

An Efficient R Shiny App for Spelling Validation in Pharmaceutical Clinical Study Documentation Review

Lufei Wang, Jianyong Tong, Dizal Pharma

ABSTRACT

Proper documentation practices in clinical research serve as fundamental requirements for achieving regulatory compliance in submission processes. Conventional spelling checkers (e.g., Microsoft Word), reliant on generic lexicons, frequently misclassify commonly used terminologies (such as CDISC CT, variable names, MedDRA terms, etc.) in routine programming documentation as spelling errors, requiring extra effort for manual verification. This paper presents a spelling validation tool developed with the R Shiny framework, which combines the hunspell spelling engine and customized dictionaries to create a comprehensive terminology-checking tool to enhance efficiency in achieving regulatory compliance for programming documentation. Key features of the application include: (1) Semi-dynamic lexicons integrating CDISC core vocabulary (covering SDTM CT, internal SDTM and ADaM metadata), specialized dictionaries (MedDRA, WHODrug), and temporary user-defined dictionaries for project-specific terms; (2) Support for different format validation of routine programming documentation (.xlsx, .docx); and (3) Generation of comprehensive validation results.

INTRODUCTION

Within clinical trial submission packages, comprehensive documentation constitutes a critical component. In Dizal's established workflow, the define.xml file is programmatically generated by extracting key metadata and structural information directly from Excel-based ADaM specification documents (.xlsx). Concurrently, clinical documents such as CSDRG.docx and ADRG.docx must be manually written by programmers. And precise and accurate wording in such files is fundamentally critical to ensure terminology consistency in submission deliverables.

A persistent challenge in clinical documentation is that standard spell-checkers flag specialized terms (e.g., CDISC controlled terminologies, MedDRA terms) as incorrect spellings. Given the volume and complexity of submission documents, manual verification demands excessive effort while introducing quality risks. Existing tools have limitations: commercial software (e.g., Microsoft Word) lacks field-customizable dictionaries, while open-source engines (e.g., hunspell) do not support clinical terminology.

To address these requirements, we developed an interactive Shiny application that performs spelling validation for both Excel (.xlsx) and Word (.docx) files. And this tool supports clinical terminology libraries (CDISC/MedDRA), and it enables dynamic dictionary customization. Thereby this spelling validation tool can accelerate documentation reviews while ensuring CDISC terminology accuracy.

DEVELOPMENT AND IMPLEMENTATION OF THE TOOL

HUNSPELL ENGINE

Hunspell is a free, open-source (LGPL/GPL/MPL) spell checking and morphological analysis library and command-line tool. Powering major applications like LibreOffice, Firefox, and Chrome, as well as Linux and macOS, it offers high-quality correction for languages with complex word structures, rich morphology, and compounding. Originally derived from MySpell (OpenOffice), its development emphasizes Unicode support, highly customizable suggestions, and advanced capabilities like handling intricate compounds, extensive affixation (64k classes), and morphological analysis/generation.

The hunspell R package provides a low-level interface to this library within the R environment. It enables direct access to Hunspell's core functions within R, allowing users to check spelling, analyze morphology, and tokenize text in individual words or structured documents.

This spelling validation tool is built upon the foundational function provided by the hunspell R package.

BUILDING CUSTOMIZED LEXICONS

The customized lexicons form the foundation of the application's validation capability. This integrated terminology resource comprises three complementary components:

1. **Core CDISC Dictionary:** Constructed from SDTM and ADaM metadata (covering domain/dataset names, variable names, controlled terminology, codelists, and specialized formats), updated upon new CDISC releases.
2. **Specialized Clinical Dictionaries:** Incorporating standardized medical terminologies (MedDRA and WHODrug), updated when new versions are published.
3. **User-Defined Dictionary:** A project-specific repository for ad hoc terms, supporting addition and deletion by users.

Collectively, these lexicons contain 29,864 unique terms essential for accurate clinical documentation validation. Table 1 summarizes their scope and maintenance.

