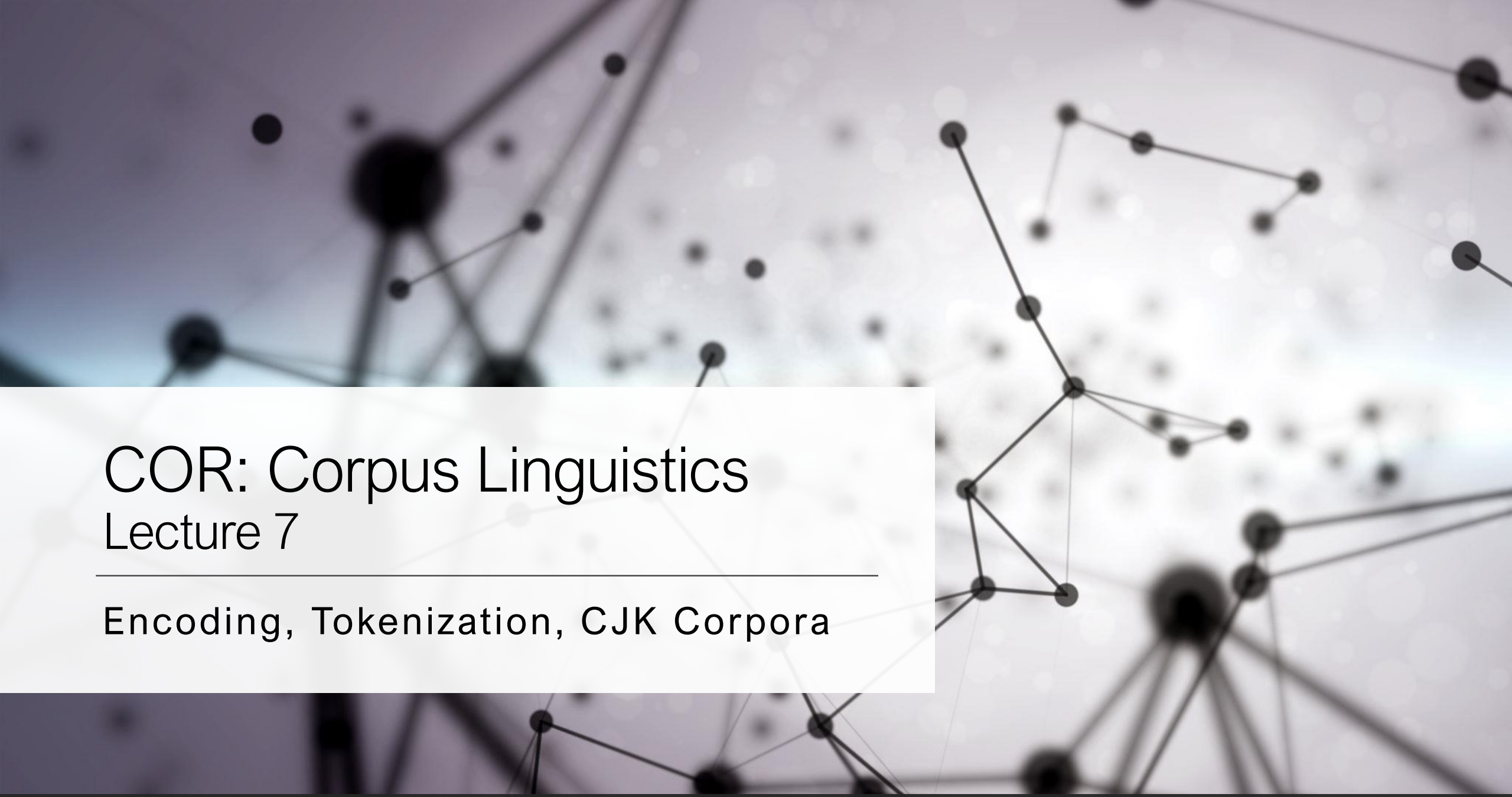


COR: Corpus Linguistics
Lecture 7

Encoding, Tokenization, CJK Corpora

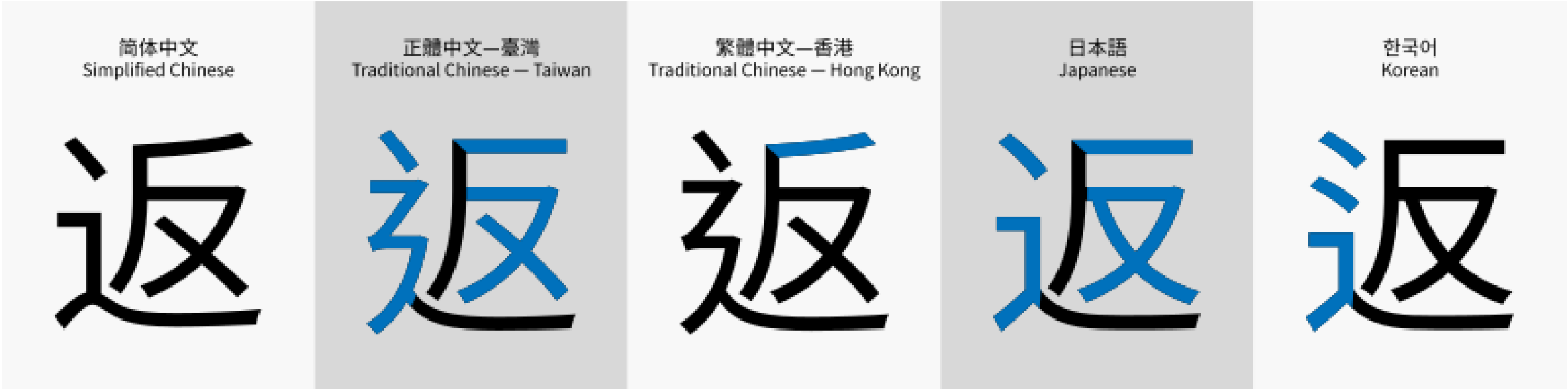
Martina Jemelková | 25.3.2024 | Department of Asian Studies, Palacký University Olomouc

Overview

- Encoding
- Introduction to the Unicode Standard for East Asian Scripts
- Character DIY
- Tokenization
- CJK Corpora + Czech National Corpus

Encoding

- assigning numerical code, i.e., code point, to written characters of any language
- CJK languages – Chinese, Japanese, Korean, i.e., languages that use Chinese characters
- different codepoints to regional variants (UTF-16: 说 8BF4; 說 8AAA)
- **Unicode** (+ Universal Character set – Han unification process)
- Mainland China, Singapore: National Standard – Guobiao (GB)
- Taiwan, Hong Kong, Macau: Big5


简体中文
Simplified Chinese
返
正體中文—臺灣
Traditional Chinese — Taiwan
返
繁體中文—香港
Traditional Chinese — Hong Kong
返
日本語
Japanese
返
한국어
Korean
返

Chinese Telegraph Code

- first codebook originates 1871
- 1 Chinese character = 4 digit numerical code (0000–9999)
- input method for computers
- codebook is arranged according to the character‘s radical and number of strokes
- *Standard Telegraph Codebook* (1st ed. 1983) – 7,000 simplified Chinese characters + Zhuyin, Latin alphabet, digits, special symbols

3500	3501	3502	3503	3504	3505	3506	3507	3508	3509
灰	灶	灸	灿	灼		灾	炊	炎	炒
3510	3511	3512	3513	3514	3515	3516	3517	3518	3519
炕	炙	炘	炸	炫	炬	炭	炮	炯	
3520	3521	3522	3523	3524	3525	3526	3527	3528	3529
炱	炳	炷		畑	烈		乌		烙
3530	3531	3532	3533	3534	3535	3536	3537	3538	3539
烘	烜	烝	烟	烹	烺	烽	焌	焙	焚
3540	3541	3542	3543	3544	3545	3546	3547	3548	3549
焜		焦	焰	然		焠	焉	煅	焊
3550	3551	3552	3553	3554	3555	3556	3557	3558	3559
炼	煊	煌	煎	煮	炜	熙		煜	煞
3560	3561	3562	3563	3564	3565	3566	3567	3568	3569
茕	煤	煥	煦	照	煩	煨		炀	煽
3570	3571	3572	3573	3574	3575	3576	3577	3578	3579
	熄			熊	熏	荧		熟	熔
3580	3581	3582	3583	3584	3585	3586	3587	3588	3589
熨	熬		热	熘	颍			熹	炽
3590	3591	3592	3593	3594	3595	3596	3597	3598	3599
				烫	燃		灯	燎	烧

