# A Cautionary Note about Saving Unicode Text from Excel

Neal Altman

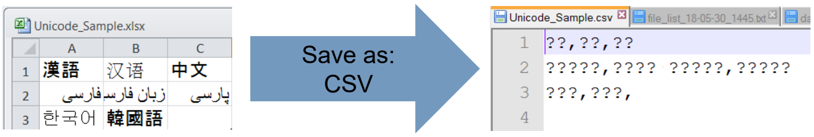
na@cmu.edu

Carnegie Mellon

Center for Computational Analysis of
Social and Organizational Systems
http://www.casos.cs.cmu.edu/

## Microsoft Excel’s CSV Save

- Non-ASCII characters may be lost when exporting in CSV (comma separated value) format:



Unicode_Sample.xlsx

|  | A | B | C |
|---|---|---|---|
| 1 | 漢語 | 汉语 | 中文 |
| 2 | فارسی | زبان فارسی | پارسی |
| 3 | 한국어 | 韓國語 |  |

Unicode_Sample.csv

```
??,??,??
?????,???? ?????,?????
???,???,
```

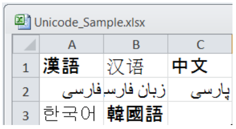
- Why? The content was saved as “Western European (Windows)” (aka Windows-1252) a limited extension to the standard ASCII character set.
- Good news! The CSV UTF-8 option was announced for MS Office build 16.0.7466.2023 and after.

  (But only if you pay for an upgrade.)

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

# "Old" Excel Workaround

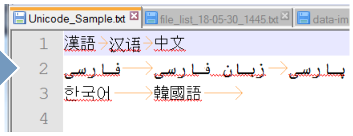
- Export as Unicode Text (saves as TSV - tab separated value format):



Unicode_Sample.xlsx

|   | A | B | C |
|---|---|---|---|
| 1 | 漢語 | 汉语 | 中文 |
| 2 | فارسی | زبان فارسی | پارسی |
| 3 | 한국어 | 韓國語 |  |

Save as:
Unicode Text

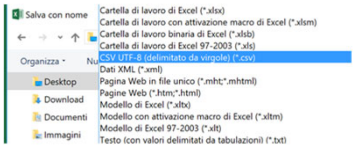


```
漢語	汉语	中文
فارسی	زبان فارسی	پارسی
한국어	韓國語	
```

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

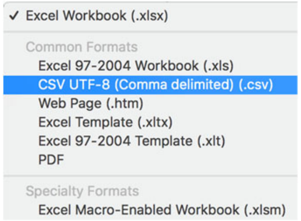
# Upgrade Workarounds

- Use a version of Excel with the "CSV UTF-8" option:



Cartella di lavoro di Excel (*.xlsx)
Cartella di lavoro con attivazione macro di Excel (*.xlsm)
Cartella di lavoro binaria di Excel (*.xlsb)
Cartella di lavoro di Excel 97-2003 (*.xls)
CSV UTF-8 (delimitato da virgole) (*.csv)
Dati XML (*.xml)
Pagina Web in file unico (*.mht;*.mhtml)
Pagine Web (*.htm;*.html)
Modello di Excel (*.xltx)
Modello con attivazione macro di Excel (*.xltm)
Modello di Excel 97-2003 (*.xlt)
Testo (con valori delimitati da tabulazioni) (*.txt)

MS Windows, Italian
(Courtesy Gian Maria Campedelli)



✓ Excel Workbook (.xlsx)
Common Formats
Excel 97-2004 Workbook (.xls)
CSV UTF-8 (Comma delimited) (.csv)
Web Page (.htm)
Excel Template (.xltx)
Excel 97-2004 Template (.xlt)
PDF
Specialty Formats
Excel Macro-Enabled Workbook (.xlsm)

Macintosh Office 365
(Courtesy Mihovil Barulovic)

- Use a different spreadsheet which supports Unicode CSV save:
  - Open Office Calc (user selects encoding)
  - LibreOffice Calc (user selects encoding)
  - Google Sheets (UTF-8 without BOM)