Lexicon Type	Scope	Update Mechanism
Core CDISC Dictionary	SDTM/ADaM Metadata (Variables, CT, etc.)	Updated on CDISC Release
Specialized Dictionaries	MedDRA, WHODrug Terms	Updated on New Release
User-Defined Dictionary	Project-Specific Terms (e.g., Abbreviations)	Real-Time User Maintenance

Table 1. Composition and Maintenance of Customized Lexicons

Lexicon Type 1: Core Dictionary

- **SDTM Component:** Extracts terms from SDTM metadata, including domain names, variable names, controlled terminology (CDISC CT), company-level codelists. Representative examples are provided in Table 2.
- **ADaM Component:** Extracts terms from ADaM metadata, encompassing dataset names, variable names, and controlled terminology. Representative examples are shown in Table 3.

Category	Count	Examples
Domain	127	"CO" "DC" "DM" "PB" "SB" "SE" "SM" "SV" "AG" "CM" "EC" "EX" "SUPPMH" "SUPPDD" "SUPPFA" "SUPPIE"
Variable	5883	"STUDYID" "DOMAIN" "USUBJID" "AELLTCD" "AEDECOD" "AEPTCD" "AEHLT" "AEHLTCD" "LCLLOQ" "LCULOQ"
Controlled Terminology	19278	"Sigmoideus" "Infrascapular" "Subclavian" "Subcutis" "Subretinal" "Subscapularis" "Supraglottis" "Supramarginal"
Company-Level Codelist	1280	"DCond" "DExp" "LBALL" "SECONDS" "sec" "LAUC072W" "LAUC072D" "LAUC120S" "LAUC120W" "LAUC120D"

Table 2. Representative Terms Extracted from SDTM Metadata

Category	Count	Examples
Dataset	24	"ADPC" "ADVS" "ADECOC" "ADRE" "ADRS" "ADCM" "ADCV" "ADTTE" "ADSL" "ADEC" "ADPP" "ADDILI" "ADDATES" "ADDV"
Variable	831	"ARM" "ARMCD" "ACTARM" "ACTARMCD" "AGE" "AGEU" "AGEGR1" "AGEGR1N" "BMIBL" "CANCRTYP" "COUNTRY" "STRATA1" "STRATA1N"
Controlled Terminology	78	"DATEFL" "DTYPE" "BLOCF" "WOCF" "BOC" "BOCF" "EXTRAP" "HALFLLOQ" "INTERP" "LLOD"

Table 3. Representative Terms Extracted from ADaM Metadata

Lexicon Type 2: Dictionary of Medical Terms

This lexicon integrates standardized medical terminologies critical for regulatory reporting:

- **MedDRA (Medical Dictionary for Regulatory Activities):** The ICH-standardized terminology for classifying adverse event information.
- **WHODrug:** The WHO's global medicinal product dictionary.

Terms relevant to programming documentation are extracted from these sources. Representative examples are presented in Table 4.

Dictionary	Count	Examples
MedDRA v26.0	1007	"Colopathy" "Hemipelvectomy" "Ankyloglossia" "Anodontia" "Enterolysis" "Paraneoplastic" "Klebsiella" "Arthralgia"
WHODrug 2023Q3	929	"ANTIDIARRHEAL" "ANTIANEMIC" "ANTIPSYCHOTICS" "ANTIFIBRINOLYTICS" "ANTITHYROID" "HYPERKALEMIA" "HYPERPHOSPHATEMIA" "OPHTHALMOLOGICALS"

Table 4. Representative Terms Extracted from MedDRA and WHODrug

Lexicon Type 3: User-Defined Dictionary

This flexible component allows users to augment the lexicons with project-specific terminology not covered by the core or specialized dictionaries. Common additions include:

- **Documentation Conventions:** Terms like "eDT", "csdrg", "adrg", "chinadrugtrials", "StudyName", "ProtocolName", "StandardVersion", "datetime", "partialDate".
- **Variable Conventions:** Patterns such as "XXSEQ", "XXSTDY", "XXENDY", "XXDY", "XXSTRESN", "XXLOBXFL", "XXBLFL", "XXSTRESC", "XXENDTC", "XXSTDTC", "PTxxH".
- **SAS Functions:** Terms like "upcase", "substr", "ne".
- **Abbreviations:** Country codes (e.g., "CHN", "USA", "JPN", "KOR") and company names (e.g., "Dizal").