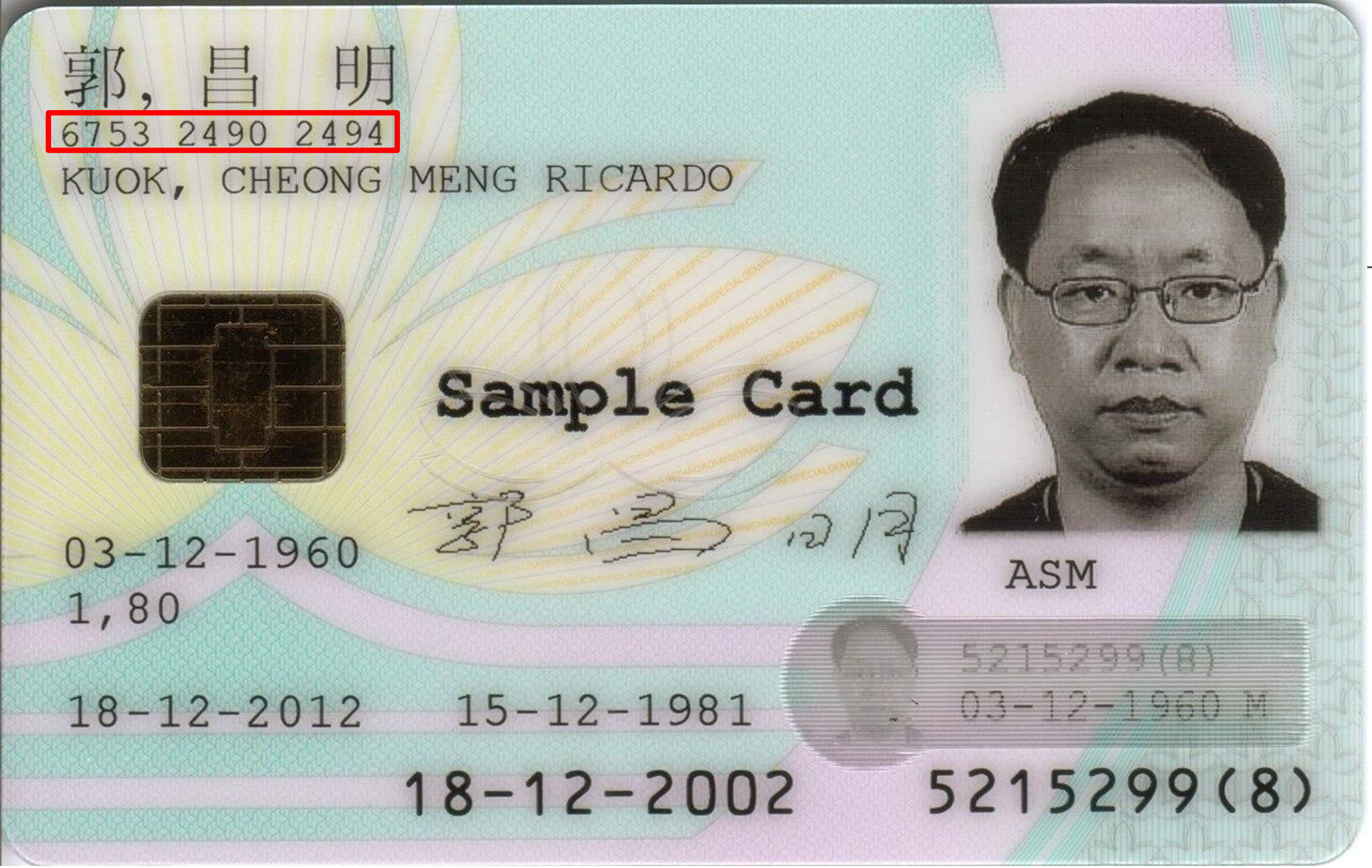
0100	0101	0102	0103	0104	0105	0106	0107	0108	0109
他	仗	付	仙	仝	仞	仡	仟	代	令
0110	0111	0112	0113	0114	0115	0116	0117	0118	0119
以	仰	仲	仳	仵	件	价	任	份	仿
0120	0121	0122	0123	0124	0125	0126	0127	0128	0129
企	伉	伊	伋	伍	伎	伏	伐	休	伙
0130	0131	0132	0133	0134	0135	0136	0137	0138	0139
伯	估	你	伴	伶	伸	伺	伭	似	伱
0140	0141	0142	0143	0144	0145	0146	0147	0148	0149
佃	但		位	低	住	佐	佑		何
0150	0151	0152	0153	0154	0155	0156	0157	0158	0159
佗	余	佘	佚	佛	作	佞	佟	佺	
0160	0161	0162	0163	0164	0165	0166	0167	0168	0169
佩	佇	佯	佳		佶	佝	佻	佾	使
0170	0171	0172	0173	0174	0175	0176	0177	0178	0179
侃	来	侈	例	侍	侏	侑	侔	仑	侗
0180	0181	0182	0183	0184	0185	0186	0187	0188	0189
供	依		伕	佰	侮	侯	侵	侣	便
0190	0191	0192	0193	0194	0195	0196	0197	0198	0199
	促	俄	俊	俎	俏	俐	俑	俗	俘


香港永久性居民身份證
HONG KONG PERMANENT IDENTITY CARD
李智能
LEE, Chi Nan
樣本 SAMPLE
2621 2535 5174
出生日期 Date of Birth
01-01-1968
女 F
***AZ
簽發日期 Date of Issue
(01-79)
15-09-03
C668668(E)


2621
李
DWV


2535
智
DTN


5174
能



6753
郭

2490
昌

2494
明

Guobiao GB (国家标准 Guójiā Biāozhǔn)

- first set: 1980 (GB 2312); last set: 2005 (GB18030-2005)
- compatible with Unicode
- originally designed for simplified Chinese characters
- latest subset also includes characters used by ethnic minorities (e.g., Tibet, Mongolia)

Unicode

- covers most of world‘s scripts (+ currency symbols, punctuation marks, mathematic and technical symbols, dingbats, and emojis)
- parallel standard: ISO/IEC 10646
- code charts, character database, annexes
- general principles, requirements for conformance, guidelines for implementers – text processing, normalization, encoding forms
- Unicode 15.1.0 (December 2023): 149 813 characters
  - Han script 97,058 ideographic characters (China, Taiwan, Japan, Korea, Vietnam, Singapore)

ode 15.1 Character Code Charts

East Asian Scripts

Bopomofo
Bopomofo Extended
CJK Unified Ideographs (Han) (38MB)
CJK Extension A (7.6MB)
CJK Extension B (31MB)
CJK Extension C
CJK Extension D
CJK Extension E
CJK Extension F
CJK Extension G
CJK Extension H
CJK Extension I
(see also Unihan Database)
CJK Compatibility Ideographs
CJK Compatibility Ideographs Supplement
CJK Radicals / Kangxi Radicals
CJK Radicals Supplement
CJK Strokes
Ideographic Description Characters
Hangul Jamo
Hangul Jamo Extended-A
Hangul Jamo Extended-B
Hangul Compatibility Jamo
Halfwidth Jamo
Hangul Syllables

Hiragana
Kana Extended-A
Kana Extended-B
Kana Supplement
Small Kana Extension
Kanbun
Katakana
Katakana Phonetic Extensions
Halfwidth Katakana
Khitan Small Script
Lisu
Lisu Supplement
Miao
Nushu
Tangut
Tangut Components
Tangut Supplement
Yi
Yi Syllables
Yi Radicals

Meetei Mayek
Meetei Mayek Extensions
Modi
Mro
Multani

sian & Philippine Scripts
nese Supplement
ian Scripts
ofo
nofo Extended
ified Ideographs (Han) (38MB)
xtension A (7.6MB)
xtension B (31MB)
xtension C
xtension D
xtension E
xtension F
CJK Extension G
CJK Extension H
CJK Extension I
(see also Unihan Database)
CJK Compatibility Ideographs

Latin Extended-D

Unicode map

- total capacity: 1 114 112 characters
- **Private Use Area** (**PUA**)
  - range: E000–F8FF
  - a space for your own characters/symbols/logos
  - cannot be easily reproduced or searched via network (i.e., view, print, read aloud)
  - decorative function; associated with particular font
  - BUT! through closed proprietary file format (.pdf) are replicable

Private Use
Private Use Area
Supplementary Private Use Area-A
Supplementary Private Use Area-B

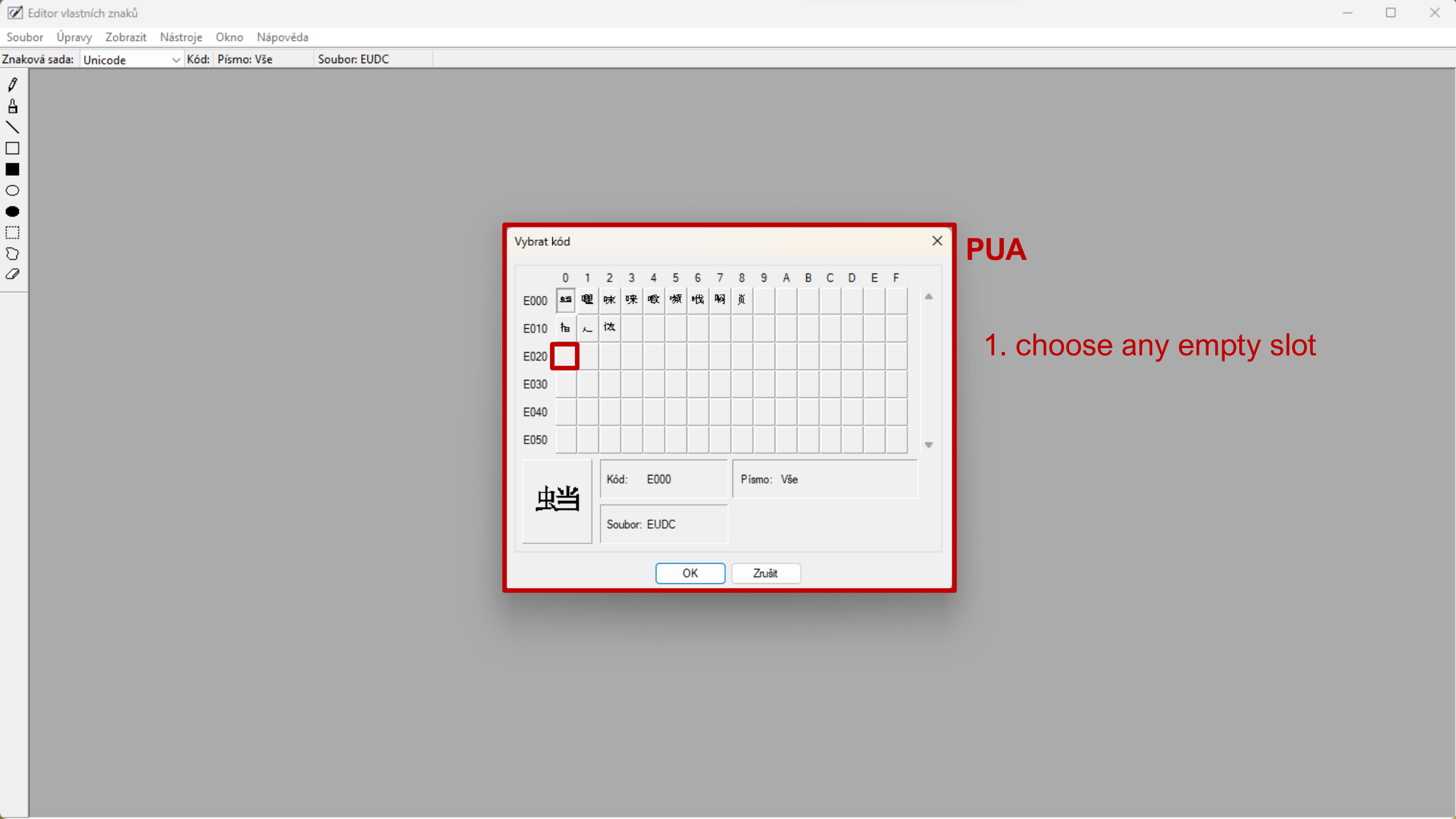
Character DIY (Windows only)

