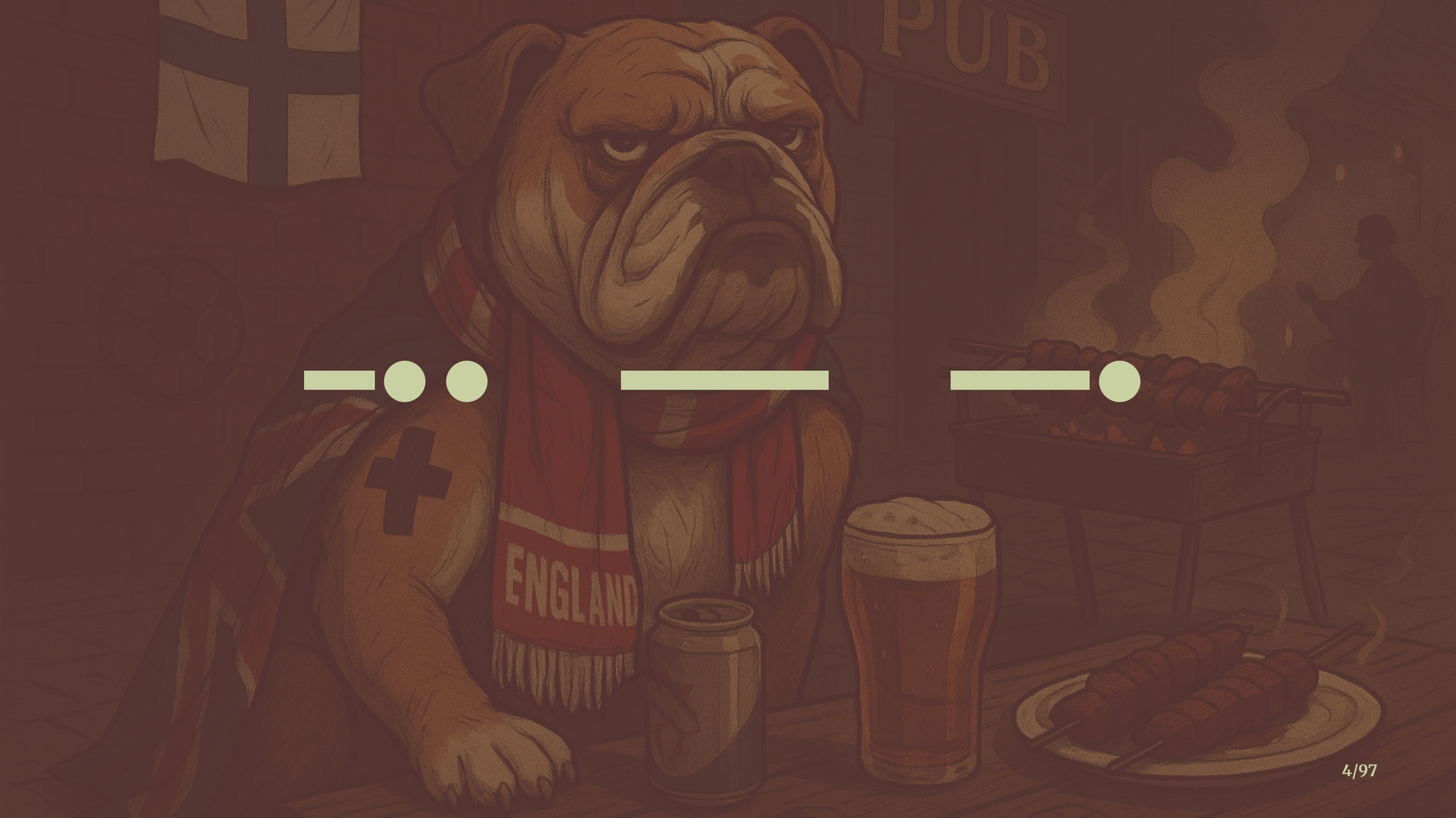