TOOL INTERFACE AND FUNCTIONAL CAPABILITIES

The core function of the application is to utilize the integrated lexicons as the validation standard to perform spelling checks on uploaded documents. Any term not present within these lexicons is flagged as a potential spelling error for user review.

The Shiny application interface is organized into four sequential modules, guiding the user through the validation workflow:

1. File Upload

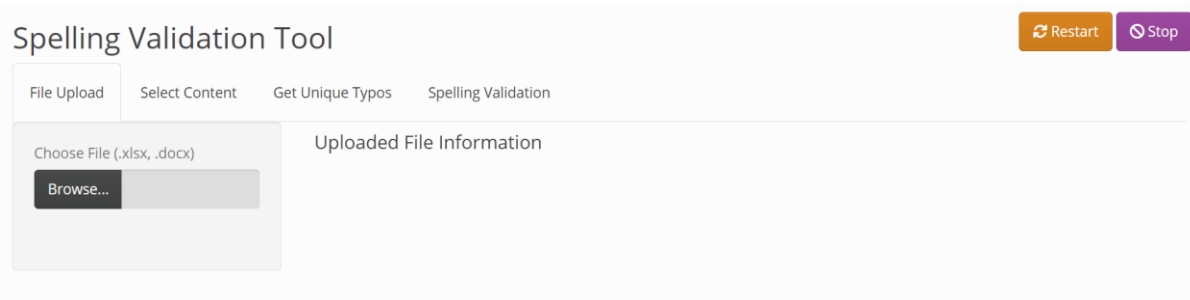


Figure 1. Tool Interface (File Upload)

- Supports input of both Excel (.xlsx) and Word (.docx) document formats.
- Upon successful upload, the right-hand panel displays essential file information including document name, size, and format.

2. Content Selection for Validation

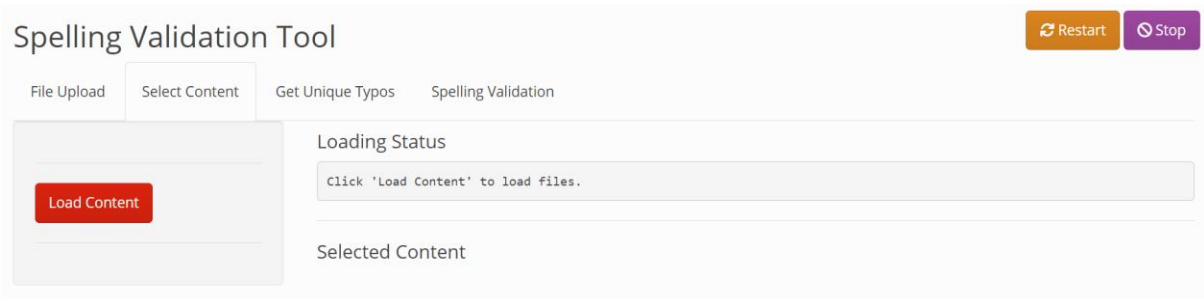


Figure 2. Tool Interface (Select Content)

- **Excel Documents:** To accommodate the complexity of Excel files (multiple sheets) and the common need to validate specific columns within specific sheets, the interface provides dynamic sheet selection. Users first activate relevant sheets, with the right-hand panel providing real-time load status confirmation. Following sheet selection, all columns within chosen sheets are automatically selected for validation by default. Users can then refine validation targets by deselecting non-relevant columns in the column selection interface located below the Selected Content panel.
- **Word Documents:** The interface provides two processing modes for Word files validation: Full Document mode and Paragraph Segmentation mode. Selection of Full Document processes the entire file as a continuous text block, while Paragraph Segmentation automatically divides the document into discrete units based on first- and second-level headings before validation. Users may select their preferred mode and dynamically switch between options within this panel during real-time spelling validation process.