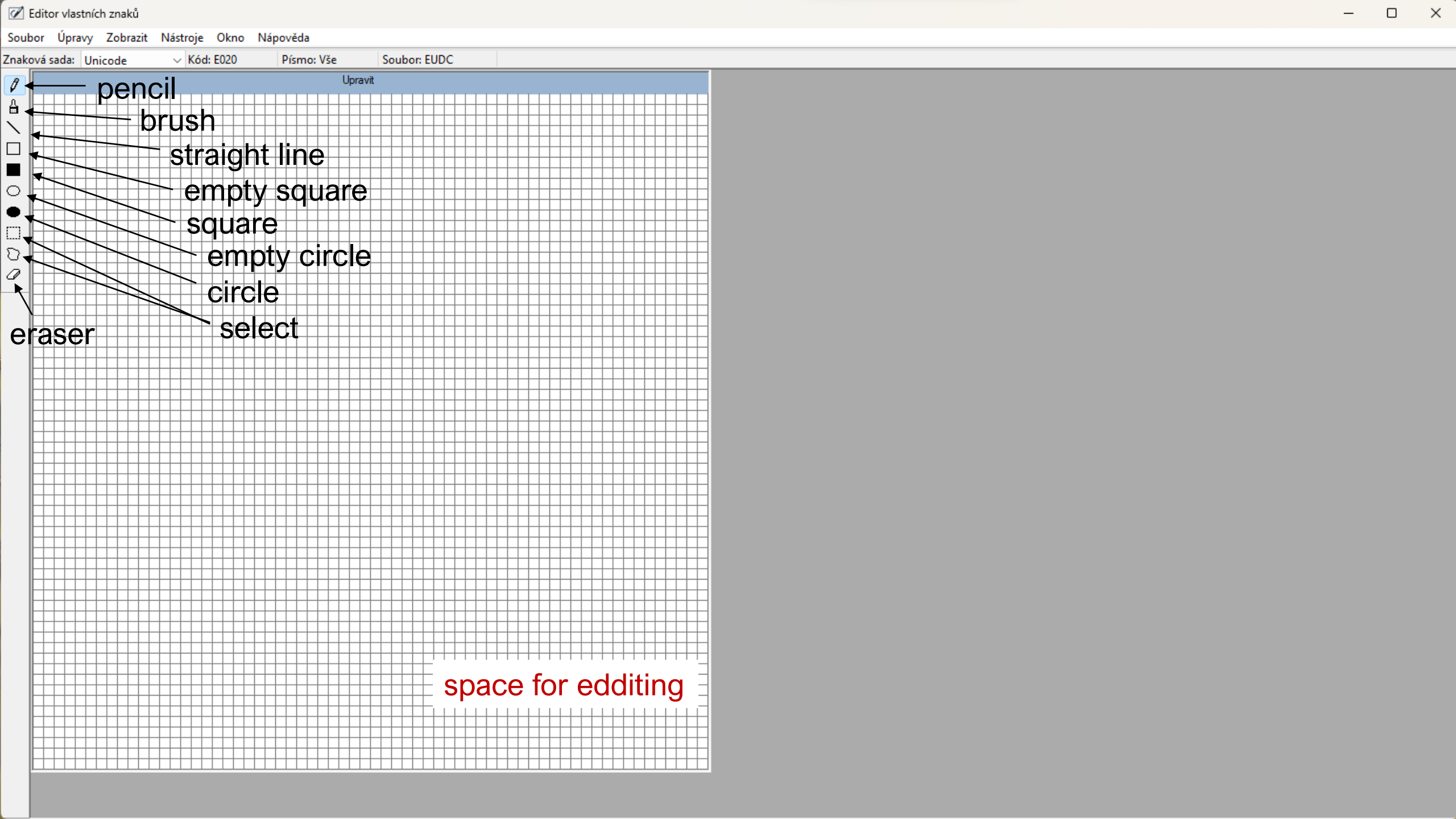
- find and open: eudcedit.exe (Private Character Editor)
  - path: Windows\System32 or Windows Run-Command

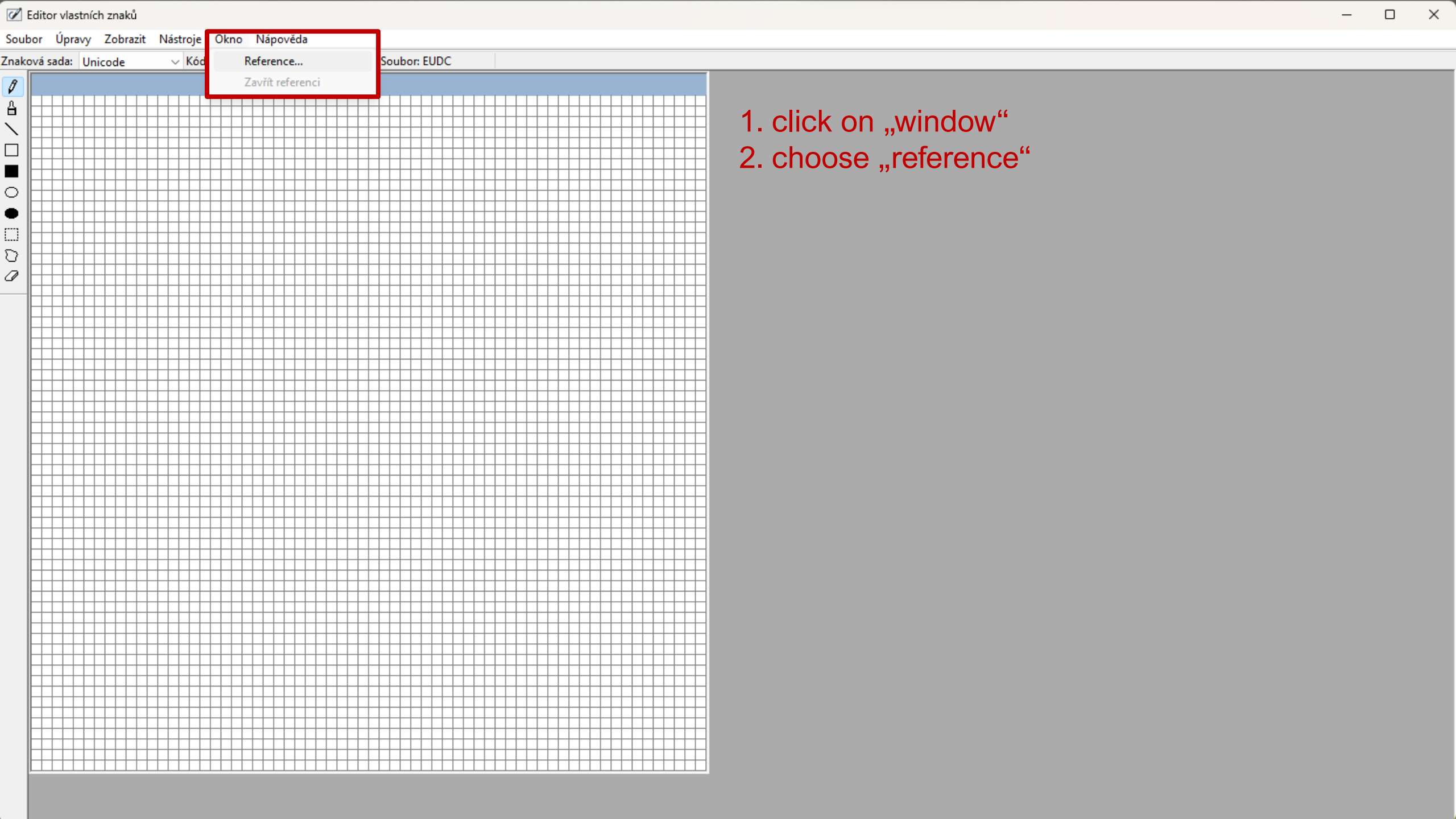
谖	蛘	缏	辖	颂	钼	颧	辗	筹	昢
chù X_1	dāng X_2	fán X_3	gé X_4	hòng X_5	jǔ X_6	kǎn X_7	kǎn X_8	kǎo X_9	kē X_{10}

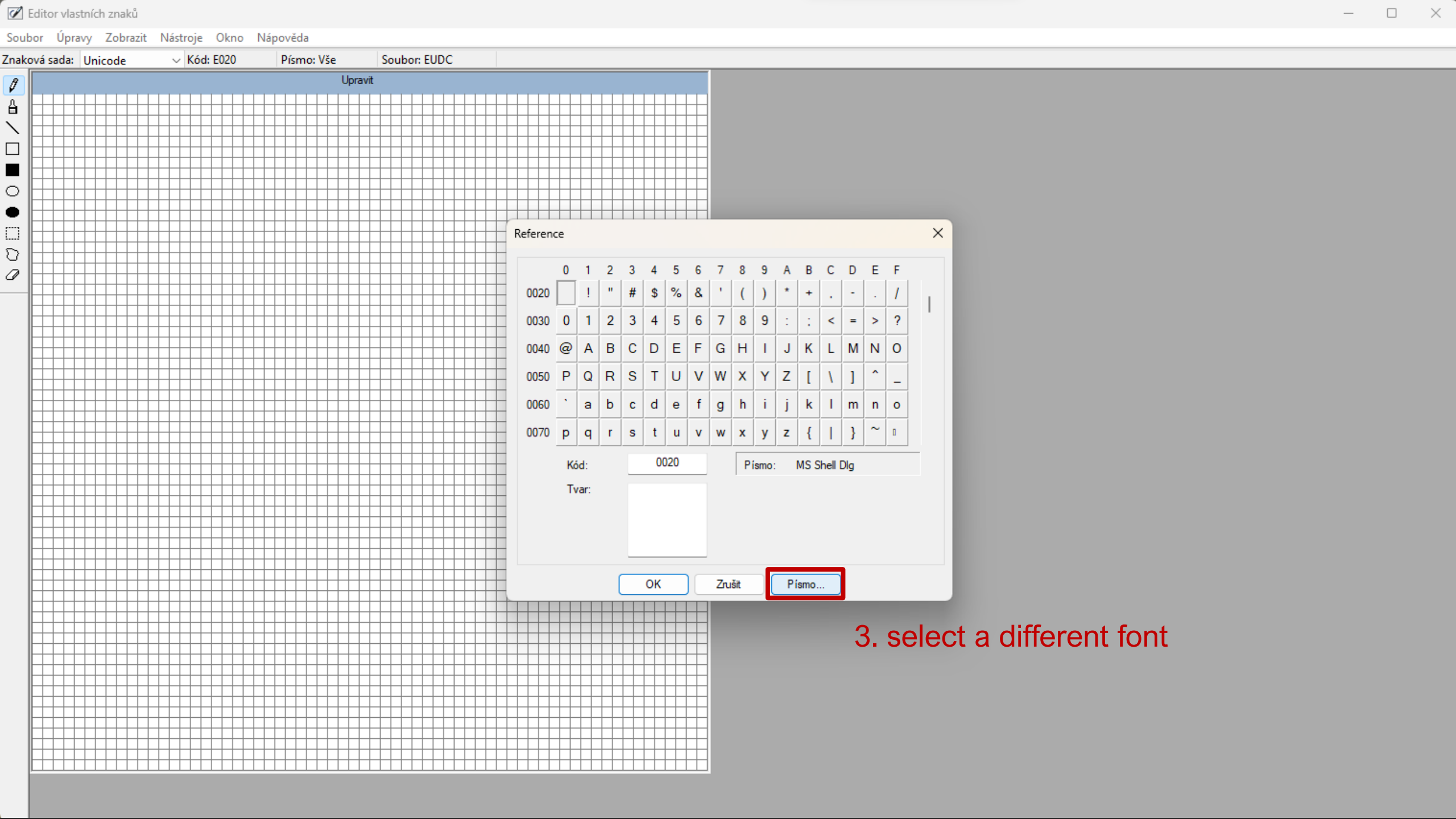
倥	旮	箩	侬	恢	隈	椹	髹	落	蹡
kuāng X_{11}	lā X_{12}	láo X_{13}	lǒng X_{14}	náo X_{15}	nì X_{16}	nì X_{17}	níng X_{18}	qià X_{19}	qiāng X_{20a}

