# Memo: RTF → Excel Bible Verse Exporter

## Goal

Convert a batch of RTF Bible book files into Excel workbooks where each row is a verse: - **Column A:** Full book name + reference (e.g., Ruth 1:1) - **Column B:** Verse text **verbatim** (as printed in the RTF), with the verse number removed from the text body.

## Inputs

* One RTF per Bible book, e.g., 08.Ruth.rtf, 14.2 Chronicles.rtf. The filename contains the full book name we should use in the reference.
* Within each RTF:
  + **Chapter numbers** are the large inline numerals that start a chapter.
  + **Verse numbers** are consistently formatted as **superscript** numerals.
  + All chapter and verse numerals share consistent styling **within a batch**, but **may change** (font name / size) in future batches.

## Output

* One Excel file per input RTF (or one combined workbook, configurable), with:
  + A header row: Reference, Text (configurable).
  + One row per verse in canonical order.

## Parsing Rules (critical)

1. Treat any **superscript run of digits** as a **verse number start** (v). When you encounter a superscript digits token, finish the previous verse (if any), then start collecting text for v.
2. Treat a **non-superscript run of digits** that matches the **chapter font spec** (see Config) and appears at a chapter boundary as a **chapter number** (c).
3. Verse text should be the characters **after** its superscript number until the next superscript number or end of chapter/file.
4. Preserve punctuation, casing, small caps, and special characters as emitted by the RTF to Unicode. Do **not** normalize quotes or whitespace except to collapse hard RTF control boundaries.
5. Verse numbers should **not** appear in the exported text column.

## Configurability (future-proofing)

At the top of the script, expose a CONFIG block to adapt to future file batches: - CHAPTER\_FONT\_NAME (optional; leave None to ignore). - CHAPTER\_MIN\_FS\_HALFPOINTS (RTF \fsN uses **half-points**; e.g., 48 = 24pt). - VERSE\_IS\_SUPERSCRIPT (bool; typically True). - VERSE\_FONT\_NAME (optional; leave None to ignore). - Filename → book-name extraction pattern. - Per-book output mode (one workbook per book vs. one combined file).

If the next batch changes fonts/sizes, we edit values in this config—no code changes.

## Edge Cases & Notes

* Some RTF producers wrap numbers in groups: watch for {\\*\pntext ...} or nested groups; the parser handles nested groups and toggles \super state correctly.
* RTF \fsN changes can occur mid-paragraph; keep a current “font state” while iterating tokens.
* Handle chapter spans with headings or section titles interleaved—ignore non-numeric superscripts for verse starts.
* Unicode escapes \'hh and \uN? must be decoded.
* If a chapter number appears immediately before a verse superscript (e.g., “1^1”), treat the large number as chapter, the superscript as verse 1.
* If a verse appears before any chapter (rare), assume chapter 1 until a chapter token appears.

## Directory Layout & Run

* Drop RTF files into an input folder (e.g., ./rtf\_in/).
* Run python rtf2excel.py.
* Outputs Excel files into ./xlsx\_out/ by default.