3. Found Unique Potential typos

The screenshot shows the 'Spelling Validation Tool' interface with the 'Get Unique Typos' tab selected. The interface includes a top navigation bar with 'File Upload', 'Select Content', 'Get Unique Typos', and 'Spelling Validation'. On the right, there are 'Restart' and 'Stop' buttons. The main content area is divided into two sections. The left section contains a red button labeled 'Get Unique Potential Typos'. The right section, titled 'Unique Potential Typos', contains instructions: 'If you want to remove words from the potential typing error results. Write down them separated by comma or pick the words in the following checkbox.' Below this is a text input field with the placeholder 'For example: MDZ, DIGOXINC'. An 'Add Words' button is positioned below the input field. At the bottom, there is a 'Temporary Dictionary' section with the text 'Current temporary dictionary words:'.

Figure 3. Tool Interface (Get Unique Typos)

- This module displays a consolidated list of unique terms flagged as potential spelling errors based on the initial scan against the core lexicons. It enables users to efficiently review the identified potential errors before final validation.
- Crucially, it provides functionality to build a **Temporary Lexicon** for the current session. Terms added to this lexicon are treated as valid for the duration of the current spell-checking run, allowing users to designate specific exclusions:
 - Source 1: Directly add terms from the displayed list of flagged potential errors.
 - Source 2: Manually input terms (separated by commas) into a dedicated text box.
- This Temporary Lexicon operates independently of the pre-defined Customized Lexicons (Core CDISC, Specialized Clinical, User-Defined). Its sole purpose is to allow users to ignore specific terms during the current validation session that they deem correct or irrelevant for this specific check.
- If no terms require exclusion via the Temporary Lexicon, users can skip this module and proceed directly to the final validation report.

4. Final Spelling Validation

The screenshot shows the 'Spelling Validation Tool' interface with the 'Spelling Validation' tab selected. The interface includes a top navigation bar with 'File Upload', 'Select Content', 'Get Unique Typos', and 'Spelling Validation'. On the right, there are 'Restart' and 'Stop' buttons. The main content area is divided into two sections. The left section contains a red button labeled 'Run Spelling Check' and a 'Temporary Dictionary' section with the text 'No words added to temporary dictionary yet.' and a green button labeled 'Download Results'. The right section, titled 'Processing Status', contains a text input field with the placeholder 'Click 'Run Spelling Check' to start processing.' Below this is a 'Results Preview' section.

Figure 4. Tool Interface (Spelling Validation)

This module performs comprehensive spelling validation using the established lexicons and generates detailed results. Key features include:

a) Dynamic Temporary Lexicon Display (If Applicable):

- If a Temporary Lexicon was created in the previous module, it is displayed on the left panel.
- Users retain the ability to modify the Temporary Lexicon entries during this stage.

b) Validation Execution:

- Spelling validation is initiated by clicking the 'Run Spelling Check' button.
- Validation based on the Semi-dynamic lexicons:
 - Only the Customized Lexicons (if no Temporary Lexicon exists).
 - Active Temporary Lexicon (if created by the user).

c) Real-Time Processing Feedback & Results Preview:

Upon clicking 'Run Spelling Check', the right panel updates Processing Status indicating validation progress and Results Preview dynamically displaying flagged potential errors.

For Excel Files:

- Processing Status: Shows the total count of potential errors found per sheet.
- Results Preview: Creates a dedicated tab for *each* sheet containing errors. Each tab displays a table with the following columns: Potential.Typo (The flagged term), Original Text (The full cell content where the term was found), Row (Row number in the source sheet), Column Name (Column header in the source sheet)

For Word Files:

- Processing Status: Shows the total count of potential errors found in the entire document.
- Results Preview: Full Document mode displays in a two-column table: Potential.Typo and Original Text. By Paragraph mode adds two columns: Heading_1 and Heading_2, indicating corresponding first- and second-level headings.

d) Report Download:

The application generates a comprehensive validation report in Excel (.xlsx) format containing:

- Dedicated sheet: listing all terms within the Temporary Lexicon (if used), providing traceability for user exclusions.
- Validation Results: same as presented in Results Preview.

