UNIVERSITY OF EDINBURGH
Edinburgh Regional Computing Centre

User Note 90

(November 1985)

Title:

SPELL on EMAS 2900

Author: Neil Hamilton-Smith

Contact: Advisory service

Software Support Category: See Note 15

Synopsis

This Note describes two programs on EMAS 2900. The first is for checking and optionally correcting the spelling of a text, and the second for checking if a text had had a word typed twice.

Keywords

CHECKLEX, DOUBLE, lexicon, LEXMERGE, SPELL, SPELLC, SPELLE, STRIPLEX

Edinburgh Regional Computing Centre

James Clerk Maxwell Building, The King's Buildings, Mayfield Road, Edinburgh, EH9 3JZ. Telephone 031-667 1081

© 1985 Edinburgh Regional Computing Centre

1 Introduction

This Note describes a program which will check the spelling of English words in a file, and (if required) generate a corrected version of the file. There are also several utility commands which are used for maintaining lists of words (these lists are called 'lexicons').

The program uses a public list of valid words (the *system lexicon*), but will also create and maintain a list of words peculiar to your own requirements (the *private lexicon*). It is possible to construct lexicons for checking words of languages other than modern English.

2 Access

Before SPELL (or any of its utility commands) can be used for the first time, the following EMAS 2900 command must be typed:

Command: OPTION SEARCHDIR=KNTLIB.GENERAL

It is needed only once, and is remembered for subsequent sessions.

3 Overview

SPELL operates by scanning the input file and building a sorted list of words. Each word is recorded together with the line number of its first occurrence and a count of its total number of occurrences. This list is then scanned, and the words compared with the contents of the system and private lexicons.

If an unknown word is found during this scan then it is reported, together with its frequency count and the place of its first occurrence. You then have several options, which are described in Section 5.

If a corrected version of the input file is required, any corrections are recorded and a second pass is made over the input file in order to perform the corrections.

4 Using SPELL

The program is entered by the command

Command: SPELL param1,param2,...

At least the first parameter (input) must be supplied with the command.

If the command

Command: SPELL ?

is typed, a summary of parameters and their defaults is given.

Possible parameters are as follows.

Parameter	Position	Use	Default
INPUT	1	Input file The file whose contents are to be checked for correct spelling.	none
OUTPUT	2	Output file The file to which a corrected version of the input file is to be sent. If this parameter is omitted, no corrected file is produced. This file can be the same as the input file, in which case it may be overwritten: it cannot be a member of a pdfile.	none
ERRORS	3	Logging file A list of errors (and any corrections) is sent to this file: it cannot be a member of a pdfile.	T#LOG
MYLEX	4	Private lexicon The file containing your own list of valid words. It is created and cherished if it does not exist: it cannot be a member of a pdfile.	LEXICON
SYSLEX	5	System lexicon The file containing the system word list.	KNTLIB.LEXICON
UPDATELEX	6	Update private lexicon This may take the value YES or NO (or any abbreviation). If the value is YES, your private lexicon is updated with any new words which you have indicated are valid; otherwise no update is done.	YES
WORKSIZE	7	Workspace size The number of words set aside for workspace. This will only need to be increased for large input files. An indication of workspace usage is given at the end of every run.	16340

Parameters, which may be in upper or lower case, may be quoted by position or by keyword, and keywords may be shortened. For example:

Command: SPELL doc11,doc11n,,dict

Command: SPELL in=doc11,my=dict,out=doc11n

5 Interaction with SPELL

During the interactive phase of a SPELL run, you are informed of any dubious words, and asked if they are in fact valid. There are several possible replies:

H Prints a short 'help' text, summarizing the possible replies.

Y The word is considered valid, and is added (if UPDATELEX=YES) to the private lexicon.

N The word is erroneous. This fact is written to the logging file, and no correction is attempted.

E The word is 'eccentric'; this means that it is valid, but does not occur in enough documents to be worth adding to the private lexicon.

W Wind through to the end, marking all other dubious words as erroneous. This can be useful for a background job which checks the spelling of a text and lists all dubious words at a specified line printer or in a logging file for later inspection.

Q Quit. The run is terminated immediately; no corrections are performed, and the private lexicon remains unchanged.

L This has a similar effect to Y, but the word is converted to lower case before being added to the private lexicon. This is useful if the word occurred (say) at the beginning of a sentence. See Section 6 below for a description of the treatment of upper and lower case words.

word The word 'word' is the correct spelling. This fact is logged, and the word corrected wherever it occurs in the input file (if corrections have been requested by the specification of an output file).

