Patterns of misspellings in L2 English – a view from the ETS Spelling Corpus

Michael Flor,
Yoko Futagi,
Melissa Lopez,
Matthew Mulholland

NLP & Speech Group, R&D Division
Educational Testing Service, Princeton, NJ, USA

mflor@ets.org

The Corpus

- Initial Motivation:
  evaluate speller performance
  by comparing it to human-annotated data.
- We needed:
  a human-annotated corpus of misspellings,
  where misspellings appear in their original context.
- Method:
  Use ConSpel to generate a corpus automatically,
  then let human annotators work on it…

The Corpus

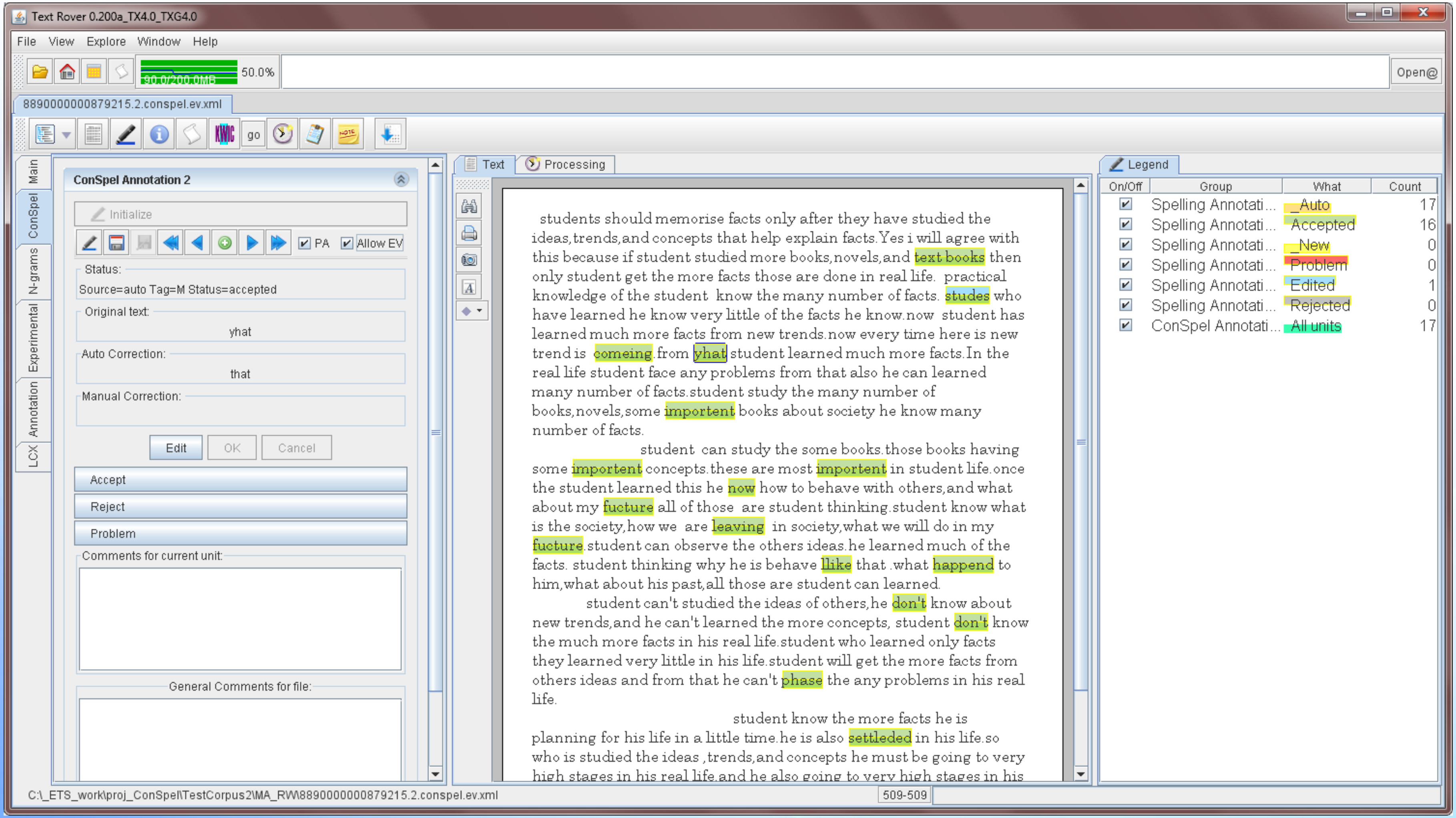
Materials – English essays written on TOEFL and GRE tests at international testing centers around the world. (computer-based delivery, QWERTY keyboard),

Program/task	Description of writing activity
TOEFL Independent	support an ***opinion*** in writing (topic assigned).
TOEFL Integrated	write essay responses based on reading and listening tasks (**summarize and compare arguments**)
GRE Issue	express ***opinion*** clearly, in writing, about a topic of general interest (topic assigned).
GRE Argument	analyze and evaluate **arguments** according to specific instructions and convey evaluation clearly in writing.

The Corpus

- 4 program/task groups
- 10 different prompts for each task
- 75 essays per prompt
- Total: 3,000 essays (963K words)
- Essay length ranges from 29 to798 words, average 321 words

Annotation software



ConSpel Annotation 2

☑ PA ☑ Allow EV

Status:
Source=auto Tag=M Status=accepted

Original text:
yhat

Auto Correction:
that

Manual Correction:

Comments for current unit:

General Comments for file:

students should memorise facts only after they have studied the ideas,trends,and concepts that help explain facts.Yes i will agree with this because if student studied more books,novels,and text books then only student get the more facts those are done in real life. practical knowledge of the student know the many number of facts. studes who have learned he know very little of the facts he know.now student has learned much more facts from new trends.now every time here is new trend is comeing.from yhat student learned much more facts.In the real life student face any problems from that also he can learned many number of facts.student study the many number of books,novels,some importent books about society he know many number of facts.

student can study the some books.those books having some importent concepts.these are most importent in student life.once the student learned this he now how to behave with others,and what about my fucture all of those are student thinking.student know what is the society,how we are leaving in society,what we will do in my fucture.student can observe the others ideas.he learned much of the facts. student thinking why he is behave llike that .what happend to him,what about his past,all those are student can learned.