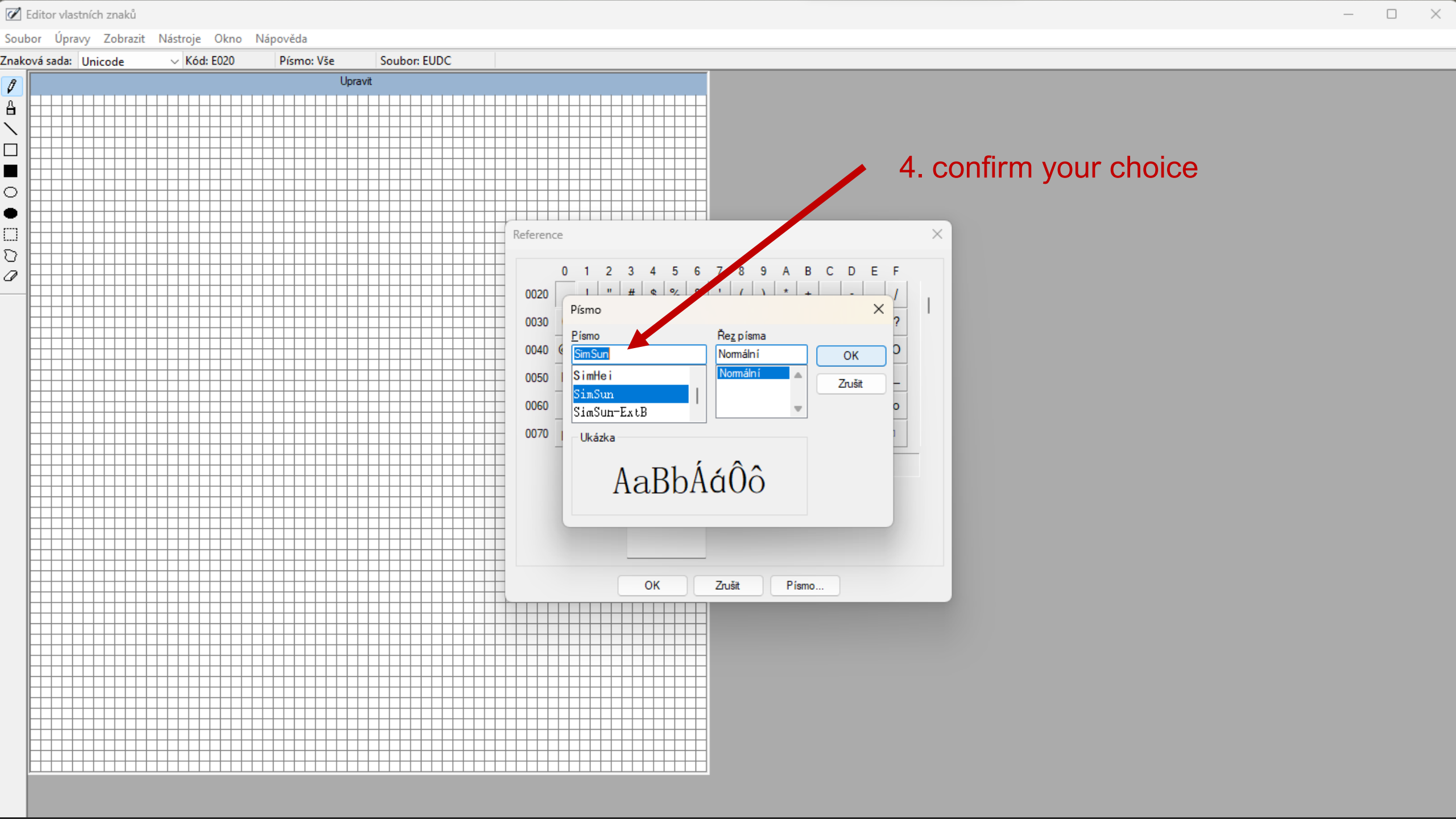
蹡	鹳	簏	霰	诡	隩	夔			
qiàng X_{20b}	qú X_{21}	sī X_{22}	sù X_{23}	yí X_{24}	yǐn X_{25}	yīng X_{26}			

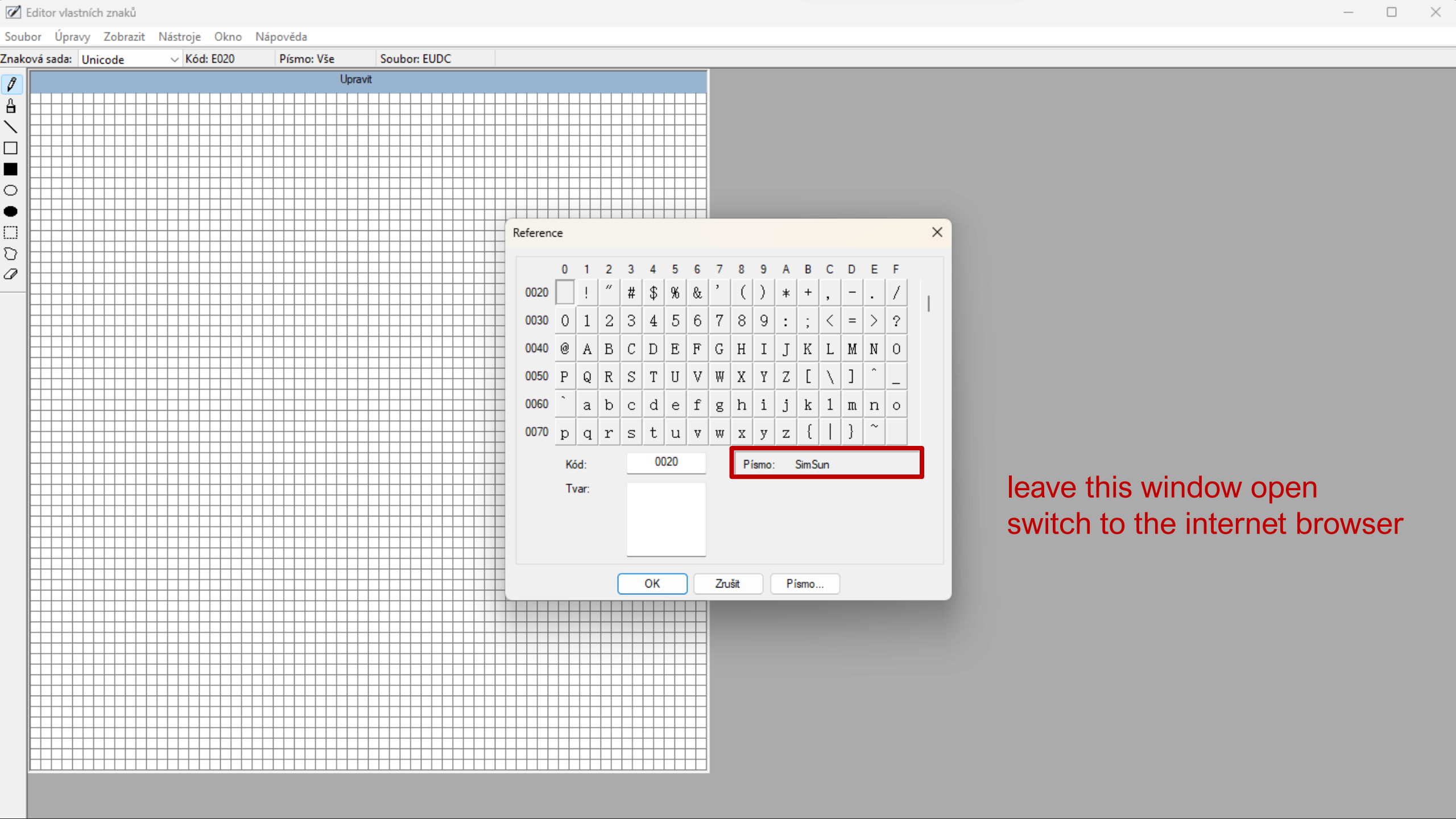

Editor vlastních znaků
Soubor Úpravy Zobrazit Nástroje Okno Nápověda
Znaková sada: Unicode
Kód:
Písmo: Vše
Soubor: EUDC
Vybrat kód
0 1 2 3 4 5 6 7 8 9 A B C D E F
E000
E010
E020
E030
E040
E050
Kód: E000
Písmo: Vše
Soubor: EUDC
OK
Zrušit
PUA
1. choose any empty slot


Editor vlastních znaků
Soubor
Úpravy
Zobrazit
Nástroje
Okno
Nápověda
Znaková sada:
Unicode
Kód: E020
Písmo: Vše
Soubor: EUDC
Upravit
pencil
brush
straight line
empty square
square
empty circle
circle
select
eraser
space for edditing


Editor vlastních znaků
Soubor
Úpravy
Zobrazit
Nástroje
Okno
Nápověda
Reference...
Zavřít referenci
Znaková sada:
Unicode
Soubor: EUDC
1. click on „window“
2. choose „reference“


Editor vlastních znaků
Soubor Úpravy Zobrazit Nástroje Okno Nápověda
Znaková sada: Unicode
Kód: E020
Písmo: Vše
Soubor: EUDC
Upravit
Reference
0 1 2 3 4 5 6 7 8 9 A B C D E F
0020 ! " # $ % & ' ( ) * + . - . /
0030 0 1 2 3 4 5 6 7 8 9 : ; < = > ?
0040 @ A B C D E F G H I J K L M N O
0050 P Q R S T U V W X Y Z [\ ] ^ _
0060 ` a b c d e f g h i j k l m n o
0070 p q r s t u v w x y z { | } ~
Kód: 0020
Písmo: MS Shell Dlg
Tvar:
OK
Zrušit
Písmo...
3. select a different font


Editor vlastních znaků
Soubor Úpravy Zobrazit Nástroje Okno Nápověda
Znaková sada: Unicode
Kód: E020
Písmo: Vše
Soubor: EUDC
Upravit
4. confirm your choice
Reference
Písmo
Písmo
SimSun
SimHei
SimSun
SimSun-ExtB
Řez písma
Normální
Normální
OK
Zrušit
Ukázka
AaBbÁáÔô
OK
Zrušit
Písmo...