REAL-WORLD APPLICATION AND RESULTS

To demonstrate the practical utility and performance of the spelling validation tool, this section presents illustrative examples using two common types of clinical programming documentation:

- Excel Input: An ADaM specification document.
- Word Input: A Clinical Study Data Reviewer's Guide (ADRG.docx) document.

These standardized submission documents necessitate absolute terminology accuracy for regulatory. The following sections detail the validation process and outcomes for each example.

ADBASE SPECIFICATION VALIDATION WORKFLOW

1. File Upload:

The ADBASE specification file (.xlsx) was loaded into the application. Upon successful ingestion, the right panel automatically displayed file information including name, size, and type.

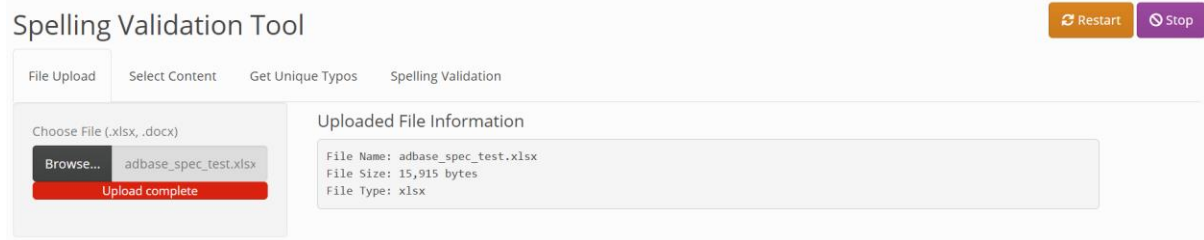


Figure 5. File Upload (ADBASE Specification File)

2. Targeted Content Selection:

Users selectively activated relevant sheets for validation. The interface provided real-time confirmation of loaded content in the status panel, enabling progress verification before advancing.

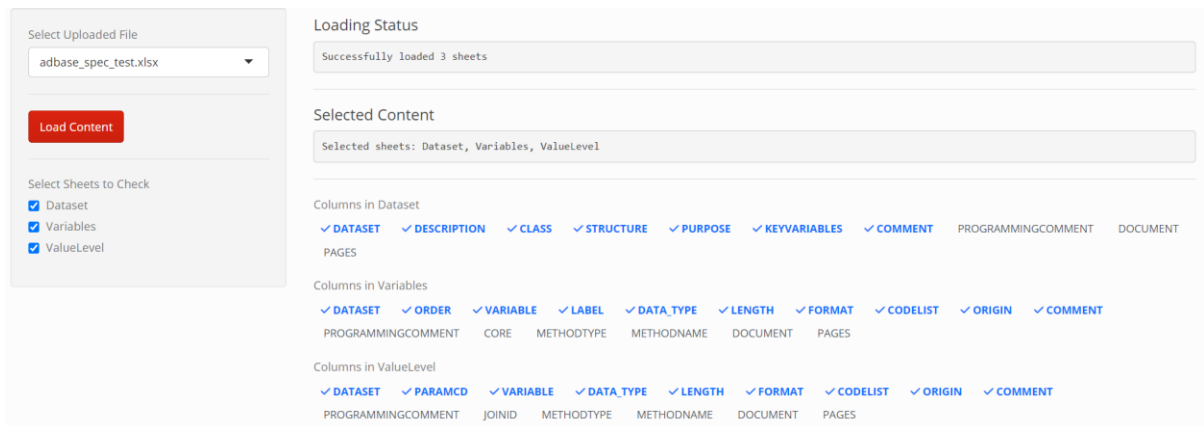


Figure 6. Select Content (ADBASE Specification File)

Within selected sheets, column selection was refined using the following approach:

- Default state: All columns selected for validation
- Custom exclusion: Columns considered no necessary for spelling validation (e.g., "PROGRAMMINGCOMMENT", "DOCUMENT", "PAGES") were removed.
- The interface generated a preliminary list of unique potential spelling errors from the filtered content

3. Temporary Lexicon Intervention:

During this case's validation analysis, several study-specific PARAMCD values ("STHYLNUM", "SURHIS", "RADHIS", "ORGCNT", "PRIORRTS", "PRIORCTS", "PRICTSAJ", "PRICTSNA", "STYPEOT", "METORGAN") were not included in core lexicons. To prevent false-positive flagging of these terms, they were added to Temporary Lexicon, thus the terms were recognized as correct for current validation

purposes, ensuring accurate processing of study-specific parameters.