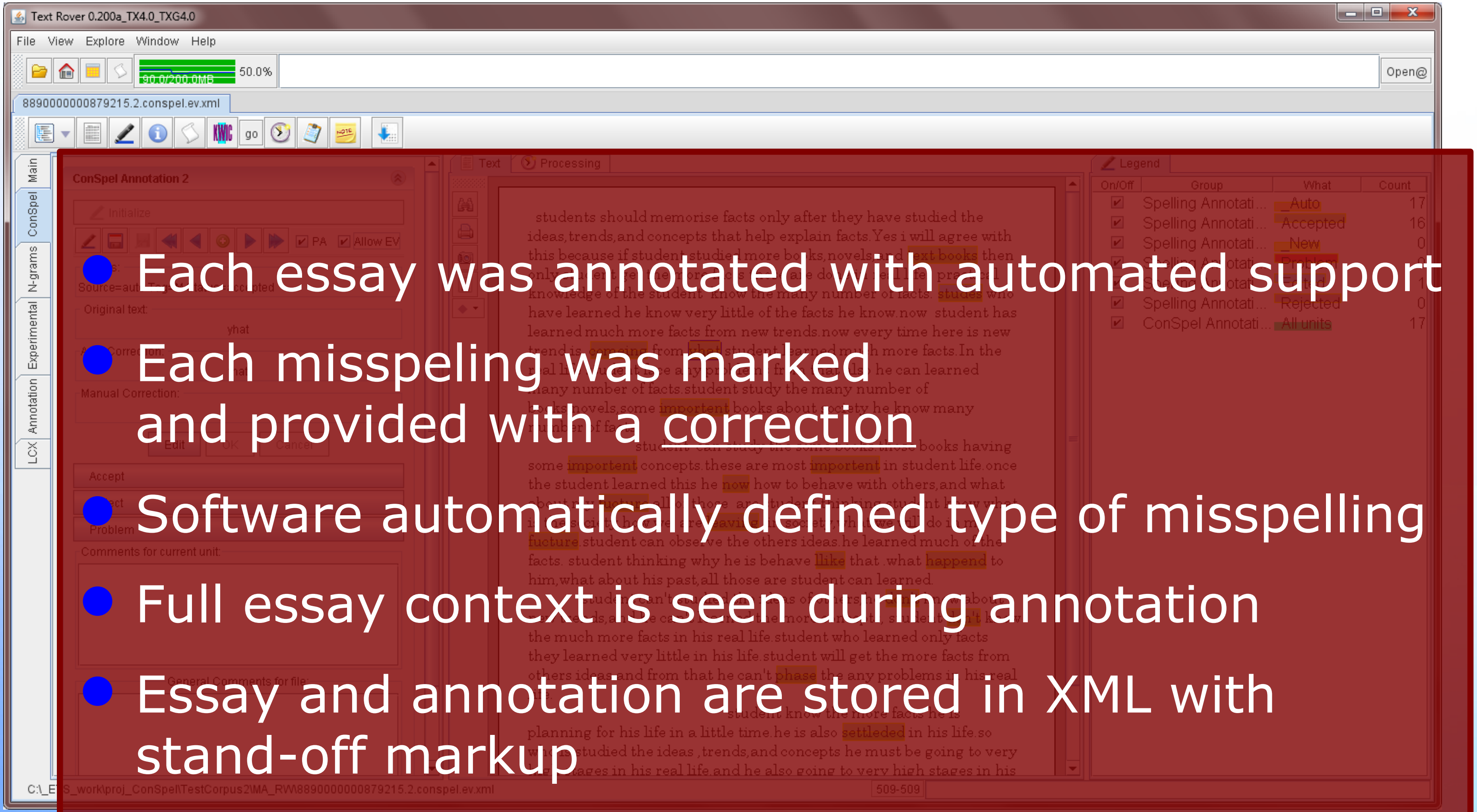
student can't studied the ideas of others,he don't know about new trends,and he can't learned the more concepts, student don't know the much more facts in his real life.student who learned only facts they learned very little in his life.student will get the more facts from others ideas and from that he can't phase the any problems in his real life.

student know the more facts he is planning for his life in a little time.he is also settleded in his life.so who is studied the ideas ,trends,and concepts he must be going to very high stages in his real life.and he also going to very high stages in his

On/Off	Group	What	Count
☑	Spelling Annotati...	_Auto	17
☑	Spelling Annotati...	Accepted	16
☑	Spelling Annotati...	_New	0
☑	Spelling Annotati...	Problem	0
☑	Spelling Annotati...	Edited	1
☑	Spelling Annotati...	Rejected	0
☑	ConSpel Annotati...	All units	17

C:_ETS_work\proj_ConSpel\TestCorpus2\MA_RW\8890000000879215.2.conspel.ev.xml 509-509

Annotation software


Text Rover 0.200a_TX4.0_TXG4.0
File View Explore Window Help
90.0/200.0MB
50.0%
Open@
8890000000879215.2.consPel.ev.xml
go
ConSpel Annotation 2
Initialize
PA
Allow EV
Original text:
Manual Correction:
Edit
OK
Cancel
Accept
Comments for current unit:
General Comments for file:
Main
ConSpel
N-grams
Experimental
Annotation
LCX
Text
Processing
Legend
On/Off
Group
What
Count
Spelling Annotati... _Auto 17
Spelling Annotati... Accepted 16
Spelling Annotati... _New 0
Spelling Annotati... Rejected 0
ConSpel Annotati... All units 17
C:_ETS_work\proj_ConSpel\TestCorpus2\MA_RW\8890000000879215.2.conspel.ev.xml
509-509

- Each essay was annotated with automated support
- Each misspeling was marked and provided with a correction
- Software automatically defined type of misspelling
- Full essay context is seen during annotation
- Essay and annotation are stored in XML with stand-off markup

The Corpus

Inter-Annotator Agreement

- Each essay was annotated by two annotators.
- Annotators strictly agreed in 82.6% the cases.
- Inter-annotator agreement was calculated over all words of the corpus: 99.3%.
- Cohen’s Kappa=0.85, p<0.001.
- All differences and difficulties were resolved by a third annotator (adjudicator).