=word This has a similar effect to just typing 'word', but the correction is recorded in the private lexicon (if UPDATELEX=YES) so that it can apply to this and *all subsequent* runs of SPELL. This is useful if a word is consistently mis-spelt.

6 Upper and lower case letters

The policy regarding upper and lower case letters is simple but effective. If a word in a lexicon contains upper case letters, then words in the input file are expected to have *at least* those letters in capitals. Thus 'Ada' in the lexicon will match an occurrence of 'ADA' in the input file, but 'APL' in the lexicon will not match 'Apl' in the input file.

7 Apostrophe and Hyphen

Neither abbreviations nor words containing apostrophe or hyphen are included in the system lexicon. You can, however, include such words in your private lexicon.

The program assumes that words ending with an apostrophe or an apostrophe and 's' are in the possessive case, and checks the spelling of the part before the apostrophe. For example, if the word is summers' or summer's then the spelling of summers or summer is checked.

The same procedure is followed for abbreviations which end with an apostrophe and 's', such as it's; these are the only abbreviations where the spelling of part of the whole word is checked.

Similarly, if a word contains a hyphen the entire word is compared with words in your private lexicon. Thus the spelling of self-contained is questioned even though self and contained are each in the system lexicon.

Example

Assume that we have one of Shakespeare's sonnets stored in EMAS:

Command: list sonnets_s018
<S 18>
Shall I compare thee to a Summer's day?
Thou art more louely and more temperate:
Rough windes do shake the darling buds of Maie,
And Sommer's lease hath all too short a date:
Sometime too hot the eye of heauen shines,
And often is his gold complexion dimm'd,
And euery faire from faire some-time declines,
By chance, or nature's changing course vntrim'd:
But thy eternall Sommer shall not fade,
Nor loose possession of that faire thou ow'st,
Nor shall death brag thou wandr'st in his shade,
When in eternall lines to time thou grow'st,
So long as men can breath or eyes can see,
So long liues this, and this giues life to thee.

Now assume that we wish to check the spelling of this sonnet, construct a private lexicon of Elizabethan English and modernize some of the spelling (changing thee to you, etc. neither improves the poetry nor helps reveal the spirit of the sonnet, but it does show how spelling may be 'corrected'!).

Command: spell sonnets_s018,s018,mylex=Elizabethan
SPELL - version E2.1
Warning - cannot open private lexicon - File ELIZABETHAN does not exist
*** Type H for help*
"dimm'd" occurs once, in line 7
Is this a word? e
"eternall" occurs twice, starting in line 10
Is this a word? y
"euery" occurs once, in line 8
Is this a word? y
"faire" occurs thrice, starting in line 8
Is this a word? y
"giues" occurs once, in line 15
Is this a word? y
"grow'st" occurs once, in line 13
Is this a word? e
"hath" occurs once, in line 5
Is this a word? y
"heauen" occurs once, in line 6
Is this a word? y
"liues" occurs once, in line 15
Is this a word? y
"louely" occurs once, in line 3
Is this a word? y
"Maie" occurs once, in line 4
Is this a word? y
"ow'st" occurs once, in line 11
Is this a word? e
"some-time" occurs once, in line 8
Is this a word? e
"Sommer" occurs twice, starting in line 5
Is this a word? y

"thee" occurs twice, starting in line 2
Is this a word? =you
"Thou" occurs once, in line 3
Is this a word? =You
"vntrim'd" occurs once, in line 9
Is this a word? e
"wandr'st" occurs once, in line 12
Is this a word? e
"windes" occurs once, in line 4
Is this a word? y
[Creating new private lexicon "ELIZABETHAN"]

(starting correction pass)
111 words read (85 distinct) in 15 lines
3 errors, all corrected
3% of workspace used

Note that the spelling of five abbreviated words has been questioned: in each case the spelling has been declared eccentric, i.e. acceptable but the word is not to be added to the private lexicon. Similarly the spelling of a hyphenated word is declared to be eccentric.

The spellings of thirteen distinct Elizabethan words (three of which occur more than once) have been questioned. Eleven of these are accepted as correct spellings and so will be added to the private lexicon. For the two other words a modern spelling has been provided which will be stored in the private lexicon: every time this private lexicon is used in the future the spelling of these words can be modernized.

Note also that in line 3 (line 2 of the sonnet) the word art (=are) is not questioned because it is a valid modern English word that is in the system lexicon. The SPELL program can only check spelling, never that a word is used in its proper context.