Editor vlastních znaků
Soubor Úpravy Zobrazit Nástroje Okno Nápověda
Znaková sada: Unicode
Kód: E020
Písmo: Vše
Soubor: EUDC
Upravit
Reference
0 1 2 3 4 5 6 7 8 9 A B C D E F
0020 ! " # $ % & ' ( ) * + , - . /
0030 0 1 2 3 4 5 6 7 8 9 : ; < = > ?
0040 @ A B C D E F G H I J K L M N O
0050 P Q R S T U V W X Y Z [\ ] ^ _
0060 ` a b c d e f g h i j k l m n o
0070 p q r s t u v w x y z { | } ~
Kód: 0020
Písmo: SimSun
Tvar:
OK
Zrušit
Písmo...
leave this window open
switch to the internet browser

https://www.unicode.org/charts/unihan.html

Unihan Database

Tech Site | Site Map | Search

Unihan Database Lookup

Lookup

5. type character you want to edit
6. press Lookup

About the Unihan Database Lookup Tool

The lookup interface on this page provides online access to property data in the Unicode Han (Unihan) database for individual ideographs via the “Lookup” button and text field above. Simply enter the four- or five-digit hexadecimal code point for the desired ideograph into the text field, or copy and paste the ideograph into it, then click the “Lookup” button. The resulting data set will contain various types of information available in the Unihan database, such as mappings to legacy encoding standards, references to dictionaries, meaning and reading information according to various authorities, links to other websites, and so on.

If you do not know the code point of the ideograph, or have no example of the ideograph to copy, the Unihan Search Page supports queries against several properties, such as those for ideograph readings. The following two indices are also available:

- A Grid Index that groups ideographs into blocks of 256 code points
- A Radical-Stroke Index

For production reasons, the version of the Unihan database that is available on this page may not yet be synchronized with the latest version of the Unicode Standard. For access to the latest version of the data files that comprise the Unihan database, download Unihan.zip from https://www.unicode.org/Public/UCD/latest/ucd/.

The Unihan database and its properties are documented in UAX #38.

Unihan Code Charts and Indices

The Unihan Radical-Stroke (RS) indices, which are documented in the

- Full RS Index
- IICore RS Index
- UnihanCore2020 RS Index

Code charts covering all of Unihan are available in PDF format, linked

https://www.unicode.org/charts/unihan.html

Disclaimers

The Unihan database is provided as-is as a public service by Unicode, Inc. (The Unicode Consortium). No claims are made as to its appropriateness for any particular purpose, and no warranties of any kind are expressed or implied.

https://www.unicode.org/cgi-bin/GetUnihanData.pl?codepoint=龙

Unihan Database

Unihan data for U+9F99

Lookup

Grid Index
<<< Previous

Radical-Stroke index (212.0-1)
Next >>>

Glyph

龙

Encoding Forms

Decimal	UTF-8	UTF-16	UTF-32
40857	E9 BE 99	9F99	00009F99

IRG Sources

Data type	Value
kIICore	AG
kIRG_GSource	G0-417A
kIRG_HSource	H-89C8
kIRG_TSource	TF-2159
kRSUnicode	212'.0
kTotalStrokes	5

Dictionary Indices

Data type	Value
kHanYu	74804.010
kIRGHanyuDaZidian	74804.010
kIRGKangXi	1537.251
kKangXi	1537.251