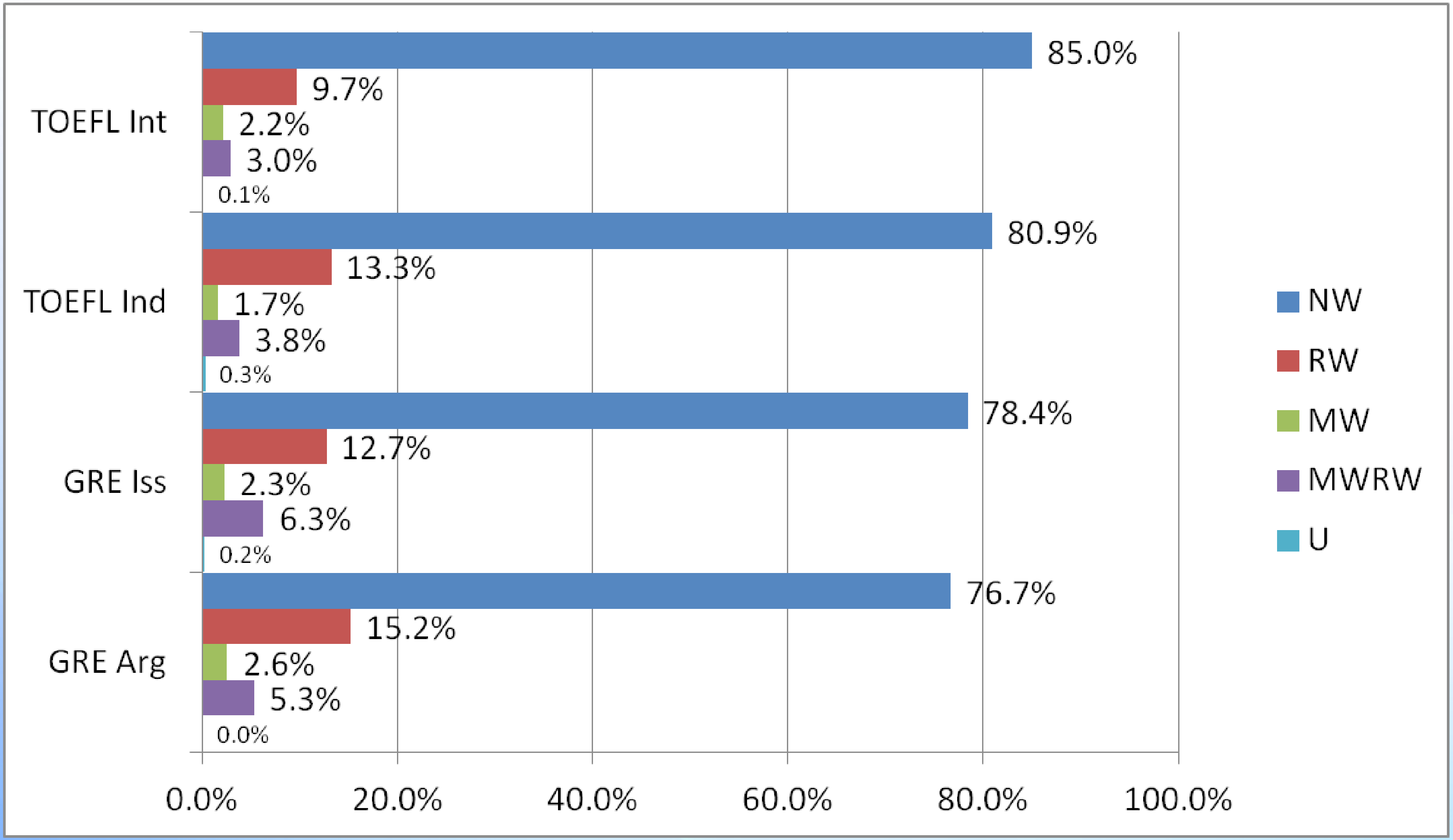
Types and counts of misspellings

	Description	Count in corpus
1	single token non-word misspelling (e.g. "**businees**") also includes fusion errors (e.g. "**niceday**" for "***nice day***")	21142 **(80.05%)**
2	misspelling (?) (non-word token for which no plausible correction was found)	52 **(0.20%)**
3	single token real-word misspelling (e.g. "**they**" for "***then***")	3393 **(12.85%)**
4	multi-token misspelling with at kleast one non-word (e.g. "**mor efun**" for "***more fun***")	574 **(2.17%)**
5	multi-token real-word misspelling (e.g. "**with out**" for "***without***")	1251 **(4.73%)**
	Total	26412 **(100%)**

Breakdown by program/task

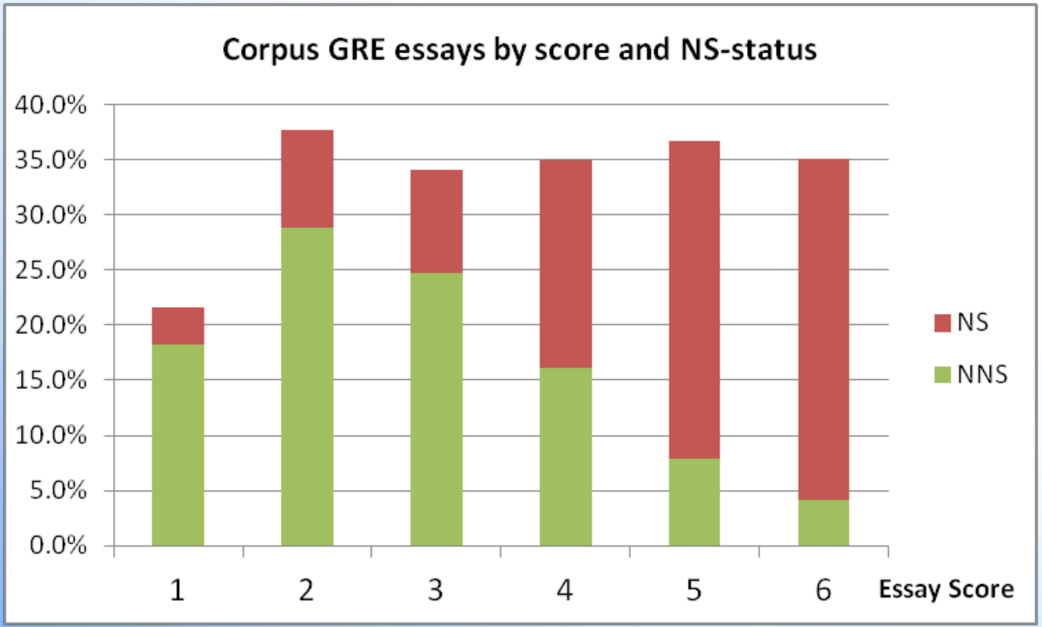
	GRE Argument	GRE Issue	TOEFL Independent	TOEFL Integrated	TOTAL
Essays	**750**	**750**	**750**	**750**	**3,000**
Without misspellings	**60**	**21**	**18**	**21**	**120**
Word Count	**263,578**	**336,301**	**212,930**	**151,031**	**963,840**
Average WC	***351***	***448***	***284***	***201***	***321***
Misspellings	**5935**	**7962**	**7285**	**5230**	**26412**
% of all words	**2.25%**	**2.37%**	**3.42%**	**3.46%**	**2.74%**

Breakdown by error-type and program/task



Count of NS/NNS essays

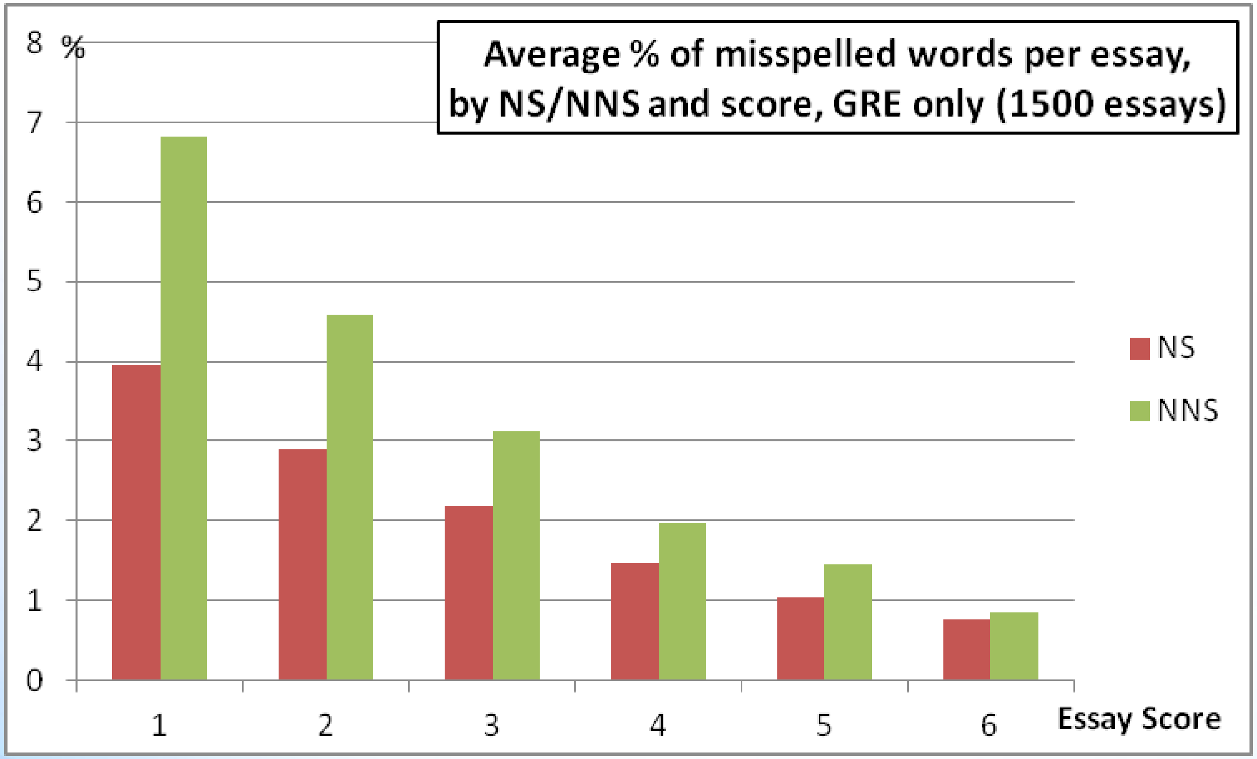
	TOEFL	GRE	Total count	Essays without misspellings
NS	19	634	653	67 **(10.7%)**
NNS	1481	866	2347	53 **(2.3%)**


Corpus GRE essays by score and NS-status

Non-native speakers of English (ELLs) are more prone to making spelling errors ?