"Words without thoughts, never to heaven go." *Hamlet III iii 101*

The statement of the number of lines read includes any blank lines.

The result of this machination is as follows:

Command: list s018
<S 18>
Shall I compare you to a Summer's day?
You art more louely and more temperate:
Rough windes do shake the darling buds of Maie,
And Sommer's lease hath all too short a date:
Sometime too hot the eye of heauen shines,
And often is his gold complexion dimm'd,
And euery faire from faire some-time declines,
By chance, or nature's changing course vntrim'd:
But thy eternall Sommer shall not fade,
Nor loose possession of that faire thou ow'st,
Nor shall death brag thou wandr'st in his shade,
When in eternall lines to time thou grow'st,
So long as men can breath or eyes can see,
So long liues this, and this giues life to you.

Now let us check the spelling of another sonnet.

Command: spell sonnets_s001,s001,mylex=Elizabethan
SPELL - version E2.1
*** Type H for help*
"aboundance" occurs once, in line 8
Is this a word? y
"beare" occurs once, in line 5
Is this a word? y
"buriest" occurs once, in line 12
Is this a word? y
"chorle" occurs once, in line 13
Is this a word? y
"cruell" occurs once, in line 9
Is this a word? y
"eate" occurs once, in line 15
Is this a word? y
"Feed'st" occurs once, in line 7
Is this a word? e
"fewell" occurs once, in line 7
Is this a word? y
"graue" occurs once, in line 15
Is this a word? y
"heire" occurs once, in line 5
Is this a word? y
"herauld" occurs once, in line 11
Is this a word? y
"mak'st" occurs once, in line 13
Is this a word? e
"neuer" occurs once, in line 3
Is this a word? y
"niggarding" occurs once, in line 13
Is this a word? y
"owne" occurs twice, starting in line 6
Is this a word? y
"Pitty" occurs once, in line 14
Is this a word? l
"selfe" occurs thrice, starting in line 7
Is this a word? y
"substantiall" occurs once, in line 7
Is this a word? y
"thee" occurs once, in line 15
Do you want it corrected to "you"? y
"thine" occurs twice, starting in line 6
Is this a word? =your
"Thou" occurs once, in line 10
Do you want it corrected to "You"? y
"thou" occurs once, in line 6
Is this a word? =you
"wast" occurs once, in line 13
Is this a word? y

(starting correction pass)
106 words read (85 distinct) in 15 lines
5 errors, all corrected
3% of workspace used

In this the spellings of two abbreviated words have been declared eccentric. Seventeen more words have been added to the private lexicon, as have two more 'corrections'. Note that one of these additions was converted to lower case before being stored. Also you are given the option of applying previously recorded corrections or leaving the spelling alone: in this case they have been applied.

Command: list s001
<S 1>
From fairest creatures we desire increase,
That thereby beauties Rose might neuer die,
But as the riper should by time decease,
His tender heire might beare his memory:
But you contracted to your owne bright eyes,
Feed'st thy light's flame with selfe substantiall fewell,
Making a famine where aboundance lies,
Thy selfe thy foe, to thy sweet selfe too cruell:
You that art now the world's fresh ornament,
And only herauld to the gaudy spring,
Within your owne bud buriest thy content,
And tender chorle mak'st wast in niggarding:
Pitty the world, or else this glutton be,
To eate the world's due, by the graue and you.

8 Storage of lexicons

Private lexicons are stored as ordinary text files; they can be edited with any text editor.

System lexicons are stored in a compressed format. The commands SPELLC and SPELLE (see Section 10) are available to compress and expand ordinary text files. It is not necessary for a system lexicon to be in compressed format, but file size considerations make it desirable to compress all but the smallest of lexicons.

Let us now look at the contents of the private lexicon Elizabethan.

Command: list Elizabethan
aboundance beare buriest chorle cruell eate eternall euery faire fewell
giues graue hath heauen heire herauld liues louely Maie neuer niggarding
owne pitty selfe Sommer substantiall thee=you thine=your Thou=You
thou=you wast windes

An unnecessary item has been included in this lexicon by the chance order in which items were added. Thou=You was added first, and the upper case T precludes thou being included in the corrections. If this item is removed by using an editor then the item thou=you will serve to correct both thou to you and Thou to You.

9 Maintenance of the system lexicon

Requests for words to be added to the system lexicon may be made to the author of this Note. They will be considered on their overall usefulness to the user community. Each word is also checked in either the Concise or Shorter Oxford English Dictionary; if it is described as an abbreviation, archaic, colloquial or slang it is excluded. Such words can, however, be included in private lexicons. It is felt that the sensible maximum size for a lexicon is 40,000 words: at the time of writing the system lexicon contains 30,538 words.