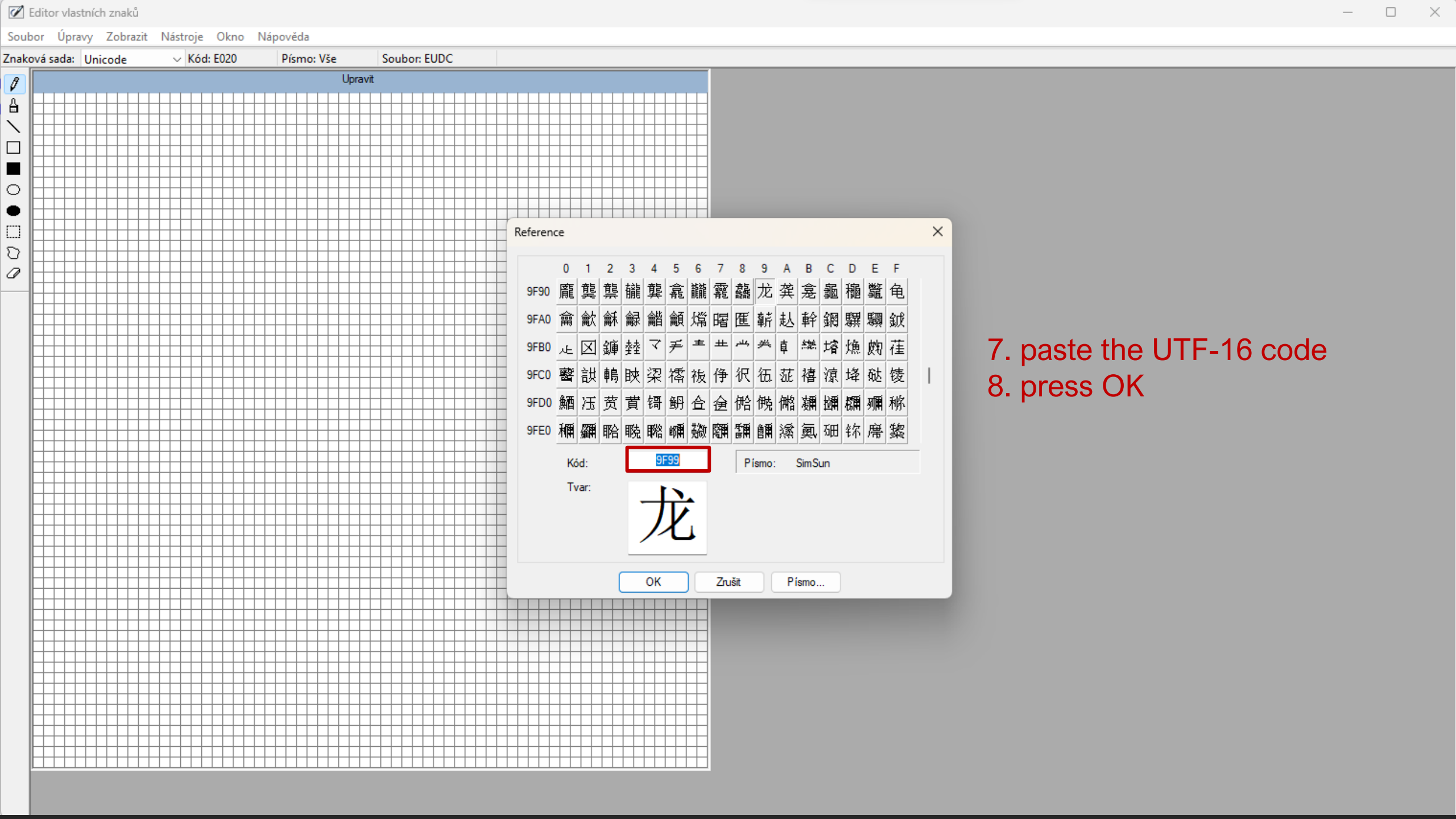
Dictionary-like Data

Data type	Value
kCangjie	IKP
kFourCornerCode	4301

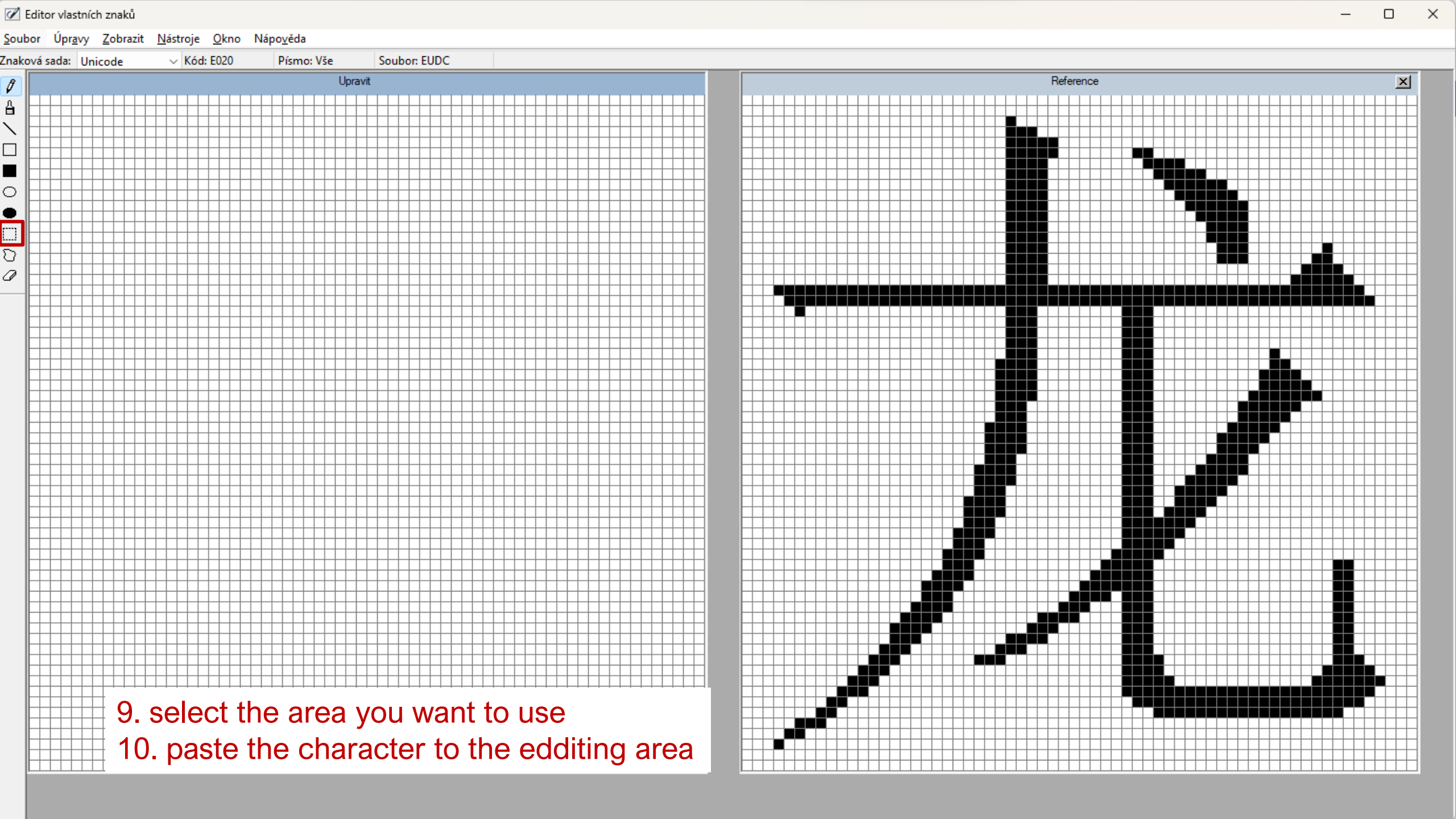
UTF-16

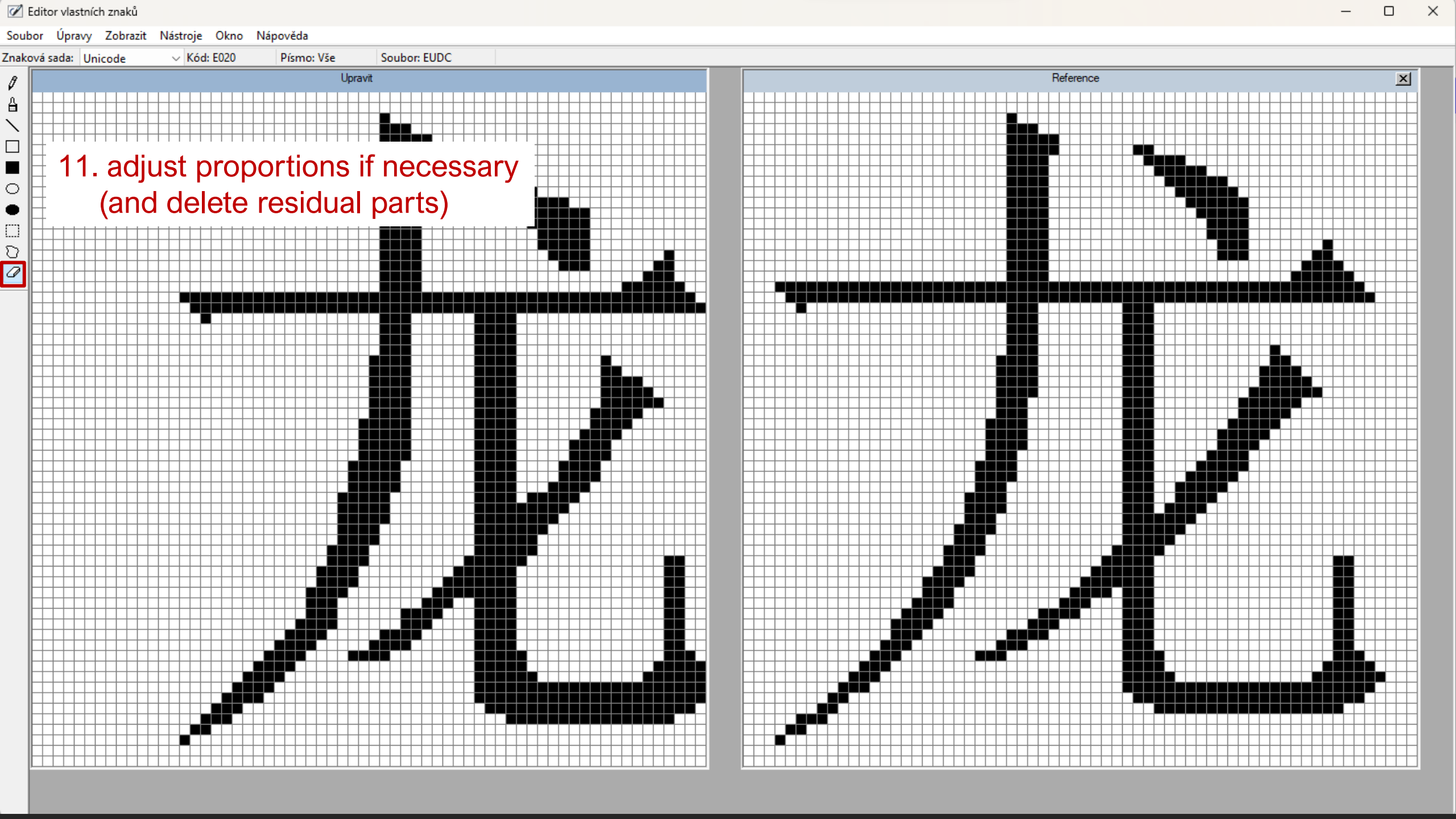
9F99

7. copy the UTF-16 code



7. paste the UTF-16 code
8. press OK


Editor vlastních znaků
Soubor Úpravy Zobrazit Nástroje Okno Nápověda
Znaková sada: Unicode
Kód: E020
Písmo: Vše
Soubor: EUDC
Upravit
Reference
9. select the area you want to use
10. paste the character to the edditing area


Editor vlastních znaků
Soubor Úpravy Zobrazit Nástroje Okno Nápověda
Znaková sada: Unicode
Kód: E020
Písmo: Vše
Soubor: EUDC
Upravit
Reference
11. adjust proportions if necessary
(and delete residual parts)

https://www.unicode.org/cgi-bin/GetUnihanData.pl?codepoint=你

Unihan Database

Unihan data for U+4F60

Lookup

Grid Index
<<< Previous

Radical-Stroke index (9.4-6)
Next >>>

Glyph

你

UTF-16

4F60

12. copy the UTF-16 code

Encoding Forms

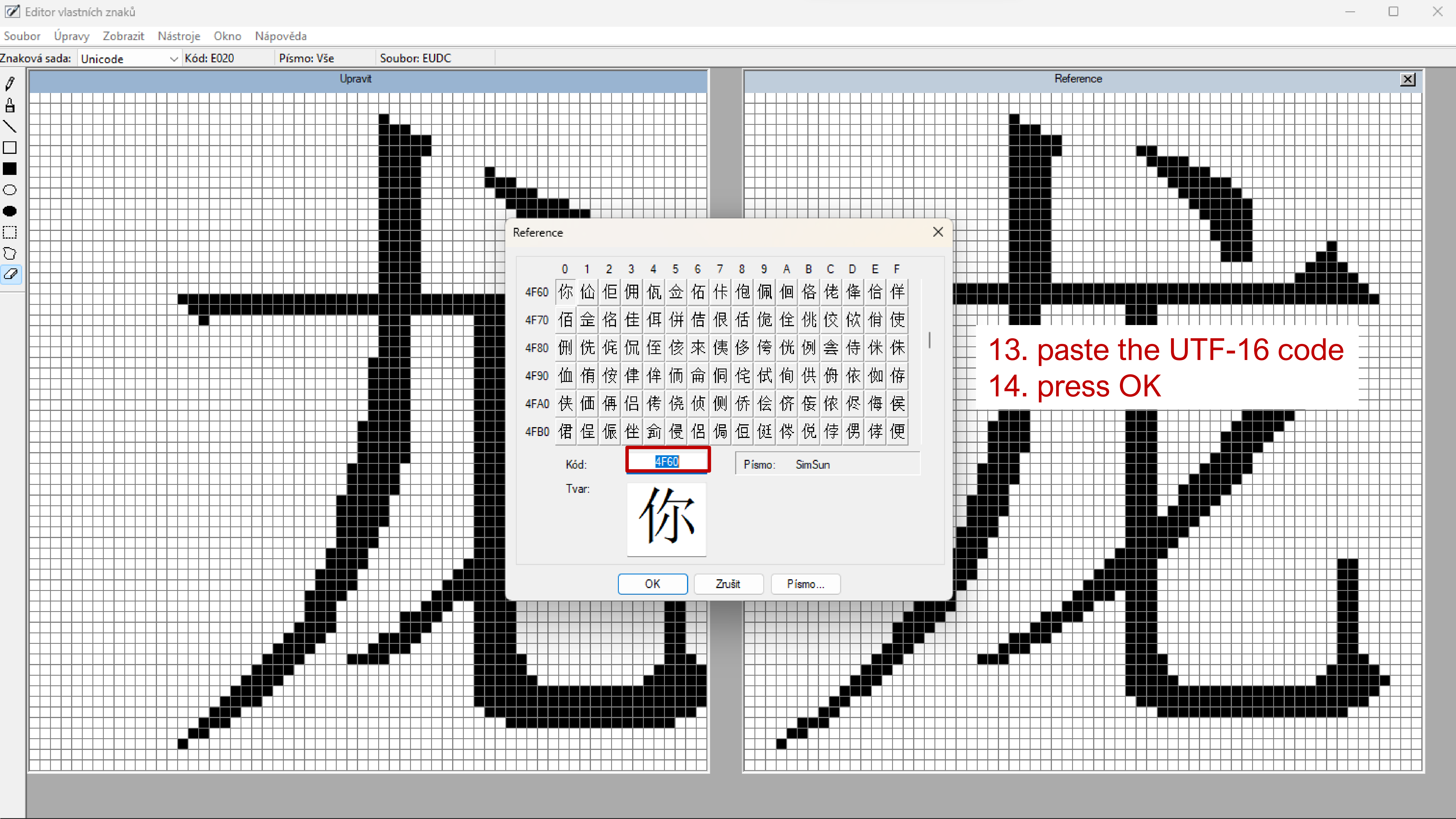
Decimal	UTF-8	UTF-16	UTF-32
20320	E4 BD A0	4F60	00004F60