Consider proficiency

Spelling Error density



- For each population, average percent of misspelled words (per essay) decreases with better proficiency

- There is a gap between NS & NNS at lower proficiencies,
  (native English speakers make less misspellings, on average)
  but the gap is closing ‘quickly’ ! (both main effects and interaction are sig., p<.0001)

How often is the first character different?

	total	1st diff	%
Non-word	21142	522	2.47%
Real word	3393	404	11.91%
Multi-token	574	10	1.74%
Multi-token RW	1251	7	0.56%

Breakdown by NS/NNS

		total	1st diff	%
Non-word	NNS	18264	465	2.50%
	NS	2878	57	1.98%
Real word	NNS	3008	361	12.00%
	NS	385	43	11.17%

Examples

Non-words

onformation	information
imerged	emerged
onther	another
htis	this
phorensic	forensic
tasttime	pasttime

Real words

write	right
equality	quality
asocial	social
affect	effect
participated	anticipated
as	has

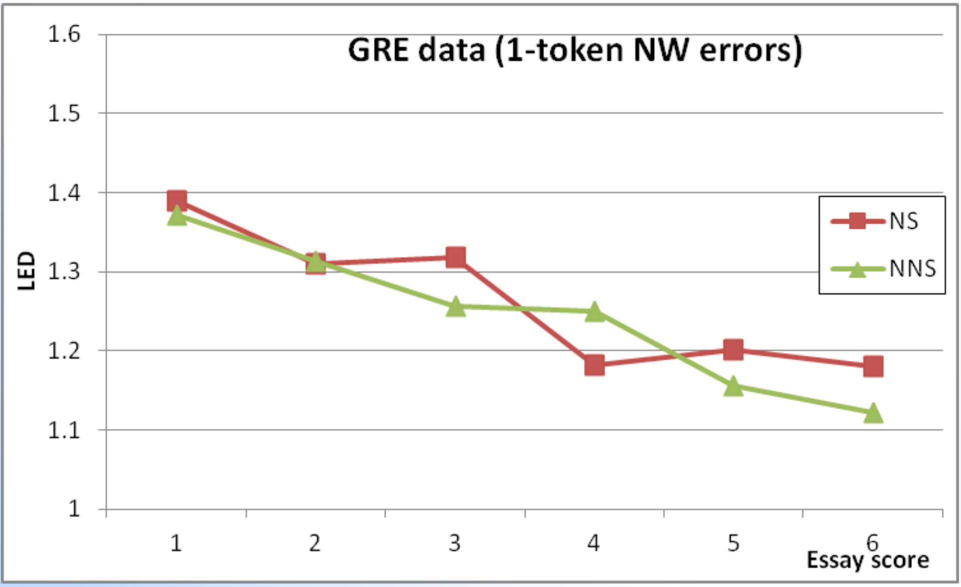
Edit distance (error-to-CorrectForm) 1-token NW

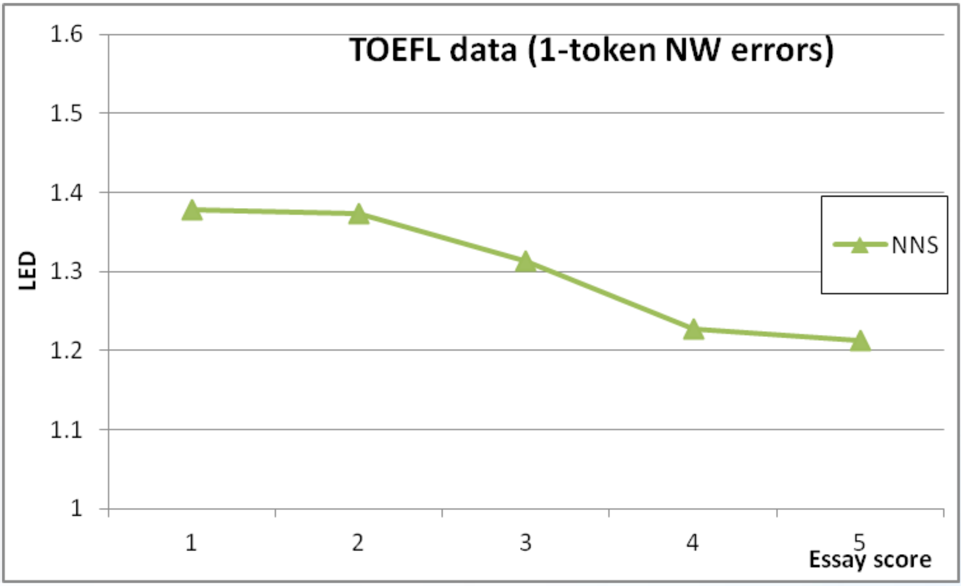
Dist. (LED)	Total tokens	Count NS	% NS	Count NNS	% NNS
1	16908	2393	83.15%	14515	79.47%
2	2957	372	12.93%	2585	14.15%
3	827	88	3.06%	739	4.05%
4	296	22	0.76%	274	1.50%
5	100	2	0.07%	98	0.54%
6	41	1	0.03%	40	0.22%
7	7			7	0.04%
8	2			2	0.01%
9	4			4	0.02%

recom recommendation (9)
unsatisfy dissatisfaction (9)
naiberhouad neighborhood (6)
chraterics characteristics (5)
voultaneer volunteer (4)
metirals materials (3)