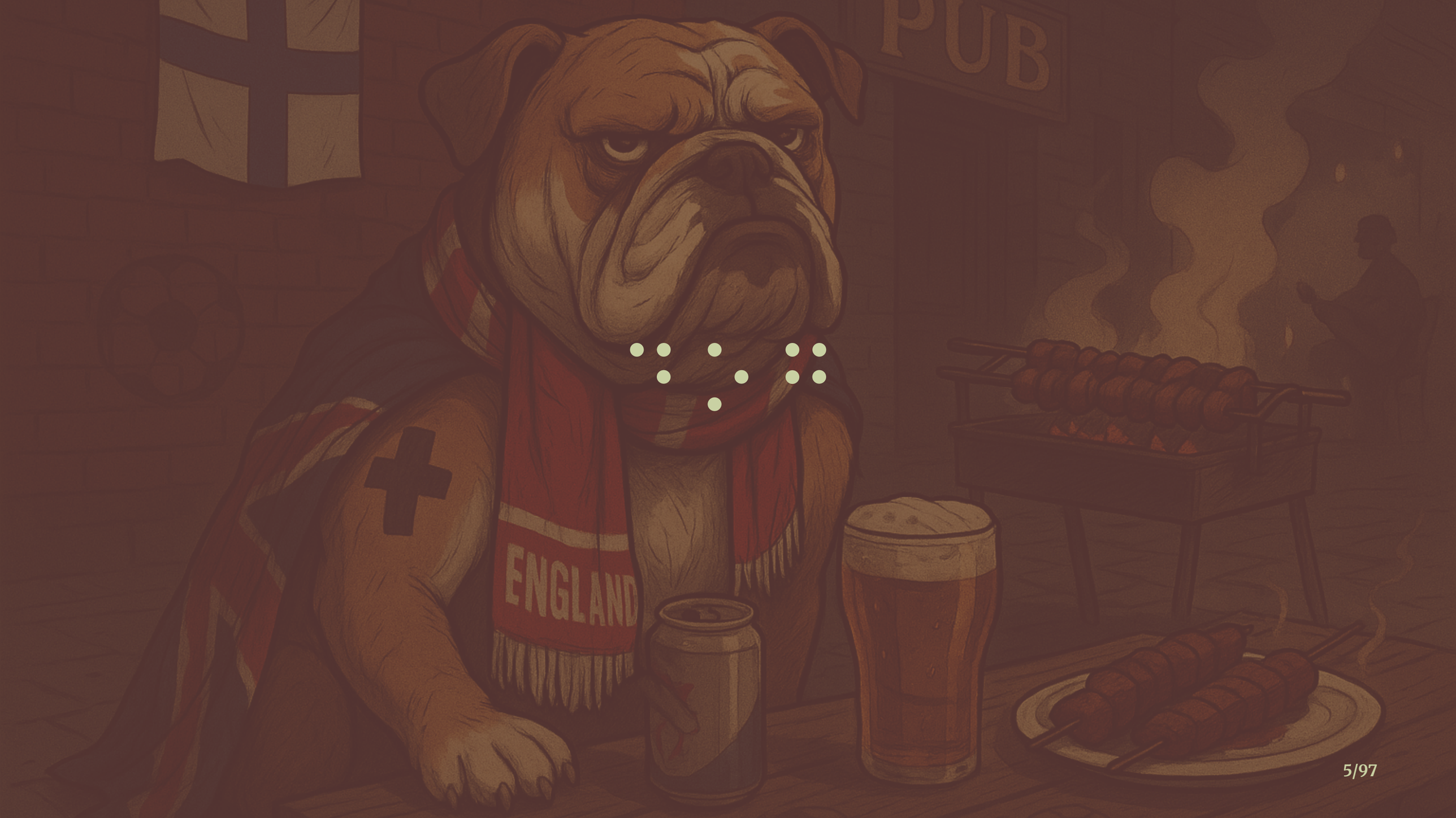