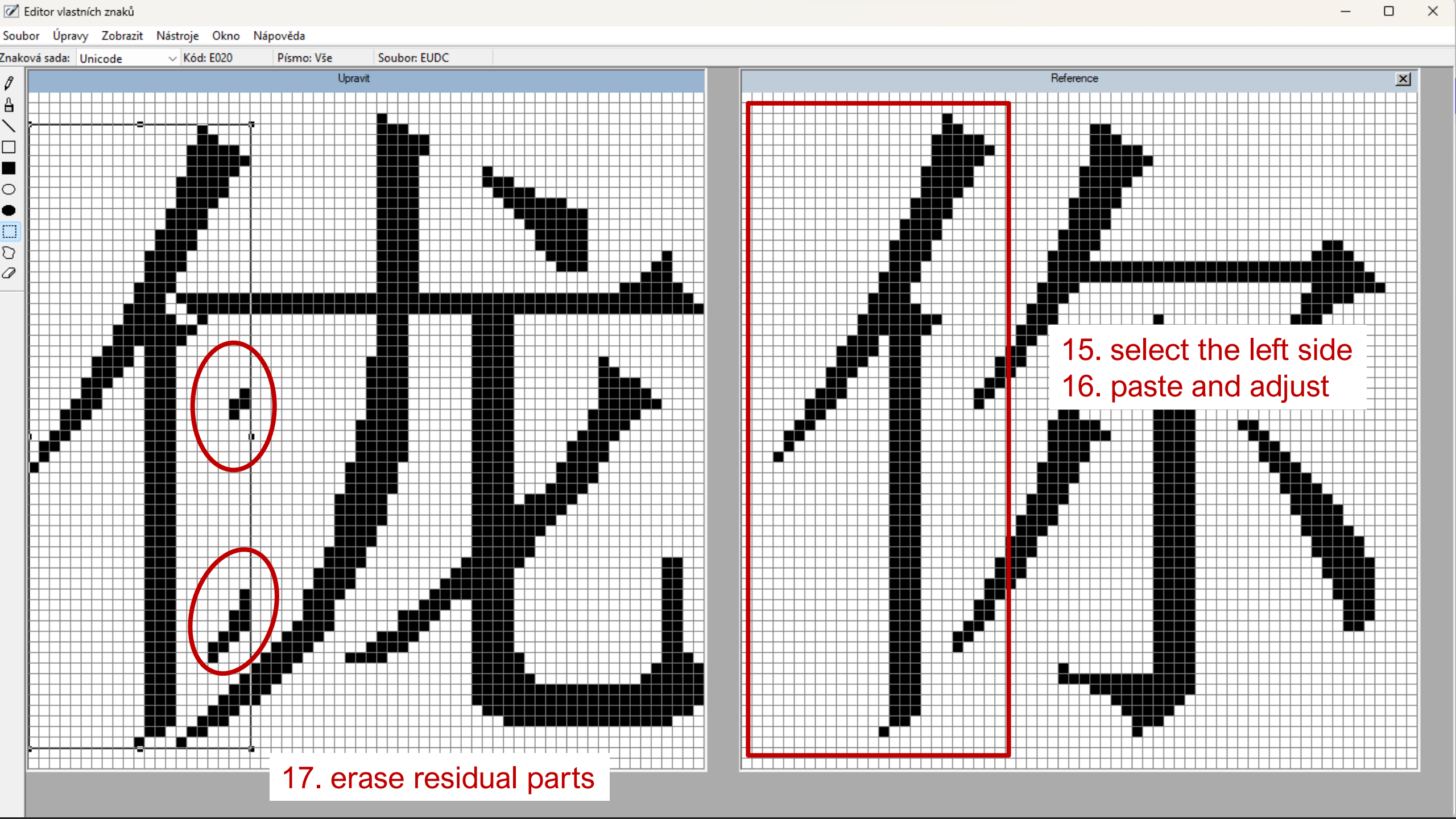
IRG Sources

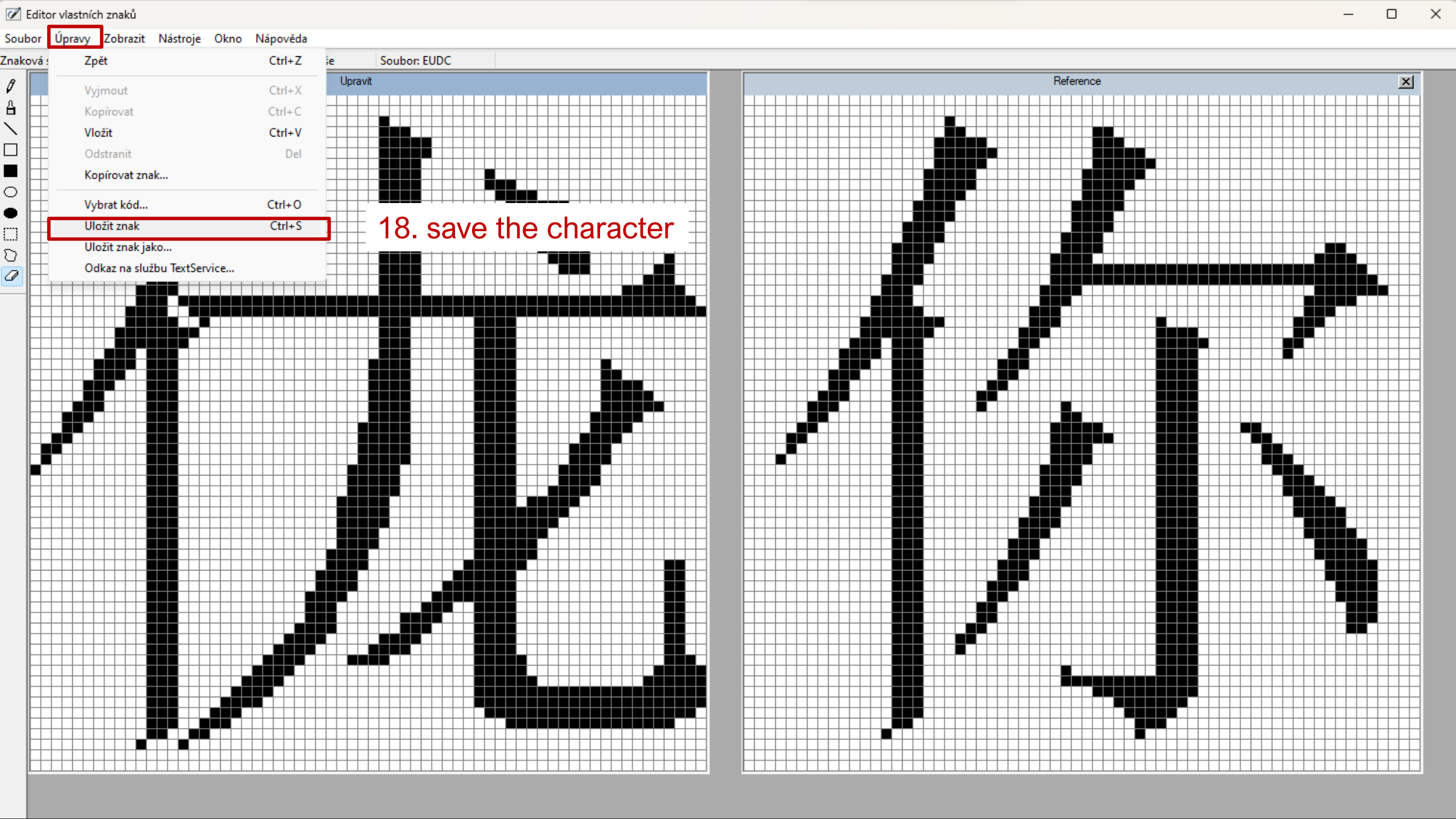
Data type	Value
kIICore	AGTHM
kIRG_GSource	G0-4463
kIRG_HSource	HB1-A741
kIRG_JSource	J13-2E2D
kIRG_KPSource	KP1-34EC
kIRG_KSource	K2-2221
kIRG_TSource	T1-4923
kIRG_VSource	V1-4B35
kRSUnicode	9.5
kTotalStrokes	7