The difference
83.1% vs. 79.4%
is significant (p <.0001),
but misleading

Edit distance (error-to-CorrectForm) 1-token NW



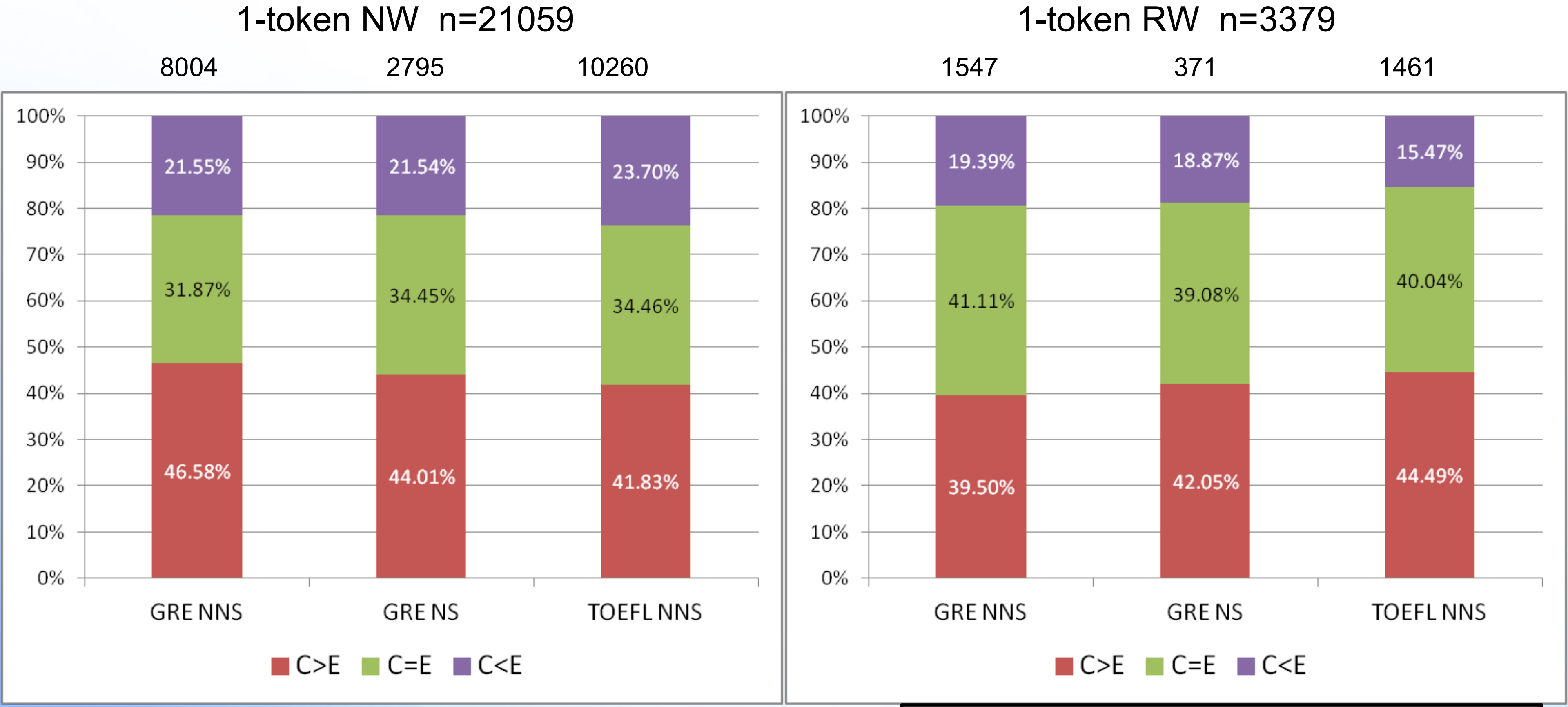


GRE data:
significant main effect of Score (p<0.001),
no effect of NS/NNS, (p=0.38)
and no interaction (p=0.155).

TOEFL data:
significant effect of Score (p<0.001).

For 1-token NW errors,
'severity of error' (DLED) depends on proficiency, not NS/NNS distinction; and yet…

Length of error-form vs. correct-form

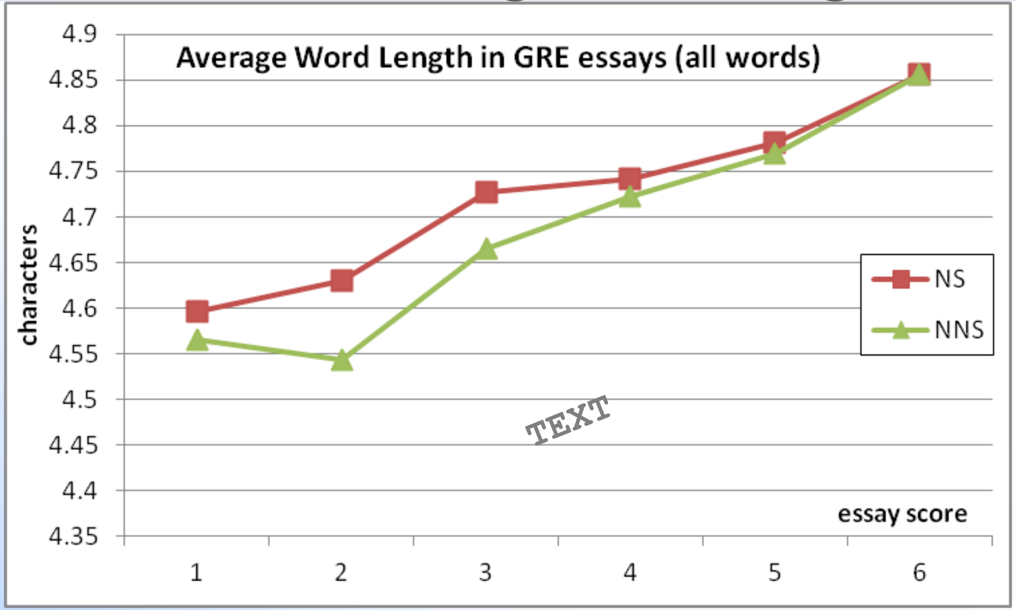


```
Onformation (=) information
as (<) has
asocial (>) social
```

For 1-token NW errors, and for 1-token RW errors:

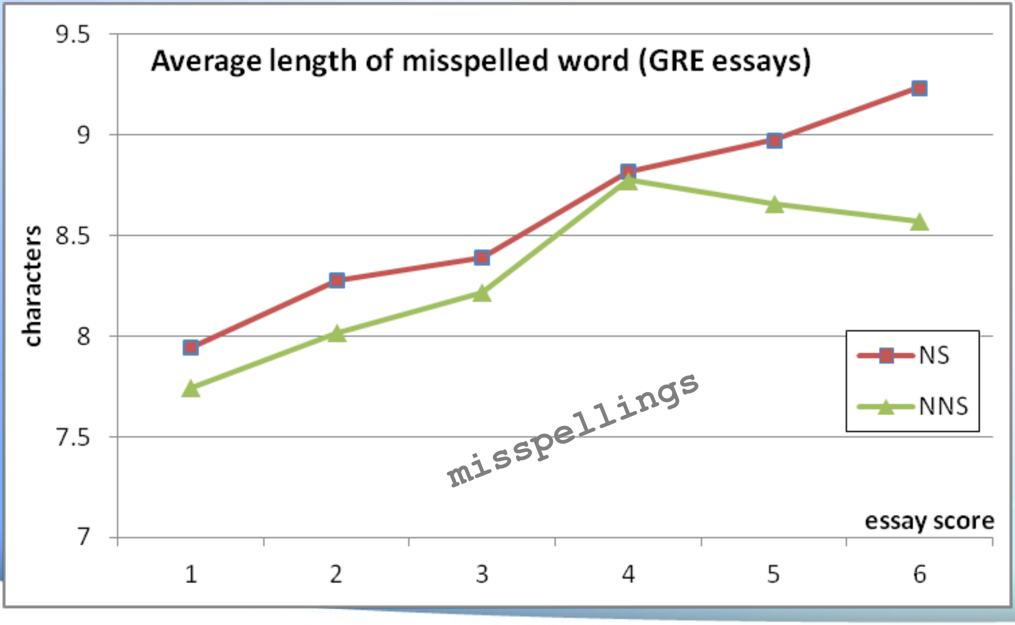
For all groups, when a word is misspelled,
there is a tendency to 'miss' characters, rather than to 'add' characters!
And a strong tendency to preserve length!