PUB
ENGLAND

dog


PUB
ENGLAND


PUB
ENGLAND

[64 6F 67]

Oracle Database Character Set

Character Set

Character Set Encoding

 dogs.txt

Hund
abwo
alabai
anjing
ashun
aso
cane
cão
câin
câine
câini
chien
chó
cyn
dog
emba
ghaddu
gom
hond
hondur
hualp
huan
hund
imba
imbua
imbwa
imbwe
inu
it
jindo
kelb
khuy
klèb
koer
koira
kotta
køter
kutya
kutta
köpek
köpeği
mbula
mbwa
mbwagh
mbwá
mbwene
mbʉ
njau
njoka
pa
perro
pes
pies
qen
qeni
sag
sagol
sobaka
suns
szczeniak
tz’i’
ukudla
welpe
xolo
zwin
āso
āšun
šuo
אבו
כלב
كلب
कुत्ता
ฆ่า
犬
狗
ᛕᛁᚦ
łééchąą’í


bcdic-a

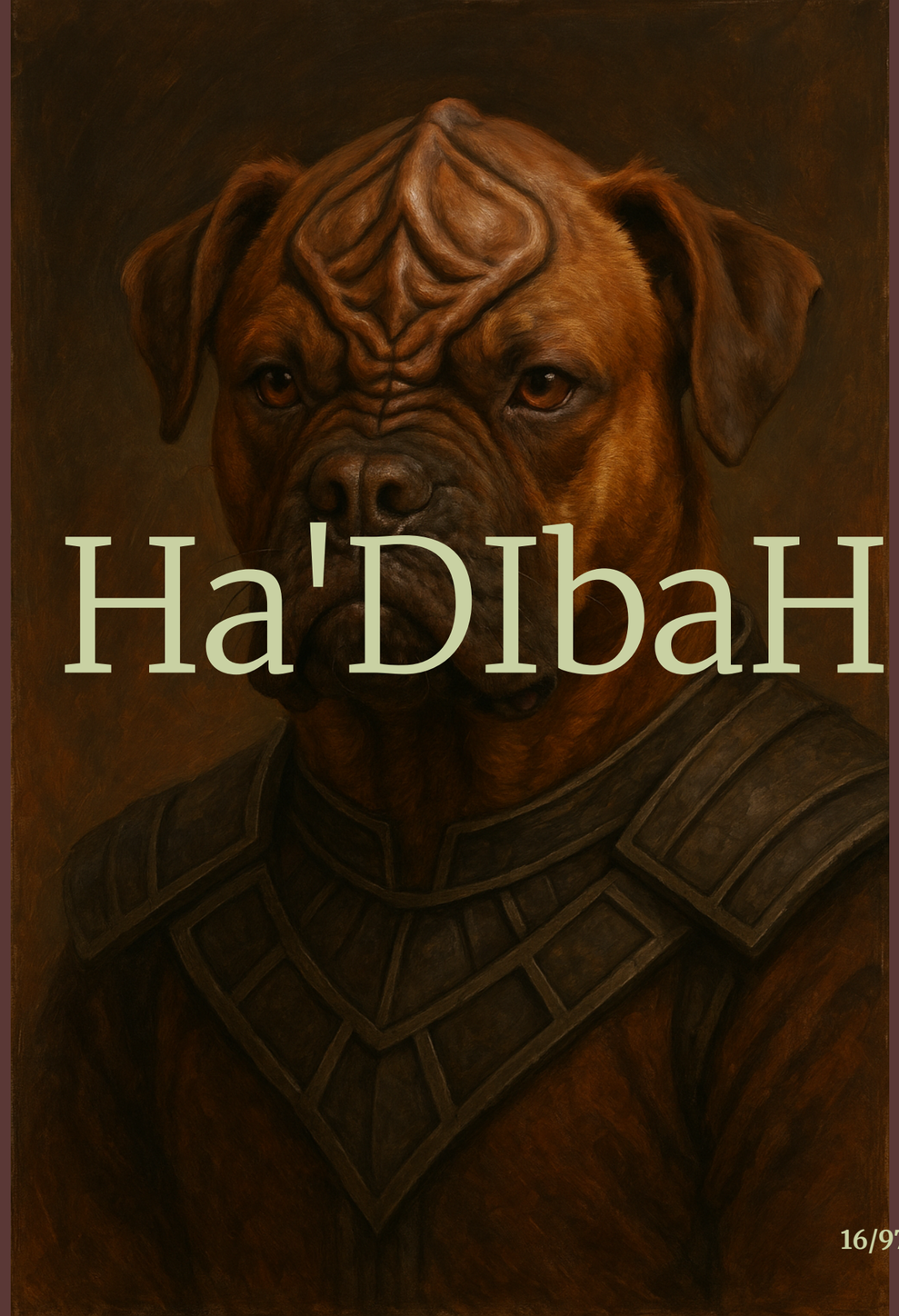
	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
0x		1	2	3	4	5	6	7	8	9	0	#	@	:	>	√
1x	₺	/	S	T	U	V	W	X	Y	Z	=	,	%	ɣ	\	₩
2x	–	J	K	L	M	N	O	P	Q	R	!	#	*	]	;	Δ
3x	&	A	B	C	D	E	F	G	H	I	?	.	¤	[	<	

ascii

ascii	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
0x	‹nul›	‹soh›	‹stx›	‹etx›	‹eot›	‹enq›	‹ack›	‹bel›	‹bs›	‹tab›	‹lf›	‹vt›	‹ff›	‹cr›	‹so›	‹si›
1x	‹dle›	‹dc1›	‹dc2›	‹dc3›	‹dc4›	‹nak›	‹syn›	‹etb›	‹can›	‹em›	‹sub›	‹esc›	‹fs›	‹gs›	‹rs›	‹us›
2x		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{	`	`	}	~

```
→ encode.py -b ascii WOOF,woof
✅ ascii:    "WOOF"=[01010111 01001111 01001111 01000110]
✅ ascii:    "woof"=[01110111 01101111 01101111 01100110]
```

Huan



Ha'DIbaH

```
→ encode.py ascii -d Huan,"Ha'DIbaH"
✅ ascii:   Good Huan [48 75 61 6e] (4 bytes)
✅ ascii:   Good Ha'DIbaH [48 61 27 44 49 62 61 48] (8 bytes)
```

```
→ encode.py ascii < dogs.txt
✅ ascii: 53 good dogs
✅ ascii: 237 chars encoded in 237 bytes, 1.0 bytes per char
❌ ascii: 27 bad dogs:
❌ ascii: cão  câin  câine  câini  chó  klèb  køter  köpek
❌ ascii: köpeğI  mbwá  mbʉ  tz'i'  āso  āšun  šuo  אבו
❌ ascii: كلب  כלב  कुत्ता  ฆ่า  犬  狗  ᛕᛁᛟ  łéédchąą'í
❌ ascii: 🦮  🐕‍🦺  🐶
```

ChatGPT 4o >

what encoding did oracle the oracle database use prior to version 6?

Prior to Oracle Database version 6 (released in 1988), Oracle used **7-bit US-ASCII** as the default and primary character encoding. This encoding was sufficient at the time because Oracle's early customer base was predominantly English-speaking, and multilingual or multibyte character support was not a focus.