Unique Potential Typos

Found 17 unique potential typos

If you want to remove words from the potential typing error results. Write down them separated by comma or pick the words in the following checkbox.

Write down specific words to be ignored:

For example: MDZ, DIGOXINC

Pick specific words to be ignored:

Charact Amivantamab Adenosquamous ☒ STHYLNUM ☒ SURHIS ☒ RADHIS ☒ ORGCNT ☒ PRIORRTS ☒ PRIORCTS ☒ PRICSAJ
☒ PRICTSNA ☒ STYPEOT ☒ METORGAN Bronchiolo TULUC ne whenPR

Add Words

Temporary Dictionary

Current temporary dictionary words:

STHYLNUM, SURHIS, RADHIS, ORGCNT, PRIORRTS, PRIORCTS, PRICSAJ, PRICTSNA, STYPEOT, METORGAN

Figure 7. Get Unique Typos and Set Up Temporary Lexicon (ADBASE Specification File)

4. Comprehensive Validation Execution:

The comprehensive validation run subsequently integrated both the core customized lexicons and the temporary lexicon. Upon execution, the Processing Status dynamically quantified and summarized validation outcomes across all sheets, providing immediate quantitative feedback on potential error distribution.

The screenshot displays the 'Final Spelling Validation' interface. On the left, the 'Temporary Dictionary' section shows a list of words to include in the dictionary, with checkboxes for STHYLNUM, SURHIS, RADHIS, ORGCNT, PRIORRTS, PRIORCTS, PRICSAJ, PRICTSNA, STYPEOT, and METORGAN. A 'Download Results' button is at the bottom. The main area is titled 'Processing Status' and shows the following summary:

- Sheet: Dataset - 1 errors found
- Sheet: Variables - 2 errors found
- Sheet: ValueLevel - 5 errors found

Below this is the 'Results Preview' section, which has tabs for 'Dataset', 'Variables', and 'ValueLevel'. The 'ValueLevel' tab is selected. It shows a table with columns for 'Potential.Type', 'Original.Text', 'nRow', and 'Column'. The table lists four entries:

Potential.Type	Original.Text	nRow	Column
Bronchiolo	If FA.FASTRESC in ('Adenocarcinoma: Bronchiolo-Alveolar', 'Adenocarcinoma (NOS)') then ADBASE.AVALC='Adenocarcinoma'; else if FA.FASTRESC='Large Cell Carcinoma (NOS)' then ADBASE.AVALC='Large Cell Carcinoma'; else if FA.FASTRESC = 'Cannot Be Determined' then ADBASE.AVALC = 'Unknown'; else set to FA.FASTRESC when FA.FATESTCD='HSTLTYP'	33	COMMENT
TULUC; ne	if TU.TULUC='OTHER' and SUPPTU.QVAL not missing when SUPPTU.QNAM='TULOCDET' and SUPPTU.IDVARVAL=TU.TUSEQ then set TU.TULOC=SUPPTU.QVAL, then set ADBASE.AVALC=(distinct number of TU.TULOC) under condition of TU.TUEVAL = 'INVESTIGATOR' and TU.TUCAT ne 'NEW LESIONS', convert to character format	35	COMMENT
TULUC	if TU.TULUC='OTHER' and SUPPTU.QVAL not missing when SUPPTU.QNAM='TULOCDET' and SUPPTU.IDVARVAL=TU.TUSEQ then set TU.TULOC=SUPPTU.QVAL, then set ADBASE.AVALC=(distinct number of TU.TULOC)	36	COMMENT
whenPR	ADBASE.AVALC=SUPPPR.QVAL when SUPPPR.QNAM='PRSTT' and SUPPCM.IDVARVAL=PR.PRSEQ when PR.PRCAT = 'PRIOR RADIOTHERAPY', only keep distinct values within subject	37	COMMENT

Figure 8. Final Spelling Validation (ADBASE Specification File)

Within the Results Preview interface, flagged potential typos were displayed alongside their original contextual content and precise source coordinates (row number and column header). This contextual presentation enabled efficient verification of each flagged term against its source documentation.