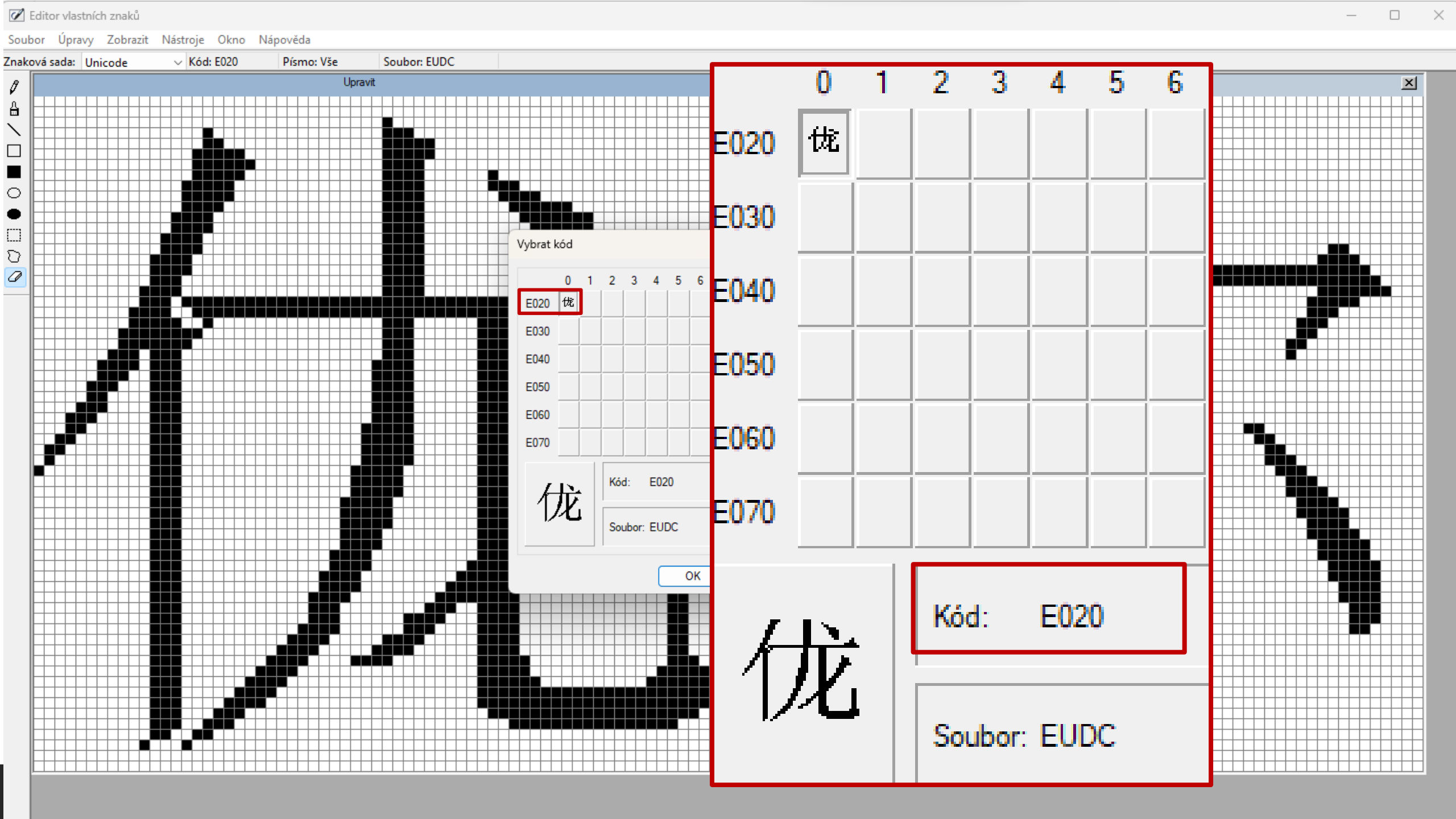
Dictionary Indices

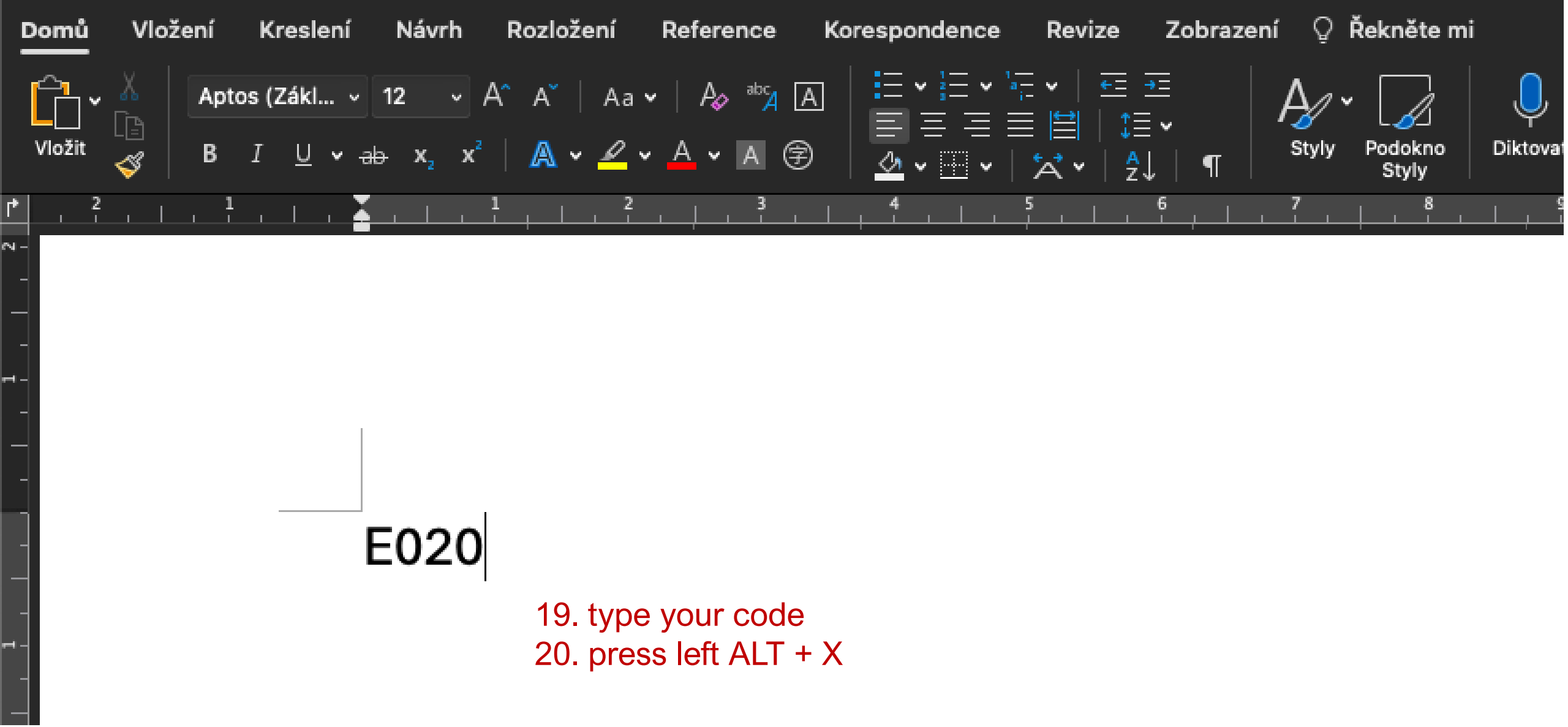
Data type	Value
kCihaiT	100.209
kCowles	2875
kDaeJaweon	0205.040
kFennIndex	360.04
kHanYu	10137.050


Editor vlastních znaků
Soubor Úpravy Zobrazit Nástroje Okno Nápověda
Znaková sada: Unicode
Kód: E020
Písmo: Vše
Soubor: EUDC
Upravit
Reference
Reference
0 1 2 3 4 5 6 7 8 9 A B C D E F
4F60
4F70
4F80
4F90
4FA0
4FB0
Kód: 4F60
Písmo: SimSun
Tvar:
你
OK
Zrušit
Písmo...
13. paste the UTF-16 code
14. press OK


Editor vlastních znaků
Soubor Úpravy Zobrazit Nástroje Okno Nápověda
Znaková sada: Unicode
Kód: E020
Písmo: Vše
Soubor: EUDC
Upravit
Reference
15. select the left side
16. paste and adjust
17. erase residual parts


Editor vlastních znaků
Soubor
Úpravy
Zobrazit
Nástroje
Okno
Nápověda
Zpět Ctrl+Z
Vyjmout Ctrl+X
Kopírovat Ctrl+C
Vložit Ctrl+V
Odstranit Del
Kopírovat znak...
Vybrat kód... Ctrl+O
Uložit znak Ctrl+S
Uložit znak jako...
Odkaz na službu TextService...
Soubor: EUDC
Upravit
Reference
18. save the character


Editor vlastních znaků
Soubor Úpravy Zobrazit Nástroje Okno Nápověda
Znaková sada: Unicode
Kód: E020
Písmo: Vše
Soubor: EUDC
Upravit
Vybrat kód
0 1 2 3 4 5 6
E020
E030
E040
E050
E060
E070
Kód: E020
Soubor: EUDC
OK
0 1 2 3 4 5 6
E020
E030
E040
E050
E060
E070
Kód: E020
Soubor: EUDC


Domů
Vložení
Kreslení
Návrh
Rozložení
Reference
Korespondence
Revize
Zobrazení
Řekněte mi
Vložit
Aptos (Zákl...
12
Styly
Podokno Styly
E020
19. type your code
20. press left ALT + X

Tokenization

- parsing text (sentences, phrases) to a smaller units, i.e., **tokens**
- **token** (words or even parts of words, punctuation, dates) – a distinct chunk of information
- in NLP – preparatory step for language modelling or machine translation
- particulary useful when building a huge dataset

Types of Tokenization

- word
- subword
- sentence
- character

Word Tokenization

- dividing text into **graphical** words (spaces)
- could be problematic in the case of East Asian Languages
  - 我的 „my, mine“ (我 [first person pron.] 的 [grammatical particle]) – one word or two?

Example:

No pictures from the mission at the surface have yet been released.

“No“, “pictures“, “from“, “the“, “mission“, “at“, “the“, “surface“, “have“, “yet“, “been“, “released“, “.“ (13 tokens)

! punctuation also counts as a separate token

Subword Tokenization

- morphological analysis
- lemma (dictionary form), tag

Example:

processing “process“, “ing“ (2 tokens)

Confucianism “Confucian“, “ism“ (2 tokens)

Sentence Tokenization

- individual sentence analysis

Example:

One of the deepest teachings of Confucius may have been the superiority of personal exemplification over explicit rules of behavior. His moral teachings emphasized self-cultivation, emulation of moral exemplars, and the attainment of skilled judgment rather than knowledge of rules.

“One of the deepest teachings of Confucius may have been the superiority of personal exemplification over explicit rules of behavior.“, “His moral teachings emphasized self-cultivation, emulation of moral exemplars, and the attainment of skilled judgment rather than knowledge of rules.“

Character Tokenization

- character-level languages

Example:

language	“l“, “a“, “n“, “g“, “u“, “a“, “g“, “e“	8 tokens
學校	“學“, “校“	2 tokens
猫头鹰	“猫“, “头“, “鹰“	3 tokens

How?

- Tokenizer (Python, online demo pages) – could be limited for Asian languages
- By yourself – make sure you have set all the criteria for your analysis in advance

Why?

- data can be used more effectively – no need for analysis of the whole text
- data summarization
- crucial step in language modelling
- good for searching for collocations, context use and function of particular words (e.g., specifical usage of sentence final particles, mimetic words, sentiment analysis)
- building and structuring corpus vocabulary list

CJK Corpora – Chinese

BCC Online Corpus – sometimes it actualy works

http://bcc.blcu.edu.cn

- Yiyan – database of Chinese/English Corpora, Parallel Corpora, GLOBE family Corpora

http://114.251.154.212/cqp/ (user ID: test; password: test)

- Alphabetical – Index (not only Chinese corpora)

https://corpus.bfsu.edu.cn/Corpora_A-Z_Beijing_Foreign_Studies_University_Corpus_Research_Group.html

CJK Corpora – Japanese

- NLT – user friendly interface, has an English version

https://tsukubawebcorpus.jp/en/

- NLB – same interface as NLT, only in Japanese

https://nlb.ninjal.ac.jp/

- Chunagon

https://shonagon.ninjal.ac.jp/ ; Corpus Usage Tools: https://clrd.ninjal.ac.jp/en/tool.html

Czech National Corpus

- accesible from: https://www.korpus.cz
- WaG (Word at Glance) – basic characteristics, word forms, frequency, collocations + text collocations → https://www.korpus.cz/slovo-v-kostce/
- KonText – concordances → https://kontext.korpus.cz/
- Treq – translation equivalents databasis → https://treq.korpus.cz

- Guidelines: https://wiki.korpus.cz/

Encoding + Unicode

Unicode 15.1.0 Standard Guidebook

https://www.unicode.org/versions/Unicode15.0.0/UnicodeStandard-15.0.pdf

Unicode 15.1.0 Character Code Charts

https://unicode.org/charts/#scripts

Unihan Database

https://www.unicode.org/charts/unihan.html

Tokenization, Tokenizers

Basic intro to tokenization with examples in Python:

https://www.geeksforgeeks.org/nlp-how-tokenizing-text-sentence-words-works/

Chinese Tokenizer (demo page):

https://yishn.github.io/chinese-tokenizer/

Chinese Telegraph Code (online)

https://www.qqxiuzi.cn/bianma/dianbao.html

Thank you for your attention!