Average word length and spelling (1-token NW)



1500 GRE essays

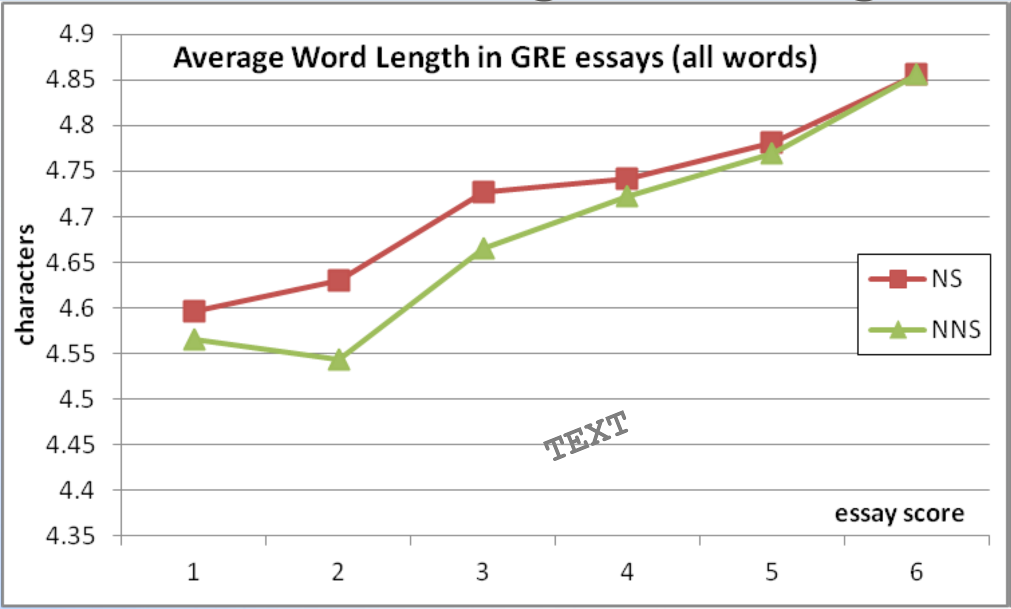
- Average word length (per essay) <u>increases</u> with better proficiency.
- NS typically use more long words
- The gap is rapidly closing with better proficiency
- (both main effects and interaction are sig., p<.0001)

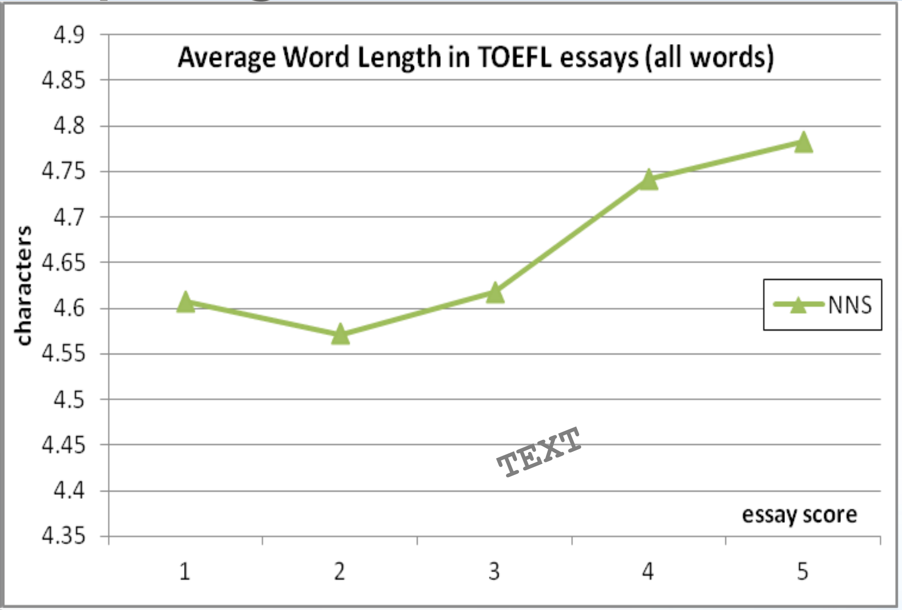


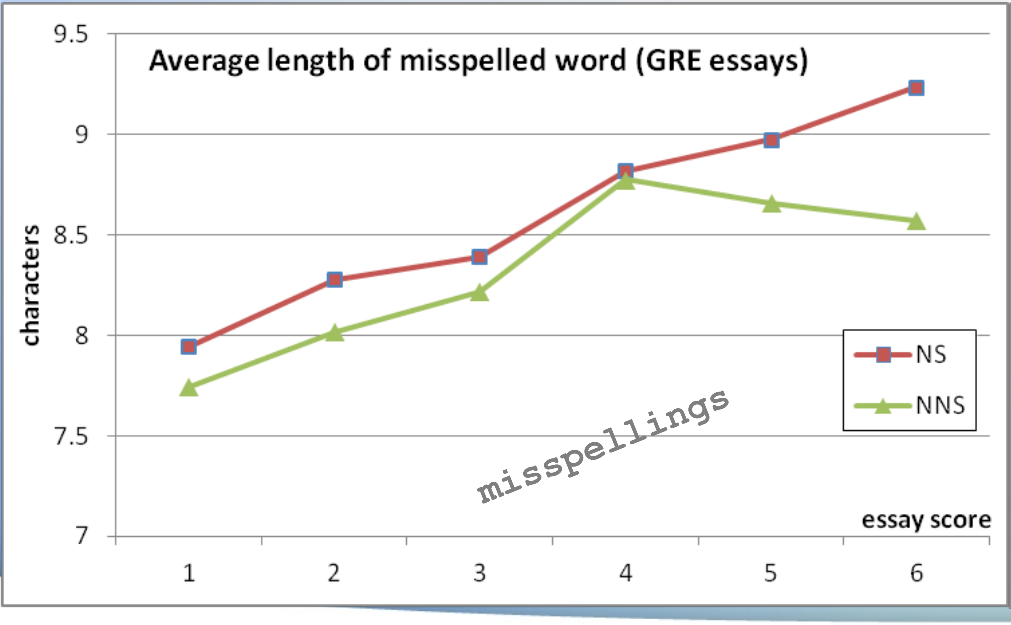
10110 1-token NW errors (GRE essays)

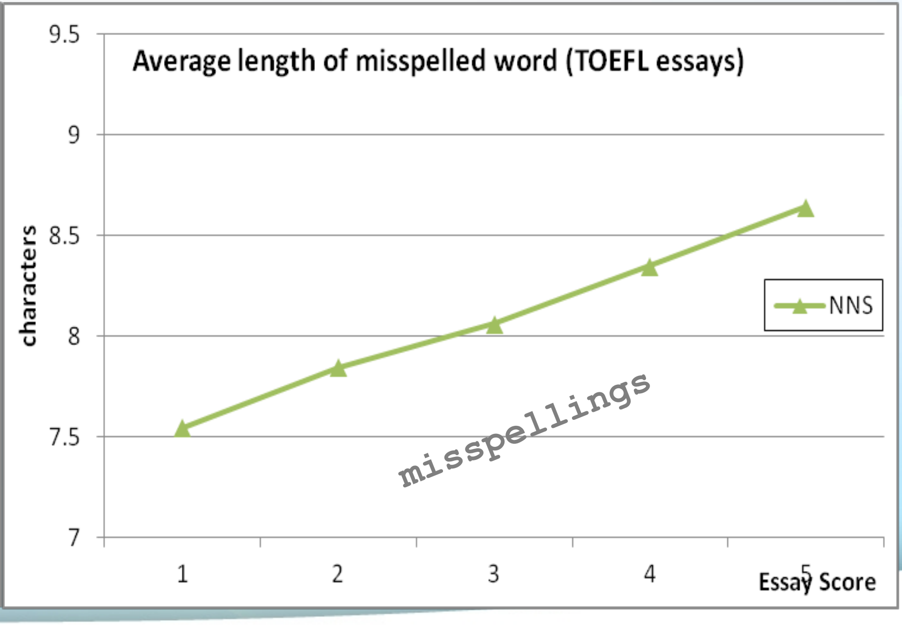
- Average length of intended word (misspelled to NW) <u>increases</u> with better proficiency.
- NS typically err in the longer words
- The gap closes at score=4, then widens!
- (both main effects and interaction are sig., p<.0001)