It is desirable that documents should spell any given word in only one way.. This is achieved by holding only one spelling of each word in the system lexicon, even if there is more than one possible way in which it is commonly spelt. In general, the spelling used is that considered most 'English' (for example 'disc' rather than 'disk'). You may well disagree with this, and are free to add alternative spellings to your private lexicons.

Whenever a private lexicon is updated, all words that appear also in the system lexicon (because the latter has been updated) are removed from that private lexicon in order to save space.

You can develop your own system lexicon, for example to check the spelling of words in a foreign language (some of the examples in this Note show how a system lexicon for Elizabethan English might be developed).

10 Utility commands

As mentioned above, there are several utility commands available which are useful for the maintenance of lexicon files. They are described in this Section. In all cases, the command may be issued with a single question mark as a parameter in order to obtain a summary of parameters and their defaults.

10.1 SPELLC

The SPELLC command compresses a lexicon, either for long term storage or for use as a system lexicon. It takes the following parameters:

Parameter	Position	Use	Default
INPUT	1	Input file The lexicon file to be compressed.	none
OUTPUT	2	Output file The file, which must be distinct from the input file, to which the compressed lexicon is to be written. If the file already exists, the existing contents will be overwritten. This file cannot be a member of a pdfile.	none

Parameters may be quoted by position or by keyword, and keywords may be shortened.

Both parameters must be supplied with the command.

For example:

Command: spellc Elizabethan,Lizlex
SPELL lexicon compress utility - version E2.2

If the command

Command: SPELLC ?

is typed, a summary of parameters and their defaults is given.

10.2 SPELLE

The SPELLE command expands a lexicon compressed by SPELLC. It takes the following parameters:

Parameter	Position	Use	Default
INPUT	1	Input file The lexicon file to be expanded.	none
OUTPUT	2	Output file The file, which must be distinct from the input file, to which the expanded lexicon is to be written. If the file already exists, the existing contents will be overwritten. This file cannot be a member of a pdfile.	none

Parameters may be quoted by position or by keyword, and keywords may be shortened.

Both parameters must be supplied with the command.

For example:

Command: spelle Lizlex,Elizabethan
SPELL lexicon expand utility - version E2.2

If the command

Command: SPELLE ?

is typed, a summary of parameters and their defaults is given.

10.3 CHECKLEX

The CHECKLEX command checks an expanded lexicon for duplicate words, and words which are not in the correct place (dictionary order). It takes the following parameters:

Parameter	Position	Use	Default
INPUT	1	Input file The lexicon file to be checked.	LEXICON
OUTPUT	2	Output The destination for any error messages. If this a file, it must be distinct from the input file, and if the file already exists the existing contents will be overwritten. This file cannot be a member of a pdfile.	.OUT

Parameters may be quoted by position or by keyword, and keywords may be shortened.

For example:

Command: checklex Elizabethan
SPELL lexicon check utility - version E2.2
No errors; 32 words in lexicon "ELIZABETHAN"

If the command

Command: CHECKLEX ?

is typed, a summary of parameters and their defaults is given.

10.4 STRIPLEX

The STRIPLEX command removes all words from a lexicon except those containing only lower case letters. Thus it could be used to remove any copies of words containing upper case letters, for example those added by responding Y instead of L when updating a private lexicon. However, STRIPLEX would at the same time remove all proper nouns. It takes the following parameters:

Parameter	Position	Use	Default
INPUT	1	Input file The lexicon file to be 'stripped'.	none
OUTPUT	2	Output file The file to which the new lexicon is to be written. If the file already exists it must be distinct from the input file and its existing contents will be overwritten. This file cannot be a member of a pdfile.	none

Parameters may be quoted by position or by keyword, and keywords may be shortened.

Both parameters must be supplied with the command.

If the command

Command: STRIPLEX ?

is typed, a summary of parameters and their defaults is given.

STRIPLEX could have been used instead of an editor to remove Thou=You from the private lexicon (see Section 8 above), but it would also remove Maie and Sommer:

Command: striplex Elizabethan,Eliz
SPELL lexicon strip utility - version E2.2
29 words in lexicon "ELIZ"
**** 3 bad words removed*

Command: list Eliz
aboundance beare buriest chorle cruell eate eternall euery faire fewell
giues graue hath heauen heire herauld liues louely neuer niggarding owne
pitty selfe substantiall thee=you thine=your thou=you wast windes

For the purpose of later examples let us assume that Thou=You was removed by an editor, leaving 31 items in Elizabethan.