Following review, the complete validation results report including the Temporary Lexicon entries for traceability (in .xlsx format) was exported via the download functionality.

ADRG VALIDATION WORKFLOW

Structural differences between document types yielded a little difference in ADRG validation pathways compared to specifications in Excel type:

1. File Upload:

Upon successful upload of the ADRG document (.docx), file information (name, size, format) was automatically displayed in the right-hand panel, confirming readiness for validation.

The screenshot shows the 'Spelling Validation Tool' interface. At the top, there are four tabs: 'File Upload', 'Select Content', 'Get Unique Typos', and 'Spelling Validation'. The 'File Upload' tab is active. On the left, there is a 'Choose File (.xlsx, .docx)' section with a 'Browse...' button and a file named 'adrg_test.docx' selected. Below this is a red 'Upload complete' button. On the right, the 'Uploaded File Information' panel displays: 'File Name: adrg_test.docx', 'File Size: 234,203 bytes', and 'File Type: docx'. At the top right of the tool, there are 'Restart' and 'Stop' buttons.

Figure 9. File Upload (ADRG)

2. Full Document Mode and Paragraph Segmentation Mode:

Consistent with the tool's design for Word documents, the entire textual content underwent automated validation with selection of full document selection while Paragraph Segmentation Mode automatically based on first and second level headings divides the document into discrete sections.

The screenshot shows the 'Spelling Validation Tool' interface with the 'Select Content' tab active. On the left, there are two identical panels. Each panel has a 'Select Uploaded File' dropdown menu with 'adbase_spec_test.xlsx' selected, a red 'Load Content' button, and a 'Process Document by:' section. The first panel has 'Full Document' selected, while the second panel has 'By Paragraph' selected. Below this, it says 'Full: Process as single text' and 'Paragraph: Process each paragraph separately'. On the right, the 'Loading Status' and 'Selected Content' panels are shown. The top 'Loading Status' panel says 'DOCX content loaded successfully! (Full Document)'. The 'Selected Content' panel says 'Processing as: full'. The bottom 'Loading Status' panel also says 'DOCX content loaded successfully! (Full Document)'. The 'Selected Content' panel says 'Processing as: paragraph'.

Figure 10. Select Content (ADRG)

3. Comprehensive Validation Execution:

Processing streamlined with no Temporary Lexicon creation being implemented in this case.

Unique Potential Typos

Found 42 unique potential typos

If you want to remove words from the potential typing error results. Write down them separated by comma or pick the words in the following checkbox.

Write down specific words to be ignored:

For example: MDZ, DIGOXINC

Pick specific words to be ignored:

Charact MDZ RDI ULN BCRP OATP digoxin rosuvastatin SCRNFALL NOTASSGN COEFL PDFL ARELPRO ARELALL
AEACNPRO FDOSRDFL FDOSITFL FDOSRED FDOSINT DOSDIS ATRTDSC OUTWINFL adsl adae adbase adex adexsum
adlb addili admh adpc adpp adre adcv adtr addeg advs adcm addv adecog adam toxgr

Add Words

Temporary Dictionary

Current temporary dictionary words:

Figure 11. Empty Temporary Lexicon (ADRG)

The final validation report flagged lowercase dataset references (e.g., adpp.sas7bdat, adpc.sas7bdat) as potential spelling errors. These identifiers represent legitimate file naming conventions within programming documentation. Users retain two resolution pathways: 1) add terms to Temporary Lexicon for systematic exclusion; 2) more conservative way to manually verify and dismiss during results review.