Average word length and spelling (1-token NW)

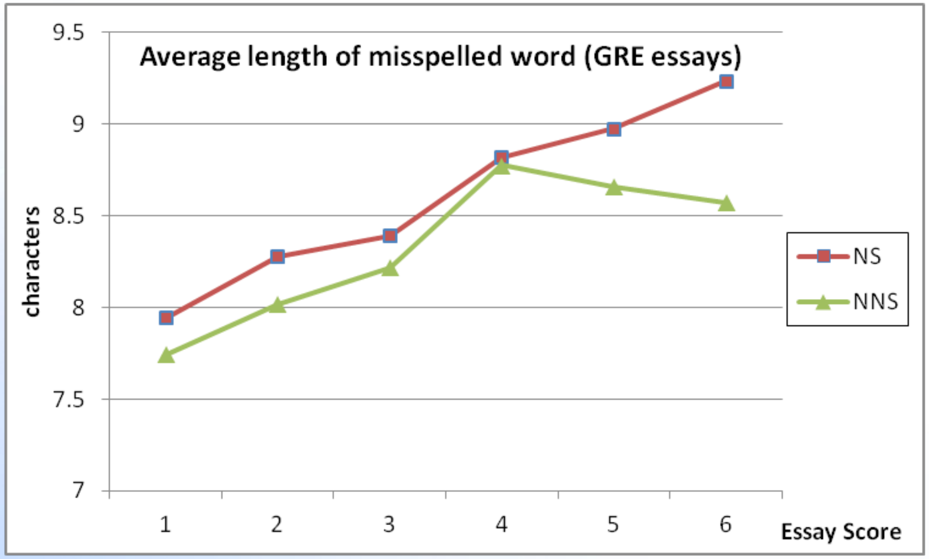


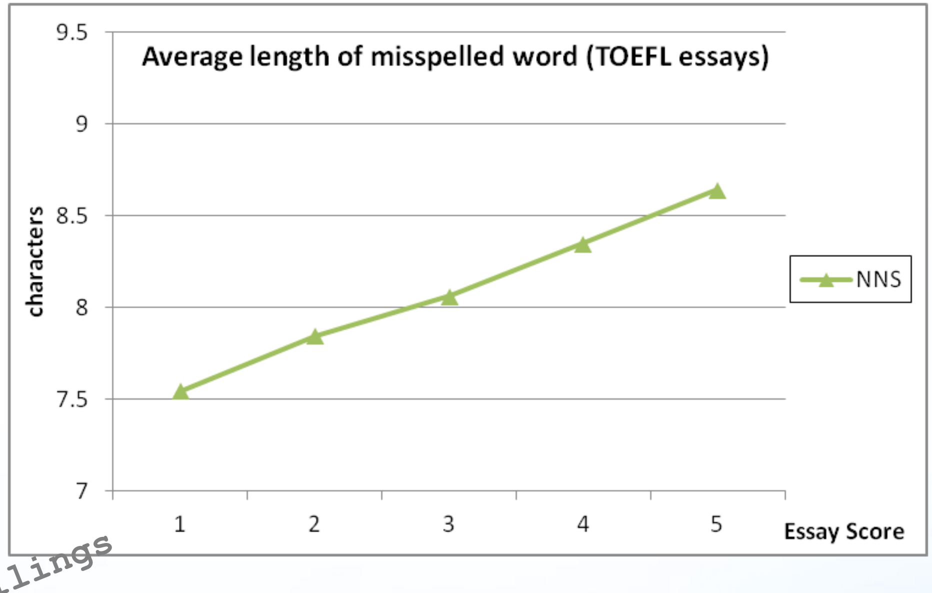




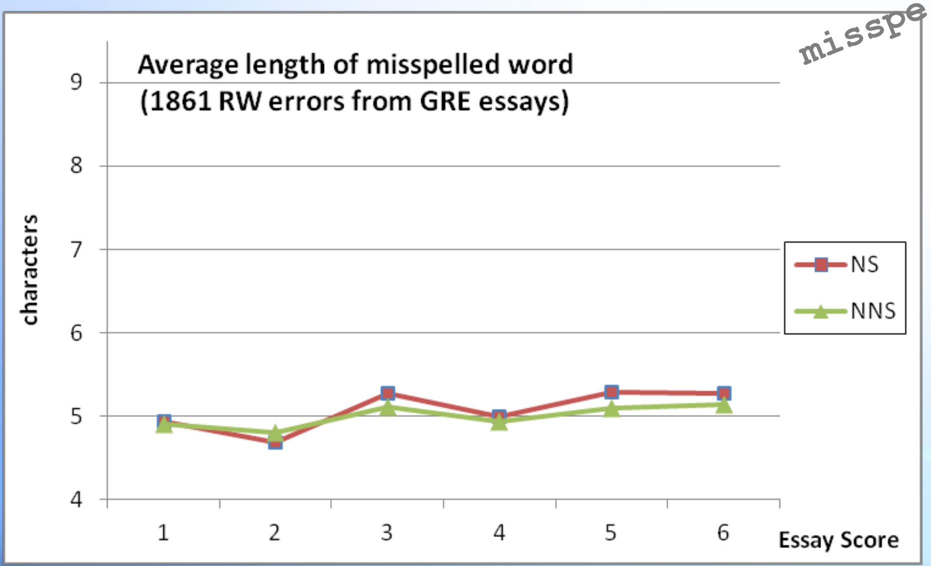


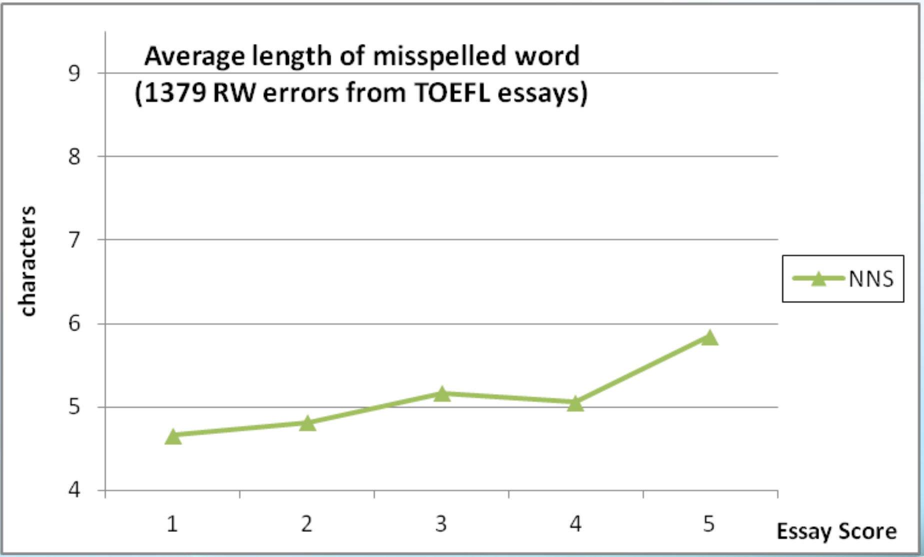
Average word length and spelling: NW vs RW


Average length of misspelled word (GRE essays)