10.5 LEXMERGE

The LEXMERGE command merges two expanded lexicons; these are assumed to be in the correct order. It is particularly useful when several people are writing a document; they can ensure, at intervals, that they pick up all the words used by other members of the team. The command takes the following parameters:

Parameter	Position	Use	Default
INPUT1	1	First input file One of the lexicon files to be merged.	none
INPUT2	2	Second input file The other lexicon file to be merged.	none
OUTPUT	3	Output file The file, which must be distinct from each input file, to which the new (merged) lexicon is to be written. If the file already exists, the existing contents will be overwritten. This file cannot be a member of a pdfile.	none

Parameters may be quoted by position or by keyword, and keywords may be shortened provided they remain unambiguous.

All three parameters must be supplied with the command.

If the command

Command: LEXMERGE ?

is typed, a summary of parameters and their defaults is given.

Suppose that the private lexicon Lizlex had been used as the system lexicon and that a new private lexicon Elizabethan had been constructed, for example:

Command: spell Lucrece,syslex=Lizlex,mylex=Elizabethan
SPELL - version E2.1
**** WARNING - Corrections will not be applied to file ****
Warning - cannot open private lexicon - File ELIZABETHAN does not exist
:
:

Then the two lexicons Lizlex and Elizabethan might be merged to form a single lexicon which could be used as the new system lexicon. Before merging lexicons it is necessary to expand any which are stored in compressed format and it is advisable to check them by using CHECKLEX.

Command: checklex Elizabethan
SPELL lexicon check utility - version E2.2
No errors; 253 words in lexicon "ELIZABETHAN"

Command: spelle Lizlex,Liz1
SPELL lexicon expand utility - version E2.2

(Liz1 is derived from the first Elizabethan which had been checked.)

Command: lexmerge Liz1,Elizabethan,Newliz
SPELL lexicon merge utility - version E2.2
279 words in new lexicon "NEWLIZ"

If any word appears in both lexicons then LEXMERGE will place only a single copy of this word in the new lexicon. Hence if the two lexicons are checked before merging then it is unnecessary to check the third lexicon after it is created. In this example five words were obviously common to each lexicon: there is no easy way to identify these words.

11 The DOUBLE Command

A common error in writing documents is the unintentional repetition of a word, particularly where the first occurrence is the last word on one line and the second occurrence is the first word on the next line. The DOUBLE command checks that no two adjacent words in a file are in fact the same. It takes the following parameters:

Parameter	Position	Use	Default
INPUT	1	Input file The file to be checked.	none
OUTPUT	2	Output file The destination for any error messages. If this a file which already exists, then the existing contents will be overwritten. This file cannot be a member of a pdfile.	.OUT

Parameters may be quoted by position or by keyword, and keywords may be shortened.

At least the first parameter (input) must be supplied with the command.

If the command

Command: DOUBLE ?

is typed, a summary of parameters and their defaults is given.

Using a poem instead of a contrived text might make a more interesting example.

Command: list Expiration

The Expiration

So, so, breake off this last lamenting kisse,
Which sucks two soules, and vapors Both away,
Turne thou ghost that way, and let mee turne this,
And let our selves benight our happiest day,
We ask'd none leave to love; nor will we owe
Any, so cheape a death, as saying, Goe;

Goe; and if that word have not quite kil'd thee,
Ease mee with death, by bidding mee goe too.
Or, if it have, let my word worke on mee,
And a just office on a murderer doe.
Except it be too late, to kill me so,
Being double dead, going, and bidding, goe.

Command: double Expiration
SPELL double word check utility - version E2.2
Two occurrences of "Goe" near line 11
1 error

Note that this matching is case dependent. Thus if the Goe at the end of the first stanza had lacked the upper case G then it would not have been matched with the Goe at the start of the second stanza.

Shakespeare correctly identified the author of this poem when he wrote "what's done is done" *Macbeth III ii 11,* but then John Donne developed this pun better in one of his own poems!

12 Acknowledgement

The program and utilities described in this Note were written by Bob Eager of The University of Kent at Canterbury and are based heavily on two versions of the SPELL program developed by Peter Robinson and Dave Singer of the University of Cambridge. For more details, see CACM, Vol. 24, No. 5, pp. 296-297 (May 1981).

This Note is based on a document written by Bob Eager.