Oracle Pre-6

» US7ASCII

```
US   7 ASCII


WE   8 ISO8859P1
WE   8 ISO8859P2

WE   8 MSWIN1252
EE   8 MSWIN1250

ZHS 16 CGB231280

AL  16 UTF16
AL  32 UTF8
```

iso-8859

Standard	Alias	Region	Year
iso-8859-1	**latin1**	**Western Europe**	1987
iso-8859-2	**latin2**	**Central & Eastern Europe**	1987
iso-8859-3	latin3	South Europe	1988
iso-8859-4	latin4	North Europe	1988
iso-8859-5	cyrillic		1988
iso-8859-6	arabic		1987
iso-8859-7	greek		1987
iso-8859-8	hebrew		1988
iso-8859-9	latin5	Turkish	1989
iso-8859-10	latin6	Nordic	1992
iso-8859-11	thai		2001
iso-8859-13	latin7	Baltics	1998
iso-8859-14	latin8	Celtic	1998
iso-8859-15	**latin9**	**Western Europe (Improved)**	1999
iso-8859-16	latin10	South-Eastern Europe	2001

ascii	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
0x	‹nul›	‹soh›	‹stx›	‹etx›	‹eot›	‹enq›	‹ack›	‹bel›	‹bs›	‹tab›	‹lf›	‹vt›	‹ff›	‹cr›	‹so›	‹si›
1x	‹dle›	‹dc1›	‹dc2›	‹dc3›	‹dc4›	‹nak›	‹syn›	‹etb›	‹can›	‹em›	‹sub›	‹esc›	‹fs›	‹gs›	‹rs›	‹us›
2x		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	‹	=	›	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{	`	`	}	~
8x																
9x																
ax																
bx																
cx																
dx																
ex																
fx																

latin1	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
0x	‹nul›	‹soh›	‹stx›	‹etx›	‹eot›	‹enq›	‹ack›	‹bel›	‹bs›	‹tab›	‹lf›	‹vt›	‹ff›	‹cr›	‹so›	‹si›
1x	‹dle›	‹dc1›	‹dc2›	‹dc3›	‹dc4›	‹nak›	‹syn›	‹etb›	‹can›	‹em›	‹sub›	‹esc›	‹fs›	‹gs›	‹rs›	‹us›
2x		!	"	#	$	%	&	'	(	)	*	+	,	–	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	‹	=	›	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{	`	`	}	~
8x																
9x																
ax	¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬		®	-	°
bx	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿	
cx	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
dx	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
ex	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
fx	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ


cão


köpek

```
→ encode.py latin1 -d cão,köpek
✅ latin1:  Good cão [63 e3 6f] (3 bytes)
✅ latin1:  Good köpek [6b f6 70 65 6b] (5 bytes)
```

```
➜ encode.py latin1 < dogs.txt
✅ latin1: 62 good dogs
✅ latin1: 275 chars encoded in 275 bytes, 1.0 bytes per char
❌ latin1: 18 bad dogs:
❌ latin1: köpeği  mbʉ  tz’i’  āso  āšun  šuo  כלב  אבו
❌ latin1: كلب  कुत्ता  หมา  犬  狗  ꓘИѲ  łééchąą’í  Ꮖ
❌ latin1: 𓃡  🐶
```

ascii	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
0x	‹nul›	‹soh›	‹stx›	‹etx›	‹eot›	‹enq›	‹ack›	‹bel›	‹bs›	‹tab›	‹lf›	‹vt›	‹ff›	‹cr›	‹so›	‹si›
1x	‹dle›	‹dc1›	‹dc2›	‹dc3›	‹dc4›	‹nak›	‹syn›	‹etb›	‹can›	‹em›	‹sub›	‹esc›	‹fs›	‹gs›	‹rs›	‹us›
2x		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{	`	`	}	~
8x																
9x																
ax																
bx																
cx																
dx																
ex																
fx																

latin2	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
0x	‹nul›	‹soh›	‹stx›	‹etx›	‹eot›	‹enq›	‹ack›	‹bel›	‹bs›	‹tab›	‹lf›	‹vt›	‹ff›	‹cr›	‹so›	‹si›
1x	‹dle›	‹dc1›	‹dc2›	‹dc3›	‹dc4›	‹nak›	‹syn›	‹etb›	‹can›	‹em›	‹sub›	‹esc›	‹fs›	‹gs›	‹rs›	‹us›
2x		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{	`	`	}	~
8x																
9x																
ax		ˇ	˘	Ł	¤	Ľ	Ś	Š	Ş	Ť	Ź	¡	¸	˙	˝	Ů
bx	ź	‘	’	“	”	•	–	—	™							
cx	Ą	˘	Ď	Đ	Ë	Ę	Ě	Í	Î	Ï	˝	Ł	Ń	Ň	Ó	Ô
dx	Ő	Œ	Ŕ	Ř	Ŗ	Ś	Ş	Š	Ţ	Ť	Ź	Ż	Ž			
ex	ą	˛	ď	đ	ë	ę	ě	í	î	ï	˝	ł	ń	ň	ó	ô
fx	ő	œ	ŕ	ř	ŗ	ś	ş	š	ţ	ť	ź	ż	ž	ÿ		

šuo

câine


šuo


câine

```
→ encode.py latin2 -d šuo,câine
✅ latin2:  Good šuo [b9 75 6f] (3 bytes)
✅ latin2:  Good câine [63 e2 69 6e 65] (5 bytes)
```

```
➜ encode.py latin2 < dogs.txt
✅ latin2: 60 good dogs
✅ latin2: 266 chars encoded in 266 bytes, 1.0 bytes per char
❌ latin2: 20 bad dogs:
❌ latin2: cão  klèb  køter  köpeği  mbʉ  tz’i’  āso  āšun
❌ latin2: كلب  כלב  אבו  कुत्ता  ໝາ  犬  狗  ꓘИѲ
❌ latin2: łééchąą’í  ⩑  𓃡  🐶
```

windows-1252

latin1	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0x	‹nul›	‹soh›	‹stx›	‹etx›	‹eot›	‹enq›	‹ack›	‹bel›	‹bs›	‹tab›	‹lf›	‹vt›	‹ff›	‹cr›	‹so›	‹si›
1x	‹dle›	‹dc1›	‹dc2›	‹dc3›	‹dc4›	‹nak›	‹syn›	‹etb›	‹can›	‹em›	‹sub›	‹esc›	‹fs›	‹gs›	‹rs›	‹us›
2x		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{	`	`	}	~
8x																
9x																
Ax	¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬		®	-	°
Bx	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿	
Cx	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
Dx	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
Ex	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
Fx	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

win-1252	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0x	‹nul›	‹soh›	‹stx›	‹etx›	‹eot›	‹enq›	‹ack›	‹bel›	‹bs›	‹tab›	‹lf›	‹vt›	‹ff›	‹cr›	‹so›	‹si›
1x	‹dle›	‹dc1›	‹dc2›	‹dc3›	‹dc4›	‹nak›	‹syn›	‹etb›	‹can›	‹em›	‹sub›	‹esc›	‹fs›	‹gs›	‹rs›	‹us›
2x		!	"	#	$	%	&	'	(	)	*	+	,	–	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	‹	=	›	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{	`	`	}	~
8x	€		‚	ƒ	„	…	†	‡	ˆ	‰	Š	‹	Œ		Ž	
9x		‘	’	“	”	•	–	—	˜	™	š	›	œ		ž	Ÿ
Ax	¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬		®	-	°
Bx	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿	
Cx	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
Dx	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
Ex	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
Fx	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

```
→ encode.py latin1,cp1252 -d cão,chó
✅ latin1:  Good cão [63 e3 6f] (3 bytes)
✅ cp1252:  Good cão [63 e3 6f] (3 bytes)
✅ latin1:  Good chó [63 68 f3] (3 bytes)
✅ cp1252:  Good chó [63 68 f3] (3 bytes)

→ encode.py latin1,cp1252 -d tz'i',šuo
❌ latin1:  Bad tz'i'
✅ cp1252:  Good tz'i' [74 7a 92 69 92] (5 bytes)
❌ latin1:  Bad šuo
✅ cp1252:  Good šuo [9a 75 6f] (3 bytes)
```

```
➜ encode.py cp1252 < dogs.txt
✅ cp1252: 64 good dogs
✅ cp1252: 286 chars encoded in 286 bytes, 1.0 bytes per char
❌ cp1252: 16 bad dogs:
❌ cp1252: köpeği  mbʉ  āso  āšun  كلب  כלב  אבו  कुत्ता
❌ cp1252: ฆ่า  犬  狗  ᛕИӨ  łééchąą’í  [illegible]  𓃡  🐶
```

windows-1250

latin2	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
0x	‹nul›	‹soh›	‹stx›	‹etx›	‹eot›	‹enq›	‹ack›	‹bel›	‹bs›	‹tab›	‹lf›	‹vt›	‹ff›	‹cr›	‹so›	‹si›
1x	‹dle›	‹dc1›	‹dc2›	‹dc3›	‹dc4›	‹nak›	‹syn›	‹etb›	‹can›	‹em›	‹sub›	‹esc›	‹fs›	‹gs›	‹rs›	‹us›
2x		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{	`	`	}	~
8x																
9x																
ax		Ą	˘	Ł	¤	Ľ	Ś	§	¨	Š	Ş	Ť	Ź		Ž	Ż
bx	°	ą	˛	ł	´	ľ	ś	ˇ	¸	š	ş	ť	ź	”	ž	ż
cx	Ŕ	Á	Â	Ă	Ä	Ĺ	Ć	Ç	Č	É	Ę	Ë	Ě	Í	Î	Ď
dx	Đ	Ń	Ň	Ó	Ô	Ő	Ö	×	Ř	Ů	Ú	Ű	Ü	Ý	Ţ	ß
ex	ŕ	á	â	ă	ä	Í	ć	ç	č	é	ę	ë	ě	í	î	ď
fx	đ	ń	ň	ó	ô	ő	ö	÷	ř	ů	ú	ű	ü	ý	ţ	˙

win-1250	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
0x	‹nul›	‹soh›	‹stx›	‹etx›	‹eot›	‹enq›	‹ack›	‹bel›	‹bs›	‹tab›	‹lf›	‹vt›	‹ff›	‹cr›	‹so›	‹si›
1x	‹dle›	‹dc1›	‹dc2›	‹dc3›	‹dc4›	‹nak›	‹syn›	‹etb›	‹can›	‹em›	‹sub›	‹esc›	‹fs›	‹gs›	‹rs›	‹us›
2x		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\\	]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{	`	`	}	~
8x	€	``	‚	``	„	…	†	‡	``	‰	Š	‹	Ś	Ť	Ž	Ź
9x	``	‘	’	“	”	•	–	—	``	™	š	›	ś	ť	ž	ź
ax		ˇ	˘	Ł	¤	Ą	¦	§	¨	©	Ş	«	¬		®	Ż
bx	°	±	˛	ł	´	µ	¶	·	¸	ą	ş	»	Ľ	”	l'	ż
cx	Ŕ	Á	Â	Ă	Ä	Ĺ	Ć	Ç	Č	É	Ę	Ë	Ě	Í	Î	Ď
dx	Đ	Ń	Ň	Ó	Ô	Ő	Ö	×	Ř	Ů	Ú	Ű	Ü	Ý	Ţ	ß
ex	ŕ	á	â	ă	ä	ĺ	ć	ç	č	é	ę	ë	ě	í	î	ď
fx	đ	ń	ň	ó	ô	ő	ö	÷	ř	ů	ú	ű	ü	ý	ţ	˙

```
→ encode.py latin2,cp1250 -d köpek
✅ latin2:  Good köpek [6b f6 70 65 6b] (5 bytes)
✅ cp1250:  Good köpek [6b f6 70 65 6b] (5 bytes)

→ encode.py latin2,cp1250 -d šuo
✅ latin2:  Good šuo [b9 75 6f] (3 bytes)
✅ cp1250:  Good šuo [9a 75 6f] (3 bytes)

→ encode.py latin2,cp1250 -d łééchąą’í
❌ latin2:  Bad łéechąą’í
✅ cp1250:  Good łéechąą’í [b3 e9 e9 63 68 b9 b9 92 ed] (9 bytes)
```

```
→ encode.py cp1250 < dogs.txt
✅ cp1250: 62 good dogs
✅ cp1250: 283 chars encoded in 283 bytes, 1.0 bytes per char
❌ cp1250: 18 bad dogs:
❌ cp1250: cão  klèb  køter  köpeği  mbʉ  āso  āšun  אבו
❌ cp1250: كلب  כלב  कुत्ता  ฆ่า  犬  狗  ᏦᏂᎾ  𐤊𐤋𐤁
❌ cp1250: 🐕‍🦺  🐶
```

Standard	Alias	Windows Equivalent
iso-8859-1	latin1	windows-1252
iso-8859-2	latin2	windows-1250
iso-8859-5	cyrillic	windows-1251
iso-8859-6	arabic	windows-1256
iso-8859-7	greek	windows-1253
iso-8859-8	hebrew	windows-1255
iso-8859-9	latin5	windows-1254
iso-8859-11	thai	windows-874
iso-8859-13	latin7	windows-1257
iso-8859-15	latin9	windows-1252

gb2312

```
→ encode.py ascii,gb2312 dog
✅ ascii:   "dog"=[64 6f 67]
✅ gb2312:  "dog"=[64 6f 67]

→ encode.py gb2312 -d 犬,狗
✅ gb2312:  Good 犬 [c8 ae] (2 bytes)
✅ gb2312:  Good 狗 [b9 b7] (2 bytes)
```

```
→ encode.py gb2312 < dogs.txt
✅ gb2312: 60 good dogs
✅ gb2312: 258 chars encoded in 266 bytes, 1.0 bytes per char
❌ gb2312: 20 bad dogs:
❌ gb2312: cão  câin  câine  câini  køter  köpek  köpeği  mbʉ
❌ gb2312: āšun  šuo  كلب  כלב  אבו  कुत्ता  ໝາ  ΚИΘ
❌ gb2312: łééchąą’í  ⩎  𓃡  🐶
```

犬

狗


犬


福
狗

```
[?? ?? ?? ?? ?? ??] (6 bytes)

→ encode.py gb2312 doggie
✅ gb2312:  "doggie"=[64 6f 67 67 69 65]

→ encode.py gb2312 狗狗狗
✅ gb2312:  "狗犬狗"=[b9 b7 b9 b7 b9 b7]

→ encode.py gb2312 肮发
✅ gb2312:  "肮发"=[b0 b9 b7 a2]
```

Oracle 6

- Selectable charset
  - **WE8ISO8859P1**
  - …

Oracle 7

- NLS framework introduced
- New Encodings
  - `WE8ISO8859P2`
  - `WE8MSWIN1252`
  - `ZHS16CGB231280`
  - `...`

Unicode 1.0

Codepoint

U+0000→U+FFFF

U+	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
000x	‹nul›	‹soh›	‹stx›	‹etx›	‹eot›	‹enq›	‹ack›	‹bel›	‹bs›	‹tab›	‹lf›	‹vt›	‹ff›	‹cr›	‹so›	‹si›
001x	‹dle›	‹dc1›	‹dc2›	‹dc3›	‹dc4›	‹nak›	‹syn›	‹etb›	‹can›	‹em›	‹sub›	‹esc›	‹fs›	‹gs›	‹rs›	‹us›
002x		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
003x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
004x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
005x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
006x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
007x	p	q	r	s	t	u	v	w	x	y	z	{	¦	}	~	

Range	Block
U+0000→U+007F	ASCII
U+0080→U+00FF	Latin Supplement
U+0100→U+024F	Latin Extended
U+0250→U+02FF	Phonetic Symbols
U+0300→U+05FF	Greek, Cyrillic & Hebrew
U+0600→U+0FFF	Arabic & Indian
U+1000→U+17FF	S & SE Asian
U+1800→U+24FF	E Asian
U+2000→U+27FF	Symbols & Punctuation
U+2800→U+28FF	Braille & Basic Shapes
U+2E80→U+9FFF	Chinese
U+AC00→U+D7AF	Korean
U+D800→U+DBFF	High Surrogates
U+DC00→U+DFFF	Low Surrogates
U+E000→U+F8FF	Private Use

Unicode Encoding UCS-2

```
"Good 狗"
=          U+0047 U+006f U+006f U+0064 U+0020 U+72d7
→         [00  47 00  6f 00  6f 00  64 00  20 72  d7] (LE)
→         [47  00 6f  00 6f  00 64  00 20  00 d7  72] (BE)
→ [ff fe 47  00 6f  00 6f  00 64  00 20  00 72  d7] (LE + BOM)
→ [fe ff 00  47 00  6f 00  6f 00  64 00  20 72  d7] (BE + BOM)
```

```
→ encode.py ucs-2le < dogs.txt
✅ ucs-2le: 77 good dogs
✅ ucs-2le: 332 chars encoded in 664 bytes, 2.0 bytes per char
❌ ucs-2le: 3 bad dogs:
❌ ucs-2le: 🦯  🐶
```

major languages modern world

U+0000→U+FFFF

all languages

U+0000→U+10FFFF

Plane	Range	Plane Name	Abbreviation
0	U+0000→U+FFFF	Basic Multilingual	BMP
1	U+10000→U+1FFFF	Supplementary Multilingual	SMP
2	U+20000→U+2FFFF	Supplementary Ideographic	SIP
3	U+30000→U+3FFFF	Tertiary Ideographic Plane	TIP
4→13	U+40000→U+DFFFF	Reserved	
14	U+E0000→U+EFFFF	Supplementary Special-Purpose	SSP
15	U+F0000→U+FFFFF	Private Use Area A	
16	U+100000→U+10FFFF	Private Use Area B	

Unicode Encoding UTF-8

U+0000→U+007F
→[0xxxxxxx]

U+0080→U+07FF
→[110xxxxx 10xxxxxx]

U+0800→U+FFFF
→[1110xxxx 10xxxxxx 10xxxxxx]

U+10000→U+10FFFF
→[11110xxx 10xxxxxx 10xxxxxx 10xxxxxx]

```
→ encode.py utf-8 Ł
utf-8:   "Ł"=[c5 81]

U+0141 ('Ł')
 =  00101000001
 =     00101   000001
    110xxxxx 10xxxxxx
 → [11000101 10000001]
    =    [c5       81]
```

```
[0xxxxxxx]
[110xxxxx 10xxxxxx]
[1110xxxx 10xxxxxx 10xxxxxx]
[11110xxx 10xxxxxx 10xxxxxx 10xxxxxx]

[0xxxxxxx] ASCII
[10xxxxxx] Continuation
[110xxxxx] Lead of 2 byte sequence
[1110xxxx] Lead of 3 byte sequence
[11110xxx] Lead of 4 byte sequence

[...10100011 10100011 11000011 10100011 01010101...]
```

```
→ encode.py ascii,utf8 -d pies
✅ ascii:    Good pies [70 69 65 73] (4 bytes)
✅ utf8:     Good pies [70 69 65 73] (4 bytes)

→ encode.py utf-8 -d cão,狗,🐶
✅ utf-8:    Good cão [63 c3 a3 6f] (4 bytes)
✅ utf-8:    Good 狗 [e7 8b 97] (3 bytes)
✅ utf-8:    Good 🐶 [f0 9f 90 b6] (4 bytes)
```

```
→ encode.py utf-8 < dogs.txt
✅ utf-8: 80 good dogs
✅ utf-8: 338 chars encoded in 413 bytes, 1.2 bytes per char
✅  No bad dogs
```

Unicode Encoding
UTF-16

```
"Good 狗"
=  U+0047 U+006f U+006f U+0064 U+0020 U+72d7
→ [00  47 00  6f 00  6f 00  64 00  20 72  d7]

"Good 🐶"
=  U+0047 U+006f U+006f U+0064 U+0020 U+1f436
→ [00  47 00  6f 00  6f 00  64 00  20    ?   ]
```

Range in BMP	Block
U+0000→U+007F	ASCII (Basic Latin)
U+0080→U+00FF	Latin Supplement
U+0100→U+024F	Latin Extended
U+0250→U+02FF	Phonetic Symbols
U+0300→U+05FF	Greek, Cyrillic, Hebrew
U+0600→U+0FFF	Arabic & Indian
U+1000→U+17FF	S & SE Asian
U+1800→U+24FF	E Asian
U+2000→U+27FF	Symbols & Punctuation
U+2800→U+28FF	Braille & Basic Shapes
U+2E80→U+9FFF	Chinese
U+AC00→U+D7AF	Korean
U+D800→U+DBFF	**High Surrogates**
U+DC00→U+DFFF	**Low Surrogates**
U+E000→U+F8FF	Private Use

```
→ encode.py utf-16-be -d 🐶
✅ utf-16-be: Good 🐶 [d8 3d dc 36] (4 bytes)

🐶 = U+1f436 - 0x10000 = 0x0f436
                       = 0b0000111101000000110110
                       = [0b0000111101][0b0000110110]
                       = [0x003d]      [0x0036]
                         +[0xd800]     +[0xdc00]
                       = [0xd83d]      [0xdc36]

   = U+d83d U+dc36
   → [d8 3d dc 36]
```

```
→ encode.py utf-16be < dogs.txt
✅ utf-16be: 80 good dogs
✅ utf-16be: 338 chars encoded in 682 bytes, 2.0 bytes per char
✅  No bad dogs
```

UTF-32

```
"Good 🐶"
=  U+0047      U+006f      U+006f      U+0064      U+0020      U+1f436
→ [00 00 00 47 00 00 00 6f 00 00 00 6f 00 00 00 64 00 00 00 20 00 01 f4 36]
```

```
→ encode.py utf-32be,utf-32le,utf-32 "Good 🐶"
✅ utf-32be: "Good 🐶"=[00 00 00 47 00 00 00 6f 00 00 00 6f 00 00 00 64 00 00 00 20 00 01 f4 36]
✅ utf-32le: "Good 🐶"=[47 00 00 00 6f 00 00 00 6f 00 00 00 64 00 00 00 20 00 00 00 36 f4 01 00]
✅ utf-32:  "Good 🐶"=[ff fe 00 00 47 00 00 00 6f 00 00 00 6f 00 00 00 64 00 00 00 20 00 00 00 36 f4 01 00]
```

```
→ encode.py utf-32be < dogs.txt
✅ utf-32be: 80 good dogs
✅ utf-32be: 338 chars encoded in 1352 bytes, 4.0 bytes per char
✅  No bad dogs
```

Encoding	Good Dogs	Chars	Bytes	Bytes per Char
ascii	53	237	237	1.0
latin1	62	278	278	1.0
latin2	60	269	269	1.0
win-1252	64	286	286	1.0
win-1250	62	283	283	1.0
gb2312	60	261	269	1.0
ucs-2	77	332	664	2.0
utf-8	80	338	413	1.2
utf-16	80	338	682	2.0
utf-32	80	338	1352	4.0

Encoding	Bytes per Character	BMP	All Planes	Self Synchronizing
ascii	1	❌	❌	
latin1	1	❌	❌	
latin2	1	❌	❌	
win-1252	1	❌	❌	
win-1250	1	❌	❌	
gb2312	1,2	❌	❌	❌
ucs-2	2	✅	❌	
utf-8	1→4	✅	✅	✅
utf-16	2,4	✅	✅	
utf-32	4	✅	✅	

Oracle 8

- New Encodings
  - `EE8MSWIN1250`
  - `UTF8` (≠ `UTF-8`)
  - `...`

UTF8 (≠ UTF-8)

U+0000→U+007F
[0xxxxxxx]

U+0080→U+07FF
[110xxxxx 10xxxxxx]

U+0800→U+FFFF
[1110xxxx 10xxxxxx 10xxxxxx]

~~U+10000→U+10FFFF~~
~~[11110xxx–10xxxxxx–10xxxxxx–10xxxxxx]~~

Oracle 8i

- National Character Set
- Character Sets
  - `UTF8` (Beyond BMP)
  - `AL16UTF16`

UTF8 (Improved?)

U+0000→U+007F
[0xxxxxxx]

U+0080→U+07FF
[110xxxxx 10xxxxxx]

U+0800→U+FFFF
[1110xxxx 10xxxxxx 10xxxxxx]

~~U+10000→U+10FFFF~~
~~[11110xxx 10xxxxxx 10xxxxxx 10xxxxxx]~~

🐶 = U+1f436

→ U+d83d U+dc36

→ [ed a0 bd ed b0 b6]

Oracle 9i

- Character Sets
  - `AL32UTF8`
  - `...`
- Length Semantics

National Character Set

- `AL16UTF16/UTF8`
- `NCHAR,NVARCHAR2,NCLOB`

(Doc ID 276914.1)


AL32UTF8
THIS IS THE ENCODING YOU ARE LOOKING FOR
imgflip.com

Byte/Char Semantics

```
NLS_LENGTH_SEMANTICS = BYTE (default)| CHAR
```

```
SQL> create table doggie(name nvarchar2(1 char));

ORA-00907: missing right parenthesis

SQL> create table doggie(name nvarchar2(1))

SQL> Table DOGGIE created.

SQL> insert into doggie values ('🐶');

SQL Error: ORA-12899: value too large for column "PDBADMIN"."DOGGIE"."NAME" (actual: 2, maximum: 1)
```

```
SQL> select length(n'🐶'), rawtohex(n'🐶');

   LENGTH(U'\D83D\DC36') RAWTOHEX(U'\D83D\DC36')
------------------------ ----------------------------
                       2 D83DDC36
```

```
SQL> create table longdog (name varchar2(4001 byte));

ORA-00910: specified length too long for its datatype

SQL> create table longdog (name varchar2(4000 char));

Table LONGDOG created.

SQL> insert into longdog values (rpad('🐶', 4000, '🐶'));

1 row inserted.

SQL> select length(name), lengthb(name) from longdog;

   LENGTH(NAME)    LENGTHB(NAME)
--------------- ----------------
           1000             4000
```

Oracle 11g

» `AL32UTF8` recommended

Oracle 12c

- `AL32UTF8` default
- Database Migration Assistant for Unicode (DMU)