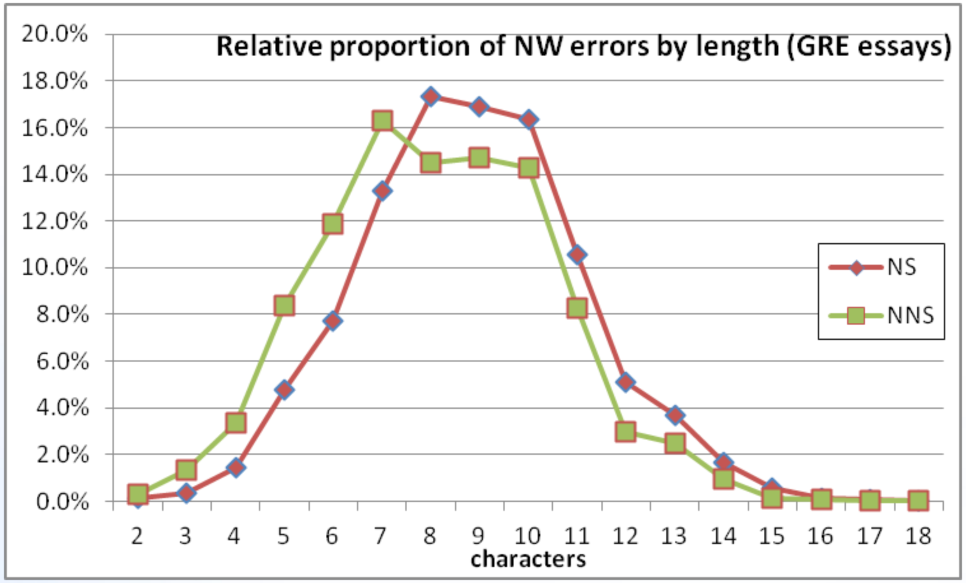

Average length of misspelled word (TOEFL essays)

misspellings

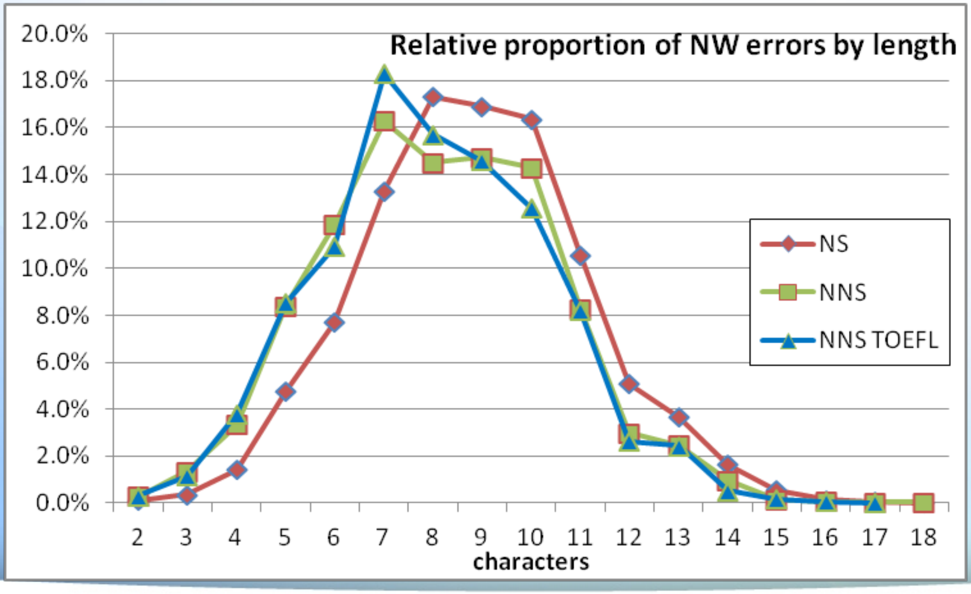

Average length of misspelled word (1861 RW errors from GRE essays)


Average length of misspelled word (1379 RW errors from TOEFL essays)

Average word length and spelling (1-token NW)

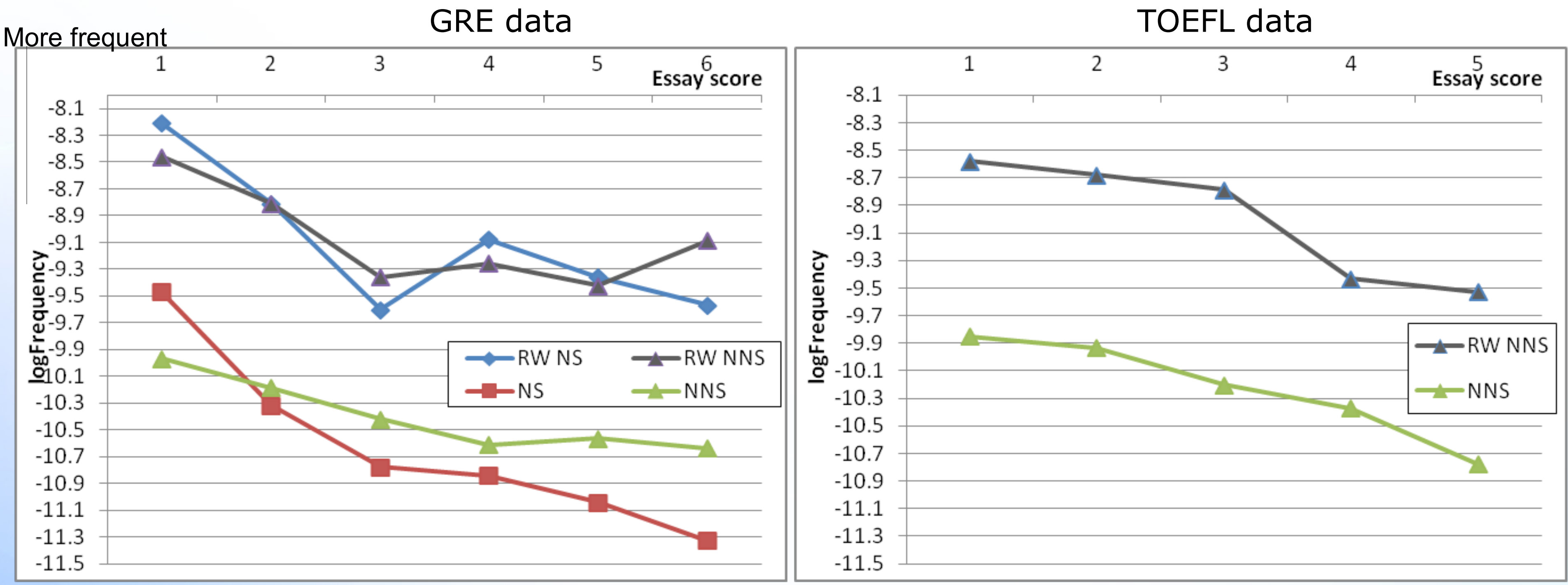


1500 GRE essays



+1500 TOEFL essays

Word frequency and spelling (1-token NW & RW)



logFrequency of the corrected-form of a misspelling

onformation information

For 1-token NW errors, GRE data: both main effects and interaction are sig., p<.002.
For 1-token RW errors, GRE data: no effect is sig. (even Score p=0.71).
TOEFL data, for each NW and RW: effect of Score is sig., p<.001.

The differences between NW and RW are sig. (p<.001) in each of 3 comparisons:
The average frequency of words where RW errors are made is higher than average frequency of words where NW errors are made.

Thank You

mflor@ets.org