Příloha č. 1: Zdrojový kód skriptu použitého na rozdělení dat z korpusu SYN2020

```
# Požádání uživatele o název souboru, který má být zpracován. Na základě něho vytvoření šablony pro pojmenovávání generovaných souborů.
text_to_open = input("Write filename of the text you want to process: ")
if text_to_open[-4:] != ".csv":
    base_name = text_to_open
    text_to_open = text_to_open + ".csv"
else:
    base_name = text_to_open[:-4]

# Požádání uživatele o informaci o zdroji dat (použitý korpus či dataset).
source = input("What is the source of the data? ")

with open(text_to_open, "r", encoding="utf-8") as loaded_text:
    # Rozdělení textu na řádky. Stanovení počtu řádků kvůli zvolení správného formátu číslování rozdělených souborů.
    lines = loaded_text.readlines()
    line_count = len(lines)
    padding = len(str(line_count))
    file_id = 1
    for line in lines:
        # Náhrada separátoru ; za ;;; v kontextech, v nichž jde patrně o separátor. Pro účely dělení jde o nepravděpodobnou sekvenci, u níž nehrozí riziko nesprávné segmentace.
        line = line.replace('";"', '";;;"')
        # Extrakce metadat a vlastního textu.
        (title, year, source_language, txtype_group, txtype, genre_group, genre, medium, context_a, kwic, context_b) = line.split(";;;")
        context_a = context_a[:-1]
        kwic = kwic[1:-1]
        context_b = context_b[1:]

        # Generace identifikační přípony souboru.
        id_suffix = str(file_id).rjust(padding, '0')
        # Vytvoření finálního souboru se zvláštními řádky vyhrazenými pro jednotlivá metadata. Obsahuje navíc řádky TAGS pro vpisování štítků a NOTES pro zápis poznámek vztahujících se k souboru.
        with open(f"{base_name}_{id_suffix}.md", "w", encoding="utf-8") as f:
            f.writelines([
                    f'SOURCE: "{source}"\n',
                    f'TITLE: {title}\n',
                    f'YEAR: {year}\n',
```

```
                    f'SOURCE LANGUAGE: {source_language}\n',
                    f'TEXTTYPE GROUP: {txtype_group}\n'
                    f'TEXTTYPE: {txtype}\n'
                    f'GENRE GROUP: {genre_group}\n'
                    f'GENRE: {genre}\n'
                    f'MEDIUM: {medium}\n'
                    f'TAGS:\n'
                    f'NOTES:\n'
                    f'CONTEXT:\n{context_a}    <font    style="color:dc320e;font-
weight:bold">{kwic}</font style> {context_b}'
                    ])

        file_id = file_id + 1
```