Processing Status

125 errors found

Results Preview

Show 15 entries

Search:

Potential.Typeo	Original.Text
adpc	adpc.sas7bdat
adpp	adpp.sas7bdat
adre	adre.sas7bdat
adcv	adcv.sas7bdat
adtr	adtr.sas7bdat
adeg	adeg.sas7bdat
advs	advs.sas7bdat
adcm	adcm.sas7bdat

Figure 11. Final Spelling Validation in Full Document Mode (ADRG)

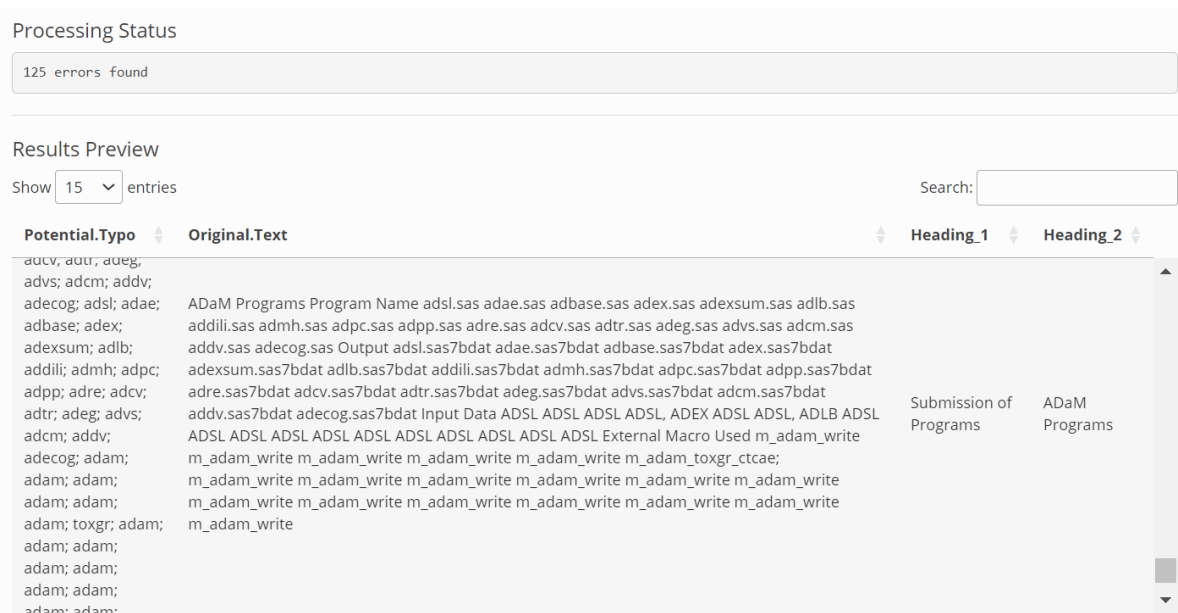


Figure 12. Final Spelling Validation in Paragraph Segmentation Mode (ADRG)

The downloaded Excel report contained two worksheets. One sheet is a placeholder for Temporary Lexicon entries (empty in this case), while the other is the comprehensive listing of all flagged terms with corresponding context strings. Based on validation findings, users determine whether source documents require revision to ensure terminology compliance.

CONCLUSION

This R Shiny-based validation tool addresses persistent challenges in clinical programming workflows by streamlining terminology verification. Its integration of clinical dictionaries substantially reduces verification effort for CDISC standards and medical terminologies, while the temporary lexicon feature provides convenient adjustments during document review.

Though currently optimized for Excel and Word formats covering most programming documentation needs, the framework could potentially extend to additional file types like PDF or XML considering user requirement.

Ultimately, this solution aims not to replace human expertise, but to enhance accuracy where traditional spell-checkers falter, serving as an adaptive assistant for quality documentation in our evolving regulatory landscape."

REFERENCES

hunspell engine available at <https://github.com/hunspell/hunspell>
hunspell R package version 3.0.6 available at <https://github.com/ropensci/hunspell>

Simiao Ye. A Shiny App to Spell Check an ADaM Specification. PharmaSUG 2022 - Paper AD-069.

ACKNOWLEDGMENTS

We would like to thank Dizal programming team for the discussion and testing on the features and functions of the tool.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Lufei Wang
Dizal Pharma
Lufei.wang@dizalpharma.com