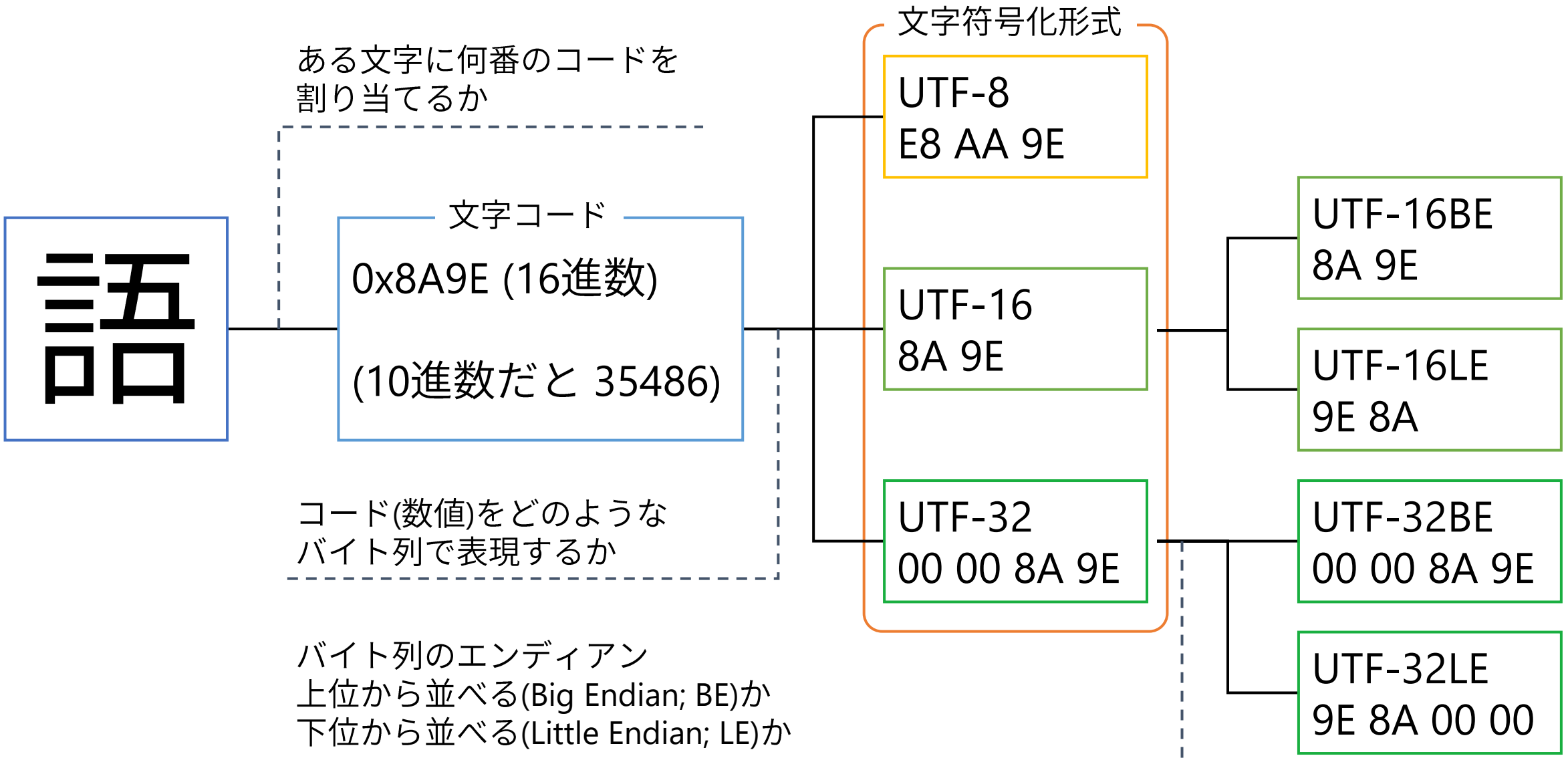
日本語の扱いと文字コード

- Pythonでは文字を表現する文字コードとしてUnicodeを使用．
  - Unicodeはもともと16ビット(65536通り)のコードで世界中の全ての文字を表そうとしたが，不足することが明らかになり，現在は符号として0x000000-0x10FFFFのおよそ110万文字分が確保されている．
  - Unicodeの文字符号化形式にはUTF-8, UTF-16, UTF-32の3種類がる．
- ファイルなど外部のデータは，様々な文字コードが使用されているため，外部入出力の際には文字セット名(character set name)を指定する．

文字コードと文字符号化形式



UnicodeとUTF-xxの関係 (まとめ)

- Unicodeの文字集合では，それぞれの文字に1つの数値(非負整数)が対応している(下の表ではUTF-32の欄にある16進数の数値).
- この数値をコンピュータ内で保持する方法(文字符号化形式)が複数存在する．UTF-8, UTF-16, UTF-32 はそれぞれ8ビット，16ビット，32ビットを単位とした符号化形式である.

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
UTF-8	A	Ω		語			😊									
	41	CE	A9	E8	AA	9E	F0	9F	98	8A						
UTF-16	A		Ω		語		😊									
	41		03A9		8A9E		D83D		DE0A							
UTF-32	A				Ω				語				😊			
	41				000003A9				00008A9E				0001F60A			

